The emergence of the data science profession

Philipp Soeren Brandt

Submitted in partial fulfillment of the
requirements of the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

The emergence of the data science profession

Philipp Soeren Brandt

This thesis studies the formation of a novel expert role—the data scientist—in order to ask how arcane knowledge becomes publicly salient. This question responds to the two-sided public debate, wherein data science is associated with problems such as discriminatory consequences and privacy infringements, but has also become linked with opportunities related to new forms of work. A puzzle arises also, as institutional boundaries have obscured earlier instances of quantitative expertise. Even a broader perspective reveals few expert groups that have gained lay salience on the basis of arcane knowledge, other than lawyers and doctors.

This empirical puzzle recovers a gap in the literature between two main lines of argument. An institutionalist view has developed ways for understanding expert work with respect to formal features such as licensing, associations and training. A constructivist view identifies limitations in those arguments, highlighting their failure to explain many instances in which arcane knowledge emerges through informal processes, including the integration of lay knowledge through direct collaboration. Consistent with this critique, data nerds largely define their work on an informal basis. Yet, they also draw heavily on a formalized stock of knowledge. In order to reconcile the two sides, this thesis proposes viewing data science as an emerging “thought community.” Such a perspective leads to an analytical strategy that scrutinizes contours that emerge as data nerds define arcane expertise as theirs.

The analysis unfolds across three empirical settings that complement each other. The first setting considers data nerds as they define their expertise in the context of public events in New York City’s technology scene. This part draws on observations beginning in 2012, shortly after data science’s first lay recognition, and covers three years of its early emergence. Two further studies comparatively test whether and in what ways contours of data science’s abstract knowledge are associated with its lay salience. They respectively consider economic and academic settings, which are most relevant to data nerds in part one. Both studies leverage specifically designed quantitative datasets consisting of traces of lay knowledge recognition and arcane knowledge construction.

Together the three studies reveal distinctive contours of data science. The main argument that follows suggests that data science gains lay salience because it relies on informal practices for recombining formal principles of knowledge construction and application, in a collective effort. Data nerds define their thought community on the basis of illustrative and persuasive tactics that combine formal ideas with informal interpretations. This form of improvisation leads data nerds to connect diverse substantive problems through an array of formal representations. They thereby undermine bureaucratic control that otherwise defines tasks in the context where data scientists mostly apply their arcane knowledge. Despite its name and arcane content, moreover, data science differs from scientific principles of knowledge construction.

The main contribution of this thesis is a first detailed and multifaceted analysis of data science. Results of this study address the main public problems. This thesis demonstrates that data science creates new opportunities for work provided that data nerds are willing to embrace the uncertainty associated with a formally undefined area of problems. The first perspective, focusing on community identification principles, furthermore allows identifying new forms of work in the ongoing technological transformation data science is part of. At the same time, the main argument supports reason for concerns as well precisely because data nerds often operate on an individually anonymous basis, despite their association with formal organizations. It has remained unclear how to address the social consequences of their work because data nerds undermine those conventional forms of control and oversight. The findings of this thesis suggest that although data nerds depart from scientific principles for identifying relevant problems, they coordinate those deviant activities through forms of discipline that qualitatively resemble those common in academic fields. Data nerds define their knowledge as a community. It follows that embedding public concerns in data science's disciplinary forms of coordination, and enhancing those forms, offers the most effective mechanisms for preserving the utility of data science applications while limiting their potentially harmful consequences.

Finally, conceptual and methodological contributions follow as well. The focus on thought communities reveals new leverage for understanding social processes that unfold as a combination of informal activities in local settings and institutional dynamics that are largely removed from individual actors. This problem is common for many instances of skilled work. This additional leverage is the result of an integrated methodological design that relies as much on qualitative observations as on formal analyses. As part of this integration this thesis has directly encoded phenomenologically salient contours into a quantitative design, effectively leading to an analysis of data science through data science.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I first of all thank my advisor, Peter Bearman. He has granted me freedom to explore the discipline and to develop my taste, and then engaged with my ideas to a level of specificity I was too often unprepared for. Peter still had the patience to let me catch up, and to point out possible steps. I thank him for the privilege of such a rigorous and complete intellectual experience. Without Josh Whitford I would not have become a sociologist. Looking back, I can only imagine how tedious it must have been to guide me on this way. I am grateful to Josh for bearing with me. David Stark showed me how to be a member of this community. There is much less use in being a sociologist without it. Gil Eyal has shaped and sharpened my thinking over the years of working on this project. Cat Turco, although she has joined my committee late, through her work provided me with the enthusiasm for taking the last steps.

Karen Barkey guided me to the main ideas with which I came to understand the case I study in this dissertation. Alondra Nelson, Andrew Schrank, Bruce Kogut and Paolo Parigi in various times and forms helped me keep up with a case that was hardly recognizable when I began studying it. Henning Hillmann provided me with the ideal environment to finish this dissertation. Christofer Edling made the whole journey possible.

Among my fellow graduate students, I first thank Matthijs de Vaan, who has become as much a friend as an intellectual peer. Byungkyu Lee has made our office an intellectually more interest and a more pleasant place for getting better work done. I also thank my cohort, the best I have seen come through the department, consisting of Kinga Markovi, Joan Robinson, Abby Coplin, Nate Dern, Jared Conrad-Bradshaw, Elyakim Kislev, adopted member Ryan Hagen, and those nearby, Fabien Accominotti, Alix Rule, Jose Arita, Sarah Sachs, Luciana de Souza Leão, Adam Obeng, Ivana Katic, Pierre Fink, Anna Kaiser and Joscha Legewie. With them, graduate school made sense.

Graduate school is an unusual experience for most, for various reasons. Mine was that that of an international student who discovered a new intellectual world as well as a new physical world. With Claudius Hildebrand I sorted out which one is which–what a trip.

I would also like to thank my friends and family. My siblings have given me the confidence that I could overcome problems I encountered along the way. My parents gave me everything I needed to embark on this journey, and to finish it. For this reason, this dissertation is dedicated to them.

Finally, I am most grateful to my partner, Natalie Schnelle, who has shared every moment with me along this way. She patiently listened to my less-than-half-cooked ideas about interesting—at least I thought so—problems to study, and still had the energy to encourage me when I did not have any. It would be hard to find a piece in this work that she has not shaped.

To my parents

1 Data science: Puzzles, privacy and work

This thesis asks how over the past few years nerds have defined a community with public relevance on the basis of esoteric knowledge. The significance of this process can be seen when we consider those occupations that have long been familiar to us. We think about mechanics, for instance, and immediately know they fix things, or of accountants, who do the books and doctors who heal diseases. Then we note quickly that it is not so clear what most nerds do for us, or anyone, although we might imagine them working on technical and arcane problems, with specialized and proprietary applications. Contrary, today's data nerds are seen to address a series of broadly relevant problems, such as commercial decisions, political engagement and public policy. The press, industry, academia, and the nerds themselves, have come to view their quantitative definition of these manifold problems as "data science." Suspiciously evocative, such consensus too easily conceals the rich and complex basis of data science's salience and therefore denies leverage over its consequences, which pertain to how we live and work.

Data scientists can and should be considered as a great surprise for a number of reasons. It is not only that most nerds work in obscurity, while data scientists have gained salience. Looking at other groups with esoteric expertise, including many from the scientific setting which data nerds take both their label and knowledge from, we find that gaining lay recognition is not so easy, that it happens neither often nor regularly, and that current solutions to data problems make their appearance particularly unlikely.

Let me just briefly address these points. The constructivist view in sociology demonstrates in so many instances the protracted processes of institutionalizing reasons that make it obvious when we need to consult a physician, lawyer or accountant, even though we lack a precise understanding of how they help us. Moreover, just a cursory look at research on the details of such processes reveals their rarity, with law and medicine considered the two key cases still today, and the rise of psychiatry and psychology in mental problems, a century ago, constituting one of the more recent groups of broad significance. Finally, we are also reminded that psychiatrists first had to persuade their patients to trust medical competence over theological explanations, with which priests had previously claimed authority over such issues.

Like knowledge of personal problems back then, data expertise has long been defined. Complicating the integrated image of data science further, separate institutional contexts have provided their own definitions of this work, ranging from government census, to insurance firms and academic disciplines.

Taken together, it is not only that the data science title is confusing, or that the processes data science undergoes are generally unlikely. The quantitative context has alternatives in place already. There is much more detail to all these aspects, which the next section introduces as it further specifies data science's challenge to them. Here I focus on the practical reasons why it is worthwhile to consider data science, aside from its genuinely puzzling emergence.

I am going to argue that data science bears public and private significance as it shapes social life and reveals a new model of work. The process of it gaining salience offers an analytical avenue to address both aspects. We can just turn to the public commentariat on data science in order to recognize its relevance to many modern social and economic activities. The behavioral consequences data science expertise is regularly applied to induce, often give reason for public concerns. By considering how data science defines its expertise, we can arrive at conclusions regarding data science's coordination, and hence ways to engage with the consequences that concern us. Second, leaving aside the question of whether data science is good or bad in these respects, I also argue that by defining such public salience, data science recovers a much older tension between the bureaucratic division of labor and autonomous work. The key implication that follows from this tension pertains to the struggle of defining a distinct expert group amid historical conditions hostile to it, and on the basis of technological resources its members could not have anticipated. Indeed, throughout the several decades computer technology has been available for, it has served to reinforce bureaucratic structures. Even with the rise of the Internet, alternative models of work have remained arcane.

1.1 What is data science?

> Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,** which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics,* similar to Knowledge Discovery in Databases (KDD).
>
> wikipedia.org

This definition is from the Wikipedia entry on data science. Even though Wikipedia may not seem like the authoritative source of choice, there are several good reasons for considering this definition, besides its convenience. For once, more established encyclopedias do not list it.[1] Analyses have also shown that errors in Wikipedia entries are about as few as in established publications (Giles 2005). Most importantly with respect to the puzzle of data science's lay salience, the Wikipedia entry would be among the first returns Google offers to those seeking to understand data science. As many might see this, it is pertinent to ask what it tells us.

As the name suggests, data science is concerned with the analysis of different kinds of quantitative data in order to learn from it. The devil is in the details, however. The definition enumerates several obscure terms, some referring to specific techniques, others to existing areas of scientific expertise. The terms and fields underlined in the quote have their own entries. For instance, the definition describes data science's goal as the extraction of knowledge and insights. While "insights" are supposedly obvious, and do not require further clarification from another entry, knowledge has a link to its own entry that introduces the philosophical and epistemological debate associated with it. We could read this to the effect that some of data science's aims are common, while others are arcane. At the same time, the idea of knowledge is commonly used as well. The same holds for data, which has nonetheless a link to an entry that describes it as a set of values, their arrangement, and several other specifics. If we are willing to overlook some details, we make some progress toward understanding data. But we also immediately see that this is not all there is for understanding data science.

The other terms are more arcane and technical. When we look just briefly into definitions data science draws on, we learn that statistics refers to the scientific discipline of analyzing quantitative data. Data mining has the same goal, but is associated with computer science. The concept of "predictive analytics" has a more specific meaning of making predictions of unknown future events, in particular for

[1] Take for example the *New Oxford American Dictionary* or the *Oxford English Dictionary*, as of July, 2016. Even Wikipedia, which accepts contributions at any time, only saw the first entry for data science on August 27, 2012.

commercial applications. It is not immediately clear whether its explicit listing implies that there are other ways of analyzing data that may not be relevant here. Similarly, “knowledge discovery in databases” links to a subsection in the entry for data mining that describes a research area. The footnotes, with the stars indicating their position in the quote, link to external websites. The first two link to an entry in the *Communications of the ACM*,[2] and a blog of few academics (first and second star). The third links to the website of a professional degree program at Northwestern University (the third star). Without studying these fields carefully, what can we take away from this for understanding data science?

While the definition seems helpful at first, it does not tell us much on its own. On a high level, we learn that data science is about the analysis of quantitative data with significant reliance on computational power and with the aim of making predictions. Some components seem to be well institutionalized already. Other parts of it are still in the process of being defined. While it is not surprising that we are unable to understand the details of the services we receive—just think of medical or legal advice—this combination indexes a more unusual problem.[3]

This direction points toward different kinds of questions. We can just recall that on Wikipedia, where we found this definition, almost anyone can contribute. Some of the areas of expertise it draws on, on the contrary, have entries in established sources, which data science did not. This description, reflecting ideas but not activities, therefore begins to recover tensions. There is a camp of those who contribute to it, who find data science sufficiently relevant to collectively work out a public definition. And then there are those who provide the new directions, indicated by the external links, but also others who have worked out for decades, if not more, some of the basis of what data science claims as well.

Several questions follow. Why would one add another layer of complexity to these questions? What are those saying who define the older fields, which others, who define data science, then draw on? How would data scientists define it themselves, and how relevant is this to their work? Even a careful reading of this entry and those it links to cannot address these questions. They are empirical. We need a clearer understanding of what data does, a way to think about it systematically, and then, most importantly, we

[2] The ACM abbreviation refers to the Association of Computing Machinery, and this is their monthly journal. Ceruzzi (2003) describes its role in the foundation of the computer science field during the 1960s.

[3] Even the questions that follow from this cursory reading of the entries index this problem. Why is it necessary to explicitly differentiate between structured and unstructured data? What is the difference between the two approaches to quantitative data analysis, as one is just called statistics and another associated with computer science? And why is data science similar to KDD, if KDD itself is a subentry of data mining, which itself is only a component of data science?

need to consider those who stitch the partly contradictory, partly just unlikely combinations together, and finally ask how the result looks from the perspective of the public settings data sciences pertains to.

1.2 Data science in public and private affairs

Data science has appeared in several contexts within just a few years, ranging from commerce, news and public administration, to political campaigns and criminal justice, as well as many technical fields. I here focus on a few highly visible moments of public concern, which also index data science's salience.

Instances in which others systematically impact our private affairs without a legitimate basis offer ample reason for concern. One such example that is often cited in the context of data science describes how analysts for a grocery store chain designed a statistical model to infer pregnancies among their customers on the basis of observed shopping behavior. The result of this could facilitate campaigns that target customers during a time that has been shown to bind them for years to come. In an evocative instance, one such campaign led a father to complain because he could not see any reason for why his daughter, a high school student, had received coupons for baby supplies. Yet it was the father who eventually apologized for his own complaints upon discovering that his daughter was indeed expecting a baby and had concealed this from him. Strikingly, the analysis inferred the pregnancy on the basis of a quantitative analysis of items pregnant shoppers tend to buy in general, and which are not explicitly related to newborns (Duhigg 2012). Other instances penetrate even subtler and no less personal affairs.

A highly popularized and polarized case has unfolded around Facebook's data science team. Together with academic researchers, this group conducted an experiment among Facebook users, who were unaware of their participation.[4] The purpose of this experiment was to understand the degree to which friends' emotions shape one's own emotions. In order to address this question, the data scientists filtered the information one group of Facebook users would see from their friend connections such that only either negative or positive messages would appear. Then the analysts considered sentiments the different groups would subsequently express in their own updates, and compare users with the different treatments. The researchers established the sentiment on the basis of the users' natural language

[4] Note that authors affiliated with Facebook emphasized that the privacy agreement users signed upon registration for membership was not violated.

(Kramer, Guillory, and Hancock 2014). In other words, quantitative methods here approximated and shaped a deeply personal matter.

Grocery store chains and Facebook are both broadly salient, but also known for seeking constant change. These instances therefore give little indication of data science's persistence. Public administration offers a more stable site, and it too has come to embrace data science. There are several data initiatives of large cities, which have for long kept records on their transportation systems, as well as commercial infrastructure and public services. In order to leverage all that data for the purpose of improving the administration's effectiveness, former New York City mayor Bloomberg, for instance, in 2013 instituted an office for data analytics (Smith 2015). Similarly, the White House has created and filled the position of a Chief Data Scientist in the Office of Science and Technology Policy, who is tasked to address problems ranging from open data to Precision Medicine. Plans for more such hires indicate the number of problems considered amenable to data science expertise (Muñoz, Smith, and Patil 2016). Finally, academia has shown growing interest in this new idea as well, as we see in the numerous programs and well-funded institutes set up to facilitate data science training and research (Miller 2013). Taken together, these instances illustrate some specific activities data science has raised concerns for, as well as evidence of its public relevance even where applications remain subtler.

Data science's public impact stretches beyond its popular and political applications. Data science has also raised attention with respect to its distinct substance and sheer magnitude, which is to indicate its application in areas where the basis for concern might not be immediately obvious. In 2011, for instance, the consulting firm McKinsey & Company published a much-noted study on the prospects of data for businesses. The report has received particular attention for its prediction of a shortage in analytical expertise, which it specified at between fifty and sixty percent in 2018 (Manyika et al. 2011). More important than the question whether the prediction is accurate or not is the prominent recognition of this arcane field of expertise. Moreover, the vague specification of the problem underlying this estimate limits its relevance. To cite just one highly visible instance, in a Harvard Business Review article from the following year a practicing data scientist described the field and defined the kind of role he has as "hacker, analyst, communicator, and trusted adviser" (Davenport and Patil 2012). A number of books have followed this early definition. They detail the underlying expertise and substantive applications in

more or less technical aspects.[5] Since that time, newspaper articles cite data scientists without further introducing their expertise in reports both on instances of the work these nerds do as well as on their privileged role in today's industries (Hardy 2016, Widdicombe 2014). In short, data science offers reasons for lay concern as it shapes public and private life directly, and because of efforts supporting its spread.

These instances could be read to suggest that data science's salience is not surprising after all. If all these areas amass data and apply it for purposes of shaping our lives, it is obvious that data science will be highly visible. At the same time, such an interpretation ignores the much more obvious alternative of defining those novel tasks in the existing occupations and organizational functions already close to them. I noted above the institutional segregation of different quantitative tasks that has so far persisted for more than a century. Even if data is available today in a greater range of sectors, there is no immediate reason for why such a division should not prevail. It follows that in order to understand the substance of data science as a public concern, we need to understand the basis on which data scientists define their tasks such that they gain salience as a distinct group.

I now argue that for that purpose, we need to reconsider our model of work that has been shaped by ongoing institutional and bureaucratic compartmentalization and subsequent specialization of tasks.[6] While the technological advances over the last few decades have produced the basis for alternative tasks and task arrangements, their esoteric status has prevented them from substantiating publicly salient models of work. Against this backdrop, data science's salience provides the grounds for a new model. Understanding the basis of its salience therefore simultaneously addresses two sides associated with modern technology: A source of public concern and a moment of individual opportunities.

1.3 Three models of work: Labor managers, technology hackers, and data scientists

We can summarize so far that in addition to data science's puzzling emergence, it is at least at times associated with worrisome consequences. Here I argue that data science also constitutes a tension between public concerns for privacy and opportunities in novel ways of defining autonomous work. In order to consider data science's historical significance as a way of organizing work, we need to take into

[5] There are many examples of this, of which I just list a few: Schutt and O'Neil (2013), Janssens (2014), Foreman (2014), Shan et al. (2015).
[6] Here I focus on their applied work, because this has raised most attention. I also consider the scientific setting in the next chapter.

account the condition of the bureaucratic division of labor, within which data science forms. Second, we need to consider the status and fate of other technology nerds. Some we find by following the bureaucratic tasks themselves, whereas for others we follow data scientists' interpretation of their own roots in the work of other nerds. These groups have undermined corporate dominance and bureaucratic compartmentalization. Contrary to data science, they have not defined public salience. This review reveals data science's relevance for revising the longstanding bureaucratic definition of tasks.

1.3.1 Corporate compartmentalization of work

Concerns over worsening jobs and working conditions for Americans fill many grim accounts. C. Wright Mills observed the degrading quality of white-collar work early on and argued in ways that make modern technology nerds directly relevant for addressing this process today. Here we need to consider two main arguments: the loss of the idea of autonomous work among those who seek employment and the removal of a basis for emotional connections to one's work (Mills 1951). These losses, in Mills's view and words, at least in part follow from the displacement of the old middle class of farmers and entrepreneurs through large bureaucracies. This process entails the demise of role models for younger generations and the compartmentalization of tasks themselves, which makes role models irrelevant. Data science's salience and technical basis address these losses.

To be sure, the main groups Mills had in mind have little substance in common with data science. The relevant processes are not tied to their specific tasks, however. Mills also notes how bureaucracies sidelined the autonomy of the learned occupations, specifically the medical and legal profession. Besides old entrepreneurs, they offered a model for careers in autonomous work as well, which corporate executives, engineers and businessmen could not provide (Mills 1951, ch.6). While data scientists neither heal diseases nor draft contracts—though their skills apply there as well—, they too rely on arcane expertise. Yet, Mills's observations of the increasing bureaucratic definitions of work in the previously free professions challenge even critical comments on the growth of capitalistic enterprises since the times of Marx and Engels, where the autonomy of lawyers has been seen as a puzzling mainstay in the division of labor (Merton 1968b).[7] The opposing interpretations index the magnitude of the shift that has followed

[7] Mills acknowledges that if anyone has managed to remain free, it is "minuscule groups of privileged professionals and intellectuals" (Mills 1951, 224).

from the spread of great bureaucracies and their principles of work.[8] At the same time, the comparison to the learned professions signals more clearly the relevance of data scientists to this debate.

Technology connects today's data science nerds with this classic view on the bureaucratic division of labor as well. Since its introduction, modern technology has reinforced the compartmentalization of work. Mills, writing in 1951, associated the qualitative change through the bureaucratic control of work directly with the spread of early office technology, remarking that "[w]e cannot fruitfully compare the psychological condition of … the old-fashioned bookkeeper with the IBM machine attendant" (Mills 1951, 227-8). IBM machines are different from modern computers, which are central to data science. This difference could therefore lead to question the significance of Mills's view today. Yet, the continuous compartmentalization of work has also been documented for the modern field of software programmers, where early generations already lost their autonomy to management (e.g., Weber 2004, Kraft 1977). A trend that deprives middle class work of its autonomy has thus begun with the growth of bureaucracies, and has so far continued with the spread of modern technology.

Computers do not reinforce compartmentalization by design, and there is evidence that this development could have unfolded differently. The early days of modern computers saw few formal experts but much lay interest (Ceruzzi 2003). Yet, even before the corporations so familiar today dominated the industry, the collective and informal engagement was quickly undercut on the basis of private property claims in software development. The lost historical opportunity of reforming the bureaucratic division of labor is powerfully illustrated in an infamous letter by Bill Gates where he asked "hobbyists" to stop sharing proprietary software. The link of the technology context to the much earlier argument on the loss of free work can be seen in that Gates's rhetoric directly invokes Mills's observation that "craftsmanship has largely been trivialized into 'hobbies'" (Mills 1951, 224). From the perspective of Mills's conceptual framework, the early spread of personal computers therefore reveals a foregone opportunity to recover the two great losses of bureaucratization; workers remain unable to build a connection to their work, and the absence of role models carrying images of autonomous work forward.

[8] Others have documented the absolute losses of professional autonomy as well (Larson 1977). The special interest and attention they receive in today's research strongly suggests, however, that professions have preserved their special status among modern occupations until today at least in relative terms (Gorman and Sandefur 2011).

This perspective thus links the public issue, which we have considered in the previous section, to individual models of work. As an occupational role gains lay salience, it becomes a personal direction for the constituents of that lay public. In Mills's time and view there was no such role left following the demise of the entrepreneur as a hero of the middle class and the bureaucratization of the learned occupations. Since then changing technology has brought new opportunities, but bureaucratic definitions persisted and reinforced compartmentalization of tasks. Instead of gaining broad appeal, the increasing specialization has favored the image of an engineer who "is part of inexorable science, and no economic hero" (Mills 1951, 22). Bureaucracies have trumped alternative models of work, but the early "hobbyists" have prevailed in the form of modern technology nerds. They constitute an important root of data science and reveal a critical tension between technology and bureaucracy, which I consider next.

1.3.2 Arcane insurgencies

The development of computers is all too quickly associated with the corporations that have put them into American offices and homes, whether it is IBM and DEC early on, or Microsoft and Apple today. While not inaccurate, focusing on just that history ignores all the computer nerds that have grown into a large community independent of these organizations. Such oversight comes easy as these nerds also often operate far away from the public eye, gaining lay attention in sporadic instances only. In other words, Bill Gates' letter did not stop the hobbyist effort, which has grown into the hacker movement that is, when visible, both celebrated and denounced today, although largely remains irrelevant. Hackers have left significant footprints in different ways by pursuing an open ideology so much at odds with the well-known corporations that dominate the industry. These nerds are diverse in their activities, but they can be broadly thought of as one type that challenges institutionalized arrangements directly, and another that competes with them by providing alternatives. Each entails models of work that differ from the bureaucratic compartmentalization considered above, yet neither has gained significant outside salience.

Activists

The first kind, and its different facets, can be summarized in the hacker Aaron Swartz. In one broadly recognized activity, Swartz gained much public attention for charges from federal prosecutors,

and his subsequent suicide, following the download of several thousand academic articles.[9] For that purpose he wrote a program that automatically downloaded the papers and hid a laptop, running it in a location with access to the database storing them. While illegal access to information in this way is often associated with hacking, several of Swartz's other initiatives are relevant as well and describe the type of work more comprehensively and appropriately (Coleman 2013).[10] For instance, Swartz contributed to developing the code underlying RSS language, which helps users to follow information from different online sources in a concise and standardized format. This standard is widely used today. Later on and in yet another project, Swartz used access through public libraries to download federal court records from a database that was otherwise charging for such access, although by distributing public records they had no copyright basis for those charges. Swartz made these records freely available, without legal backlash. His activism has also gained the attention of congress as he mobilized protests against the proposed Stop Online Piracy Act (SOPA) legislation.[11] After fruitless hearings of legal representatives from corporations such as Google and MasterCard, frustrated congressmen entertained the idea that it might be necessary to "bring the nerds in and get this right" because "[w]e are basically going to reconfigure the Internet and how it is going to work without bringing in the nerds, without bringing in the doctors" (United States. Cong 2011b). In addition to the ideology of access and openness, the examples illustrate technical principles of this community, which emphasizes improvising in the otherwise highly specialized technological context. This group embraces unconventional solutions, "hacks," as a means for solving practical problems. While these examples represent sporadic instances, others are systematic.

Organizers

Against Gates's trivialization of early computer nerds as "hobbyists" and denouncement of their ability to design stable software, the movement has grown to reach a striking scale and systematic infrastructure. The free and open operating system "Linux" is their signature achievement (Weber 2004, Kelty 2008). As a complete operating system, it directly competes with Microsoft's, and hence Gates's, "Windows" system. It started as a personal project in which Linus Torvalds, a Finnish university student at the time, tried to compensate for his lack of funds to pay for an operating system that would utilize the

[9] These activities and events provided the basis for a documentary on Swartz (Knappenberger 2014) that received prominent attention (Wu 2014).
[10] Indeed, hackers often just focus on legally legitimate work, referring to those peers who do not as "crackers."
[11] This bill aimed to prevent copyright infringement in the online setting (United States. Cong 2011a).

capacities of his new computer. Torvalds developed his own operating system instead and released the source code into the public domain. His effort met enthusiasm and quickly attracted support from a growing community. Within a few years the group has attracted thousands of contributors, spread out globally, who develop further applications and improve the code in a way that has created a robust software that is used widely today (Weber 2004). Contributors collaborate through a system of detecting problems and integrating new applications, rigorous checking of contributions, and distributed responsibilities for combining them with the overall infrastructure. Contrary to the type of hacking Swartz's actions represent, here we see a widely spread, coordinated and continuous effort of a community of nerds that directly compete with the dominance of bureaucratic organizations. Although they have a rudimentary formal system in place, specific tasks are freely chosen.

Aaron Swartz and Linus Torvalds both illustrate a powerful antithesis to Mills's observation, and Gates's vision, of the dominance of large bureaucracies and their continuously refining definition of tasks for workers in ways that deprive them of autonomous skills. They stand for a community of nerds, or hackers, who take great initiative with their work and form a close relationship to it (Coleman 2013). They nevertheless do highly specialized work, like their corporate peers. Because of the resulting arcane status, their working ethos remains with limited bearing on subsequent generations and all those not directly part of this community. With its salience, data science offers a sharp contrast, which I consider next.

1.3.3 Disciplined deviance

Data scientists directly invoke hacking as the basis of their skills, and sometimes introduce themselves as members of this movement. At the same time, there are several reasons to question the status of data scientists as hackers, both of the type we could see in Aaron Swartz, as well as in that of Linus Torvalds. In order to see the differences, we can just recall the public impact of data science, considered before. The image of a hacker who challenges large corporations as well as government legislation is clearly at odds with that of data scientists who apply their expertise in precisely these settings. Considering data science activities across all these areas, it is also difficult to recognize a formal system of collaboration of the kind we found is underlying the Linux development community. Yet, the claims are not entirely without substance either. Here it suffices to think of the fact that systems that were

designed to collect data, for instance in grocery store chains or at Facebook, neither resemble each other, nor directly connect to software designed to analyze such data. Therefore, these data science applications require unconventional solutions, typical of hackers of both kinds. Through data science's salience the similarities and differences reveal a new model of work amid bureaucratic hierarchies across which data scientists apply their expertise.

The similarities directly recover one of the two features of work Mills saw lost. As data scientists build their unconventional solutions, they engage in a type of work that has been associated with psychological, if not emotional attachment to these activities. This can be seen in one of the most surprising findings of anthropologists who have studied highly technical hacker communities to discover a casual and humorous engagement with the arcane problems hackers address, and deep personal attachment to them (Coleman 2013). This attachment stems from the tasks data science and hackers have in common, which suggests that they share this quality as well. In other words, six decades following Mills's eulogy on the entrepreneurial hero of the old middle class, data science nerds, at the forefront of modern technology, have prominently established themselves with at least one of those principles long thought lost.

That data scientists remain solidly embedded in large bureaucracies renders their project much less heroic compared to Aaron Swartz's activism or Linus Torvalds's entrepreneurship. It also seems more accessible and more consistent, however, and thereby invites once again to recall the learned occupations as a model for work.[12] Open projects like Linux remain exclusive to the extent that they operate through formally defined administrative mechanisms, hierarchical decision structures and boundaries, as the next chapter considers in more detail. These processes ensure the functionality of the overall project. They also reinforce its arcane status. Data science, on the other hand, while also excluding others on the basis of social processes, faces no technical requirements to impose such an infrastructure and attracts broad attention. Swartz's activism encounters no technical limitations either. Yet, it also lacks any systematic basis beyond the ideological connections between projects. There is therefore no infrastructure for a durable community. That we see data science spanning across different substantive and institutional contexts already suggests the presence of such a basis. Integrating expert

[12] This is not to say that either law or medicine is socially inclusive.

autonomy in a way that gains lay salience as a distinct group addresses the second aspect of free labor Mills thought lost, the role model for subsequent generations.

The comparison to the revolutionary activities of hackers could make the data science project appear dull and docile. I argue that such an interpretation overlooks the struggle entailed in defining a widely salient community of experts on the basis of arcane knowledge and across institutional boundaries. Tactics for overcoming such uncertainty might indeed benefit from such discipline. Data scientists neither resort to ideological nor to organizational grounds for devising principles of membership and coordination. Understanding the principles by which they resolve the uncertainty of defining autonomous work remains the chief task of the subsequent chapters.

1.4 Implications

What is at stake? Over the last few years a group of technology nerds, conceived of as “data scientists,” has gained substantial public salience in ways that are not well understood. The next chapter focuses on this puzzle directly. Here I have argued for two practical reasons to consider the data science case. I have shown instances in which data science work shapes social activity and inflicts upon privacy. While some of these activities benefit us, others raise concern. Without understanding data science’s salience across these cases, we cannot address our concerns comprehensively. Second, I have argued that data science’s salience recovers a model of organizing work that is inconsistent with the bureaucratic principles in which labor has been compartmentalized and detached from individual effort throughout the last century. Understanding data science’s salience therefore provides the basis for engaging with its concerning consequences as well as with its appealing opportunities.

First, data science impacts lay life in overt and subtle ways. Most instances that impact our lives in concerning ways are associated with clearly defined actors, most often organizations. Here we can just think of car safety or environmental pollution. Because organizations define the basis of those consequences, they can be held responsible. This also applies to some of the instances we have seen in the context of data science, such as Facebook, the grocery store chain and public administration. Contrary to other cases of public concern, data science has consistently appeared across these actors, and many others. This suggests that at least part of the basis for these consequences is not defined by

the organizations themselves.[13] It follows that while we can still hold them accountable, this would not address the problem comprehensively. Instead we need to understand how data science defines its tasks and applications in order to devise ways for us to raise such concerns and for it to define them as part of this expertise as well. This requires understanding the basis of its distinctive salience.

Second, data science provides a new model for expert work. To be sure, data science is not reversing the decline of autonomous work. Many data scientists themselves are not as autonomous as conforming to a strict image of professional work would require them to be. Some commentators even predict the commodification of data science as a field. Moreover, the principles of data science will not translate into occupations lost in specialization already. In all these respects, data science offers few implications.

At the same time, we need to consider that the technological transformation data science is part of still unfolds. Understanding the principles which data science operates through might be relevant for younger generations, who search for paths for participating in this transformation. Historical instances such as law or medicine, which still provide the basis of modern models of expert work, conceal much of the substance that seems relevant in the definition of new models of work. Data science for the first time allows us to observe and experience such a process directly. It follows that considering the ways in which data science struggles with the uncertainty of its scope, substance and purpose reveals a model of defining work that bureaucratic processes have long suppressed.

Although these two reasons directly oppose one another, they raise a common question, which this thesis aims to address: How has data science gained lay salience on the basis of arcane knowledge? This question is relevant to those who seek autonomous work, as well as to those who aim to ensure such autonomy rests on the right reasons.

1.5 Conclusion

Data science changes how we live in overt as well as in subtle ways. We have seen this in those instances where quantitative data analysis has uncovered and shaped deeply private information and intimate behavior. These practices raise many ethical and moral questions. Their greatest sociological

[13] Engineering has codes of ethics for those purposes. Yet, engineers work on such specialized parts of projects with concerning consequences that the responsibility falls back to the larger organization. Data scientists work on much more compact problems. Data nerds indeed entertain the idea of a Hippocratic Oath, known from the medical profession.

significance emerges at a more abstract level, as seen in association with the role of the data scientist, which there was no definition of just a few years ago. Indeed, for most problems we face, we either know who to turn to for help, or quickly identify the villain causing them in order to organize a response. Data problems are different. Here we find a series of concerns, and reason for expecting more such concerns, that appear to be related with an expert role in arcane knowledge of quantitative data analysis, which used to be deeply embedded in different institutional and substantive contexts. It follows that in order to understand how these data nerds have gained salience, we need to consider how they define and interpret their expertise such that it pertains to so many different problems. In other words, gaining leverage on the practical problem of how data changes our lives requires that we address the analytical puzzle of how nerds, who have been largely invisible in the past, gain broad salience.

We have also considered individual opportunities as another reason for data science's significance. Whereas obscure and highly specialized forms of work seem mostly obvious today, C. Wright Mills recovers a time and process when salient forms of work guided aspirations in the division of labor and he notes their loss as a result of enhanced bureaucratic definitions of work. This shows the significance and magnitude of data science's salience from the perspective of a second reason. At the time, before the rise of bureaucracies, Mills saw such salience in both the entrepreneur as well as in the learned professions. Both lost it, in his account. The latter still offers analytical leverage here, as professions share with data science their reliance on, and control over, abstract knowledge. The question of the kind of division of labor that is associated with such stocks of knowledge therefore complements its consequential impact on lay life. Yet, both perspectives require that we understand the basis on which this community of data nerds defines its contours such that it gains lay salience.

Mills's historical observations map onto modern work. We could see their relevance in observations of the mostly obscure work in today's technological transformation. Bill Gates, Linus Torvalds and Aaron Swartz are all nerds that have gained individual prominence albeit representing or promoting radically different, in some ways opposing, definitions of work. Here we can just recall how Bill Gates' letter directly attacked the kind of work both Torvalds and Swartz stand for. It was also clear from the kind of projects which they have gained recognition for that their work unfolds differently. Moreover, their personal prominence should not be read to suggest that individual tasks are salient in the work they stand for. It is

rather the opposite, that they represent the top of the career chains that remain invisible otherwise, and, as Mills bemoans, therefore offer no guidance for pursuing them. In short, the products gain prominence, not the tech nerds and contributors.

These accounts cannot directly lead us to consider data science. Times have changed since Mills's dark account. Data science is salient as an expert group doing in many ways similar work as all those who gain no salience working for Bill Gates, contributing to Linus Torvalds's Linux project, or doing work like Aaron Swartz. It follows that we need a more refined understanding of the kind of practices associated with technical expertise, as well as ways to study them systematically. Those I consider next.

Roadmap

First we need to put these images of technology nerds on a conceptual footing, and devise ways to measure them systematically. For this purpose, in the next section I introduce literature on quantification, professions and other experts, and some general principles underlying the thought communities they form. These principles rarely occur in ways that we can index easily. I therefore introduce a family of methodological strategies for recovering expert groups, data nerds or otherwise, from the ways in which they define and apply their stocks of abstract knowledge. With them we can consider empirical settings.

The conceptual complexity of expert work requires that settings allow for taking multiple perspectives. First, we study data nerds in New York City, and how they articulate their expert role in public. We approach this question in several moments, focusing on the associations of those expertise definitions with their structural and institutional context, ranging from technology and organizations to specific projects, skills they require, the community defining them and the principles by which the specialized definitions hang together. These moments represent a continuum of levels, beginning from its concrete technological footing to its abstract definition. Each step covers a moment in the transition from macro and meso-processes of modern technologies and organizational applications, to the micro level of skills, and back up again, to meso-level of community formation. Throughout we ask how much nerds in their accounts show evidence of contours of a distinct group of experts across these moments. This first analysis provides a rich image of the attempts data nerds make in order to define relevant expert knowledge. It does not give any direct indication of how these principles compare to other groups that

have gained lay salience as well. The focus on the abstract definition of arcane data science expertise extends to include two more studies that take a comparative perspective.

Taking a comparative perspective, next I consider the principles of knowledge application and definition across several salient and obscure cases over time and in economic and academic settings. Data science pertains to applied problems and academic work, each of which I consider in two separate studies. In chapter nine, the first one in part two, I analyze lay interpretations of data science expertise across different sectors in comparison to the definition of the autonomous and anonymous legal profession as well as bureaucratically defined occupations. In the academic context, which the second chapter of part two scrutinizes, I consider a community of scholars, who are seen to address data science problems with respect to how they define relevant knowledge. I compare this community to one that defines relevant legal problems today and a third one that has just begun to define a novel field of problems in the natural sciences.

I conclude with implications that follow from the understanding of data science as a thought community. Here I address lay consequences, individual opportunities, and the sociological approach to understanding expert work in the technological transformation following the ongoing rise of the Internet, digitation and mobile devices and services.

2 Quantitative problems, expert knowledge, and the struggle with uncertainty

In the following sections, I review literature that addresses the basis of data science's arcane expertise and the processes by which it could have gained lay salience. This review leads to a key argument that follows two shifts of perspectives from reconciling distinct debates in the existing literature in the context of the modern technological transformation. First, I propose research on expert groups to focus on thought communities[14] and their public recognition instead of institutionalization and state certification as relevant outcomes. Second, I suggest that we gain analytical leverage when we move toward considering stocks of knowledge and the principles underlying their construction as relevant explicans.

This argument follows from a series of observations. To begin with, the significant research on the components of data science shows no indication on how it comes that these components have become considered as part of a novel expert role. Clear guidance further suffers from a dispute in the literature on familiar expert groups, where one side emphasizes formal process and the other informal processes. A third position offers a framework that is able to reconcile the previous opposition as it recovers arcane and mundane thought communities and their distinct and often anonymous coordination processes that operate through defining shared views of relevant and irrelevant problems, ideas and histories.

On the basis of these accounts I design a strategy for studying data science directly. I propose a qualitative design in order to capture data science's struggle of defining the contours of a novel thought community. Moreover, I devise a set of formal methods that combine the established conceptual ideas from the following sections with modern tools I introduce in the next chapter in order to compare the principles of data science expertise to those of salient and irrelevant stocks of abstract knowledge. Finally, I argue that data science's emergent status, together with its specific substance, offers a rare view into moments of systematic uncertainty. Data science enables us to study how actors coordinate their application of arcane knowledge to lay problems when the relevant skills are not yet clearly defined. Analyzing data science in this way, I suggest, reveals the processes by which data nerds navigate the

[14] Let me stress the focus on a "thought" community, as it sets this focus apart from early views that considered professions as communities with shared values (Goode 1957), which subsequent empirical analyses could not find support for (Larson 1977, Heinz and Laumann 1982). Understanding expert groups as thought communities as I define more specifically below does not share the earlier view.

struggle that is part of defining separate areas of arcane expertise as a distinct stock of knowledge that is widely recognized as theirs. Contrary to historical perspectives, considering a case as it undergoes this process captures many of the emotional moments that are part of such developmental transitions.

2.1 Numbers, algorithms and computers in professional and private life

Whereas the introduction has focused on the practical problems and opportunities data science confronts us with, let me now move on to the analytical puzzle: Data science appears novel, although it draws on many old and established ideas, and salient, despite these ideas' arcane status. For designing strategies that resolve this puzzle we have to bear in mind that sociology rarely faces social objects as they emerge.[15] It is more common that significant contributions reveal how problems we already recognize for their importance unfold differently from the way we thought they did. The importance of finding a job was clear before Granovetter (1974) pointed out the role of weak ties in that process, as was school and work attainment before Blau and Duncan (1967) revealed its relation to parent positions. In some respects, data science is familiar as well, and thus could be approached from this direction that focuses on the case directly.

Experts have begun developing methodologies for and conducting systematic quantitative analyses over a century ago, and shaped modern societies that way since that time. For instance, the American constitution already defined the collection of systematic census data, expanding to include economic activity in 1850, albeit without systematic design improvements until the Congressional mandate for its own staff and support from statisticians in 1902 (Conk 1980). Economic enterprises have also relied on systematic quantification at least since the institutionalization of bookkeeping in the middle ages (Carruthers and Espeland 1991) and even more profoundly since the expansion of finance (Muniesa 2014). Finally, statistics and computational analyses have, thanks to Alan Turing, famously shortened World War 2 as they helped to decipher Nazi communication and conceal this achievement in order to preserve its utility. These cursory historical markers already begin to indicate the breadth and depth in

[15] There are examples for sure, although most take historical perspectives. Below we will consider research on the formation of psychiatry with some detail. Another prominent case is economics (Fourcade 2009, 2006), and Collins (2000) studied the formation of philosophical schools. One could also turn to other substantive areas, such as organizations, states or people. Whereas ideas from the origin of states are useful and will be considered below, the other instances emerge in institutionalized patterns. It follows that salience, which is puzzling in data science, is part of the scripted process. The same holds to a lesser degree for state formation, where lay salience is institutionalized in security and taxation. An important and non-historical instance has been the transgender child (Meadow 2011).

which quantification has penetrated our society until today (Porter 1995). We also see that social science research has studied these processes and some of their consequences with great detail and rigor, which I turn to below. Meanwhile, this set of historical accounts leaves no clear indication of how the activities they document could feature jointly in discussions around the problems data science is seen to apply to. The current state of this literature therefore rather enhances than resolves the initial puzzle.

The heterogeneity of these empirical problems complicates identifying and navigating the relevant social scientific literature that pertains to data science. To be sure, there is a considerable and rapidly growing amount of writing directly on data science. Two main groups contribute to that literature, which describes some of its technical basis, background, purpose and prospects. Data scientists themselves write much to communicate their work and purpose, as we have considered above. There are also more popular perspectives, which nonetheless offer comprehensive overviews and arguments (e.g., Baker 2008, Ayres 2008, Pariser 2011). Despite the inside perspective many of these contributions have in common, and which is often reflexive and critical, the impression management role these internal accounts play rule most of them out for a discussion here. They do offer important perspectives for us to consider as evidence in the empirical analyses.

Next, there is no social scientific literature on data science directly. This is to say that we have no account of data science's formation or consequences that this study could extend, specify or revise. There is no direct basis for a conceptual and methodological design either. The most appropriate point of departure therefore seems to be those accounts that I have introduced already, which describe its roots in older analytical and technical fields. Research has addressed these constituting areas, albeit in separate accounts.

Similar to how data science is seen to address problems in heterogeneous substantive areas, a range of separate intellectual contexts have defined the expertise it draws on for those applications. Since the combination of the data science label begins with data, the substance of this area work, I begin with literature on data here and turn to science, or the practice of working with this substance, in the next section. Ubiquitous data collection easily seems obvious today, but research on the early spread of quantification has discovered important variation across the key drivers of this development, which include state institutions with interests in taxation, but also engineering projects, and quantification for

insurance purposes (Porter 1986, 1995). With these foundations in place, statistics constitutes the most widely discussed component of data science. Its origins are well documented as well. The pure, mathematically shaped appearance statistics has in modern academia roots in external political currents, specifically the eugenics movement at the turn to the twentieth century, as influences on the construction of analytical techniques (MacKenzie 1978, MacKenzie 1981) and the effect of statistics on the economy today (Didier 2007). Meanwhile, data science's key promise is not to rely on statistics alone. Turning to the introduction of computational power, the work on artificial intelligence maps out some important ideas. One key argument it makes is that significant parts of our knowledge only translate imperfectly into computer code. It therefore proposes that the impression of more human-like machines may result from a more computer-like understanding of humans (Collins 1992). Quantitative analyses have come to replace the technology this research had considered, however, at least in problems that relate to data science. Applications that were impossible then are easy today. Despite their diversity, these different views provide a rich basis for beginning to consider data science and underscore the question of how modern experts integrate these diverse roots.

From a technical perspective, the concept of "algorithms" enters many of these arguments in one way or another. Algorithms play a central role in modern social life (Healy 2015) although their bearing has been most clearly recognized in financial transactions (e.g., MacKenzie 2014, 2016). How can we understand algorithms in order to address sociological problems? In these instances, algorithm refers to a "computer program running on a physical machine" (MacKenzie 2016, 4). In finance, these computer programs engage in financial interactions. They do so by evaluating information of offers and bids for investment instruments in order to predict price changes and to place their own bids. In the social context, in one ubiquitous application these programs evaluate the relevance of different websites or user profiles in order to rank them as search results (Healy 2015). Both instances demonstrate that algorithms cannot be viewed as shorthand for a kind of technology. MacKenzie (2016) explicitly bases his definition of algorithms on that of his interviewees, and Healy (2015) cites papers that specify algorithms and publicized them. In other words, this literature suggests that while algorithms in this modern context refer to implementations of abstract sets of rules in computer programs, they are not meaningful independent of the context in which they operate.

Two implications follow. On the one hand, when data scientists speak of algorithms we need to pay close attention to descriptions of the problems these algorithms address and the means of doing so, that is, the ideas encoded in computer problems. On the other hand, where these explanations are missing although data nerds speak of their algorithms, we need to consider the possibility of either strategic or naive references in that it may serve as a tactic to disguise specific strategies, or signal incompetence of what an algorithm does. This understanding of algorithms in some ways departs from much of the attention they receive today as a consequence of social biases encoded in them (e.g., Crawford 2016). My focus is on understanding what algorithms mean for those who use and create them, in order to then derive implications for better addressing the consequences.

Besides these technical conjectures, this literature has organizational implications as well. One important difference of all these areas and data science is their distance to users and clients. This difference is less profound in the final component of data science I consider here, where experts contribute to software tools. Because software is relatively easy to change after it has been produced, clients and users can be considered in the production processes more directly than is possible with other products (Neff and Stark 2004). Whereas these interactions pertain to the entire organization of software production with its non-technical aspects, data science at least claims to contribute in more specific and distinct ways. Studies of software engineering projects specifically have found, for example, that managerial roles in open source projects result more from structural positions than substantive contributions (Ferraro and O'Mahony 2012). While their organization is clearly distinct, it also unfolds without client problems. In other words, data science not only integrates areas considered separate before, it also connects them to lay problems more directly than previous instances. These aspects enhance the puzzle of these applications and the consequential salience compared to those groups that have kept those problems isolated.

The details that lead us to better understand data science's roots simultaneously take us further away from understanding data science itself. We learn from this research that contextual factors matter even for analytical procedures, and that social interactions shape technical projects amid abundant formal features and processes. None of these studies directly pertains to data science as neither developing their own methods is central to their work, the process of which we could understand from research on

statistics, nor do data scientists collaborate around larger software projects or other clearly defined technology efforts, as seen among software engineers. This leaves us once again with no clear understanding of how these accounts of disparate origins and applications of technical knowledge might explain the salience of data science as a distinct expert role. If anything, the dominance of specific historical actors and applications in this literature enhances the grounds for surprise about data science's broadly visible appearance. Public salience has not been a concern of the literature so far.

Contrary to the compartmentalization of expert work all these diverse professional contexts indicate, we can find more likely grounds for the broad salience of quantitative experts when we turn to the lay public directly. Two important debates in this area focus on quantitative skills and knowledge in the general public and how individuals use it in their everyday lives (von Roten and de Roten 2013, Lave 1988, Callon 2008, Vollmer, Mennicken, and Preda 2009). The first question, of statistical awareness in society, addresses an important problem in the context of increasing quantification. This literature finds, for instances, that 66 percent of Americans understand the concept of the mean and that lower shares of college students correctly read different visual presentations of statistics (von Roten and de Roten 2013). In other words, one of the most basic statistical ideas is relatively widely familiar, while competencies in interpreting even visual representations of results show severe limitations. The second debate considers how people use quantitative strategies. Callon (2008), for example, describes the specification of products as countable features, and outlines implications for consumption behavior. Complementarily, Lave (1988) describes in minuscule detail the specific calculative strategies individuals use in purchasing decisions and other activities based on quantifiable components. Together this research shows how much and the ways in which quantification shapes our everyday lives, and thus leads to important implications for data science's lay salience.

One immediate conclusion that could follow from the quantitative awareness documented here suggests that it makes recognizing new quantitative experts more likely because the public knows that quantitative problems are abundant and thus obviously require experts to deal with them. Even if that is the case and the lay public is more likely to direct its attention to a data science role, it is not for the public to define the role initially. The literature considered just before enhances the puzzle of data scientists' salience further as it recalls the generations of quantitative experts that were not recognized for those

purposes.[16] Nevertheless, this literature gives some reason for which we can expect at least receptiveness for the data scientists, once they have articulated their competencies. The literature could also benefit from understanding the ways in which data scientists define their expertise as their central promise, at least implicitly, aims to conceal the quantitative analysis tasks from the public more profoundly.

On a final note it is important to keep in mind that all this literature predates some major technological developments that define the context of data science. Over the past few years, faster and cheaper computation technology has become available and has diffused into a continuously and rapidly rising number of industrial and personal devices (Smith 2015). The enhanced ubiquity and availability of data and analytical power that follows from this trend, and which data science is associated with, constitutes just one set of questions and problems among many that range from the use of smart devices in our homes to modern living and working conditions in smart cities. Quantification offers a useful lens into these broader technological changes and their implications and consequences as it pertains to many aspects of it, even where it is not of central concern. By understanding how data science constitutes an expert role that is comfortable in this modern setting and reconciling its challenges with much older knowledge, we may be able to gain a better understanding of ways of organizing work where bureaucracies are not the primary coordinating mechanisms.

The literature considered so far historically contextualizes the problems data science is seen to address today. This context has been characterized by disparate professional activities in the past and thus denies the question of how data science is different from what we thought it was. An alternative strategy leads us to turn to cases that may be substantively different from data science, but have the advantage of resembling it in that they have also formed a distinct professional role on the basis of heterogeneous problems.

I consider such cases and findings from studies of them next.

[16] Von Roten and de Roten (2013) cite concerns of the British statistician regarding their low standing in the public eye, indicating that not only the knowledge data science draws on has existed for long, but also the struggle between its arcane status and lay recognition.

2.2 Approaches to complex problems

This review so far underscores the initial puzzle of data science defining a novel combination of problems as its own. We could see that at least its key promises articulate technical and substantive problems differently to how research has found them unfold in the past. It follows that this research offers insufficient direction for explaining how data science has emerged to be seen as a distinct expert role. Their position in the division of labor indicates an alternative basis. Because data science is seen to help others find and address arcane problems, the literature on expert groups promises some directions. I consider those here.

How data science fits this literature can be seen when considering how data science presents itself and how it is recognized more broadly. Especially the schools and institutes created around data science remind of other professions with distinct university training, such as in law and medical schools. This similarity directly invokes literature on professions. In addition to the canonical legal and medical case, studies have also described the more recent formation of psychiatrist as an expert group that, in the context of massive urbanization a century ago, has taken the mentally ill away from the clergy who had offered uncontested consultation before (Abbott 1988). At the same time data science differs in the strong relationships that persist between the expertise data scientists apply to practical problems and various institutionalized academic disciplines, as nerds apply experiences from arcane work to practical problems. The informal boundaries data science defines instead invoke research on expertise movements, where processes of formal institutionalization are much less central. One prominent case here has been the autism epidemic. Studies have revealed how an entire class of patients was discovered as a result of both institutional shifts but also the collaboration between medical experts and patient groups (Epstein 1995, 1996). The cases this literature focuses on, however, often unfold in more substantively constrained contexts than we see data science in, even at this early stage. Finally, the combination of broad scope and informal boundaries resonates with a third literature that is not primarily concerned with expert knowledge. It has instead defined a range of socially defined cognitive processes that give rise to "thought communities" broadly, including "churches, professions, political movements, generations, nations" (Zerubavel 1997, 9). While powerful in its ability to tease out consequential and widely diffused

cognitive forces, which need not be defined formally, this view quickly grows so all-encompassing as to not specify the processes by which data science gains more salience relative to other expert movements.

Although all three views promise useful ideas for understanding data science, their substantive as well as conceptual orientations vary too much as to consider them jointly. The first two sets of literature just considered, on professions and expertise, debate with each other over the most appropriate approach to understand expert work (Eyal 2013, Ben-David 1971). I discuss them together in order to preserve this tension. The third idea, of thought communities, does not directly engage with the others and I discuss it separately.

Data science fits oddly into the debate between the professions and expertise literatures. It shows few signs of institutionalization, which we associate with professional power. Its scope and magnitude simultaneously challenge the notions of informal groups and relations the expertise literature has discovered elsewhere. Besides discussing their respective positions on expert work and the contexts in which it unfolds in more detail, the following review shows that both camps emphasize knowledge production as a way of defining the respective groups each side focuses on. This leads to another view, which I discuss thereafter, of data science as a thought community. Such a framework is better able to identify markers of a cohesive group amid no clear indication of formal or informal processes. Moreover, whereas the professions and expertise debate index salience through institutionalization and recognition, the thought communities approach offers a more nuanced set of ways to define public salience. Here I begin with considering the literature on professions and occupations in order to derive the kind of processes underlying these groups. The review first revisits the main directions classic accounts proposed for these problems and then focuses on law and psychiatry as they have formed in conditions that seem most useful for understanding data science.

2.2.1 Occupations, professions and expert groups

One of the first major empirical studies of professions remains one of the most ambitious ones.[17] Its scope is telling. In their foundational account, Carr-Saunders and Wilson (1933) surveyed lawyers, patent agents, doctors, dentists, nurses, midwives, veterinary surgeons, pharmacists, opticians, masseurs and

[17] Professions have been of concern for many early sociologists, including Parsons (1939), Goode (1957, 1960, 1961), Wilensky (1964), Merton (1968b), and Hughes (1963). Those mostly theoretically grounded arguments have shaped the subsequent empirical approaches which I consider below.

biophysical assistants, the merchant navy, mine managers, engineers, chemists, physicists, architects, surveyors, land and estate agents, and auctioneers, accountants, actuaries, secretaries, public administration, teachers, journalists, authors and artists, brokers and, finally, a collection of occupations ranging from bankers to brewers.[18] Many of these professions still concern the sociological debate today.[19] In that early study, Carr-Saunders and Wilson covered all these groups on the basis of historical transitions they have undergone to then identify some general processes including task definitions, training, mutual oversight and discipline, the formation of associations, among others. Classic contributions also emphasize the community aspects of professions, defined by a sense of identity and shared values (Goode 1957) and esoteric services and knowledge (Hughes 1963). The major revisions to those initial studies have rejected the sense of identity-based communities and emphasized self-interest of individual members through status and financial rewards of professions (Larson 1977) as well as the monopoly and authority over problem areas they construct and defend against intruders on the basis of abstract knowledge (Freidson 1986, 2001, Abbott 1988).[20] This work, albeit not the most recent, shapes today's approaches most significantly. It has begun to present the formally defined professions in more nuances than one might think given the obvious way in which the lay public turns to them with some of their most intimate problems.[21]

Professions have undergone major transitions in terms of their substance and image until they have taken the shape so familiar today. The richness and diversity of this debate, and its tensions, specify and index these processes, ranging from high-status jobs to relentless competition. We also see that the early work already covered all the major occupations, promising little direction for the processes that

[18] Carr-Saunders and Wilson did not systematically delineate professions from non-professions (1933, 284), but doing so became a major focus later on (Etzioni 1969). Occupations, often defined by a lack of the autonomy professions secure for themselves (see Gorman and Sandefur (2011) for an extensive comparison), have become part of an entirely different branch or research. This research partly stems from the attainment literature and particularly focuses on the role of different occupations in that process. These analyses often focus on census or survey data, which, by definition, fails to address questions of data science's salience. Since it is only gaining salience now, it is not yet part of such surveys.

[19] Law will be described in more detail and with appropriate references below, as will recent work of engineers. Through Freidson's work, the medical case has defined the modern debate of professional work (Freidson 1960, 1961, 1988), most recently revisited with new questions by Menchik (2014). Civil servants, especially in their increasingly international capacities, are broadly studied (Mudge and Vauchez 2012, Seabrooke and Tsingou 2014), as are architects (Kreiner 2012) and accountants (Gill 2009). Physicians and chemists appear at least indirectly in the rapidly growing body of work on the sciences (e.g., Foster, Rzhetsky, and Evans 2015).

[20] Abbott (1988) defines the debate until today. I discuss his central ideas in more detail when I describe law and psychiatry. These ideas, however, have recently been directly challenged (Eyal 2013), as I review below.

[21] The now somewhat dated yet still recognized literature saw in the computer transformation significant promise for professional formation (Abbott 1988), but even early evidence documented the rule of management over technical expertise (Kraft 1977, Ensmenger 2010). This development directly resonates with the kind of industry Bill Gates sought to impose, as we considered in the introduction.

might explain the emergent data science case. Two prominent exceptions to this continuity can be found in psychiatrists and economists,[22] which in this capacity might shed some light into the formation of data science. I focus on psychiatry below for economists rely more on intermediary processes for implementing their expertise compared to data science.[23] Before considering this relatively juvenile case—at one hundred years of age—, I revisit some of the key ideas in this literature through the lens of law as a canonical baseline that, albeit having completed its emergence long ago, has established itself in the context of large bureaucracies, like data science.

Law

Neither of the two cases still most prominent today, law and medicine, would seem to offer an obvious model for data science. Law's path to prominence and autonomy, unlike that of medicine, is closely interwoven with the rise of large bureaucracies that followed industrialization. Thus, although substantively very different from data science, law shares with it a relatively similar structural position with respect to the types of clients. The relationship to corporate clients is repeatedly found to shape the legal field most significantly, as it responds to the different problems of large and small organizations, and individual clients (Heinz and Laumann 1982). Importantly, the segregation of legal services by individual and corporate clients follows from no practical limitations. Rather, the relationships to these clients have been found to define which subsets of services lawyers can legitimately offer (Phillips, Turco, and Zuckerman 2013). Furthermore, the influence of competition that the first wave of revisions of classic professions scholarship emphasized appears again in recent work that documents the effect of public rankings on the curriculum design of law schools (Espeland and Sauder 2007, Sauder and Espeland 2009). The key ideas of jurisdictional conflict (Abbott 1988), still shaping research on professions today, follow from an analysis of the legal field. In his seminal contribution, Abbott (1988) has shown how the defense against intruders led American lawyers to focus on some problems, such as drafting wills, while losing the administrative work thereof, but also internal processes such as the detrimental clerkship system that slowed the expansion of British solicitors. While their different substantive bases prevent us

[22] Whereas most professions face regulation on the level of the nation state, Fourcade (2006, 2009) demonstrates how economics has transcended that institutional level and shaped policies of nation states instead. This introduces a promising direction for research on professions, although at this early point in data science's emergence such a focus would be premature.
[23] This is not to say that data scientists offer the personal services psychiatrists are known for. Because data scientists design tools with which clients and users can interact directly, psychiatrists seem like the more relevant case.

from drawing conclusions from law for data science, the interactions between a group of arcane experts with clients, the market and each other all reveal processes an analysis of data science has to consider. In spite of its role in inspiring one of the most longstanding arguments in this research, the legal field has also provided grounds for a novel perspective critical thereof.

Recent work has shifted toward the specificities in which legal work and interactions unfold. In a study of courts, Sandefur (2015) demonstrates the importance of familiarity with procedures how to interact with judges over substantive legal knowledge. Other work reminds, however, that only a small fraction of lawyers work in capacities that involve the court, and that they instead most significantly impact society through the contracts they design (Howarth 2013). This direction resonates with Stinchcombe's (2001) little noticed but comprehensive analysis of the features and effects of formality in legal work. The key idea here, which is consistent with the sources of success in jurisdictional conflicts, is that law's central activities involve translating the substantive problems clients and the public present it with into formal legal knowledge. This explanation specifies the ubiquitously recognized importance of abstract knowledge in professional autonomy. With references to the ongoing revisions of arcane principles on the basis of concrete cases in appellate courts, this analysis clarifies that such abstraction is not to be confounded with theoretically defined knowledge. Despite their substantive differences, in the end we therefore still find several directions the legal case offers for considering data science.

These findings from the legal case invoke a more general consensus in the literature, and can be summarized in those terms. From early research we know of the decisive role arcane knowledge plays, and that professions apply it in order to gain autonomy over an area of practical problems. They control this autonomy through associations and formal education requirements, and, most significantly, through mutual oversight within these boundaries. We also learn that there is variation in arcane knowledge, seen in the abstraction from substance in legal expertise compared to theoretically defined scientific knowledge. Data science's emergent status shows few signs of the associations that facilitate self-control, whereas defining abstract knowledge would not require such channels. Meanwhile, law also shows limitations as a model for data science. Similar to data science, legal education in the US has started out outside of the traditional university context. Unlike data science expertise, however, legal knowledge has been defined independent of academic science to begin with (Stevens 1983). I turn to psychiatry next,

which resembles data science more in these respects and has also developed a strong autonomous basis of applied practice and expertise.

Psychiatry

Until a century and a half ago, when cities grew larger and more densely populated, those feeling disturbed by the greater proximity to others would still intuitively turn to their priests for relief from their mental distress. After all, who else could offer consultation? Today, it is obvious to see a psychiatrist (Abbott 1988). The questions pertaining to the origins of such a new field, as well as how they could persuade patients to turn away from the church for mental consultations have defined significant contributions to our understanding of professions (Abbott 1988, Ben-David and Collins 1966). This shift of responsibility over the personal problems jurisdiction originated in scientific ideas and careers developed in Germany (Ben-David and Collins 1966), which were also adopted in the US. In the US, a coalition of several research areas that ranged from psychiatry and psychology to neural studies, and in part just the specific scholars founding them, helped to construct a robust community that also gained the recognition from established medicine (Abbott 1988). Such a change in scientific thinking is unlikely to begin with, if we consider that disciplines tend to reward specialization in problems which scientists only rarely deviate from (Foster, Rzhetsky, and Evans 2015, Ben-David 1971). Finally, church and science represent one of the sharpest institutional oppositions shaping arcane knowledge today. For understanding the loss of influence of priests in this specific area of mental problems, however, Abbott (1988) suggests to move beyond those salient differences and consider the shape of theological and psychiatric knowledge directly. We then note that the latter interpreted all problems through their implications for the relation of the individual to God, a centralized system, whereas psychiatry allowed for separate mechanisms for different problems. Although psychiatrists treat individual patients whereas data scientists solve organizational problems, their formation during a relatively recent period that resembles more of our own time than that of the legal profession, for instance, suggests several processes to consider in the emergence of data science.

Psychiatry has highlighted some processes that differ from the broadly representative legal case. It reflects an instance where experts systematically circumvent the disciplinary mechanism that control the work in the science and thereby captures a decisive moment in the often latent but at the same time well

known tension between academic and applied knowledge (Abbott 1988). By documenting the intuition behind psychological mechanisms that accommodate distinct personal problems patients experience, standing in opposition to the centralized way of organizing problems around God, this case clarifies the variation in what we can consider abstract knowledge. It also reveals some social processes associated with this variation. Theology, science, psychiatry and the law are all abstract, yet whereas theology and the sciences construct their knowledge around doctrine and theories, psychiatry and the law[24] define knowledge on the basis of problems their patients and clients experience. For data science this suggests that we also focus on the concrete combinations in which data science defines relevant expertise and skills, such as centralized and heterogeneous-but-integrated arrangements, instead of remaining on the level of broad fields of knowledge it often invokes in general presentations.

Other research points out, however, that defining relevant knowledge often unfolds through local and informal processes. Findings in this area, which explicitly focuses on the formation of expertise instead of the formal groups claiming it, challenge accounts of professions in that they offer alternatives to the formal and institutionalized processes and structures. Implications of this research for understanding data science can be seen both in its basic consensus as well as from specific instances.

2.2.2 Expertise

Carr-Saunders and Wilson's list of professions has remained remarkably stable since their publication almost a century ago in the literature that followed it. Much expert knowledge has not. Today we treat diseases that were not even defined then (Eyal et al. 2010, Epstein 1996), our understanding of the environment has changed (Wynne 1992), and, looking from a different angle, it seems that important processes unfolding within the canonical professions are much less defined by their institutional scripts than previously accounted for (Sandefur 2015). Instead of entering public recognition through novel and monopolized responsibilities or mandates for treatment, these instances gain salience by redefining the view of existing problems.

As I indicated already, the autism epidemic has been studied as one of the most prominent cases. Here Eyal and colleagues (2010) have described in great detail the crucial role of alliances between parents of children with disorders and physicians seeking to treat them in more appropriate ways than the

[24] I strictly refer to the American legal system.

crumbling institutional matrix of mental retardation had to offer. Similarly, gaining recognition of the growing HIV+ population as a serious medical problem followed as the lay patients acquired medical expertise and thereby directly shaped expert discourse (Epstein 1994, 1996). Turning to different substantive settings, this research has also shown how local residents develop expertise over their environment which specialists could not derive from their formal training (Wynne 1992). Even the evidence for law in a certain class of cases shows that at least court decisions tend to be shaped by familiarity with informal court interactions rather than formal knowledge of legal proceedings (Sandefur 2015), already cited above. Furthermore, evidence from interviews for legal positions (Gorman 2005), could be interpreted to indicate that formal knowledge is not so important even in the cases for which it has been shown in the past.

These findings, from a different set of cases and with a different emphasis in familiar cases, provide a range of somewhat competing processes that could undergird data science. Whereas professions scholarship, as we saw in law and psychiatry, reveals various levels of formal interactions and sees the groups as a whole in conflict with organizations and other groups, the expertise research focuses on informal processes among the members directly. In one of the more programmatic arguments Collins and Evans (2002, 2007) synthesize this research to suggest that we need to reconsider the boundary between experts and the lay public. Instead of focusing on the role of clearly delineated groups, they argue that we understand the effect of technical and arcane knowledge on society better if we differentiate between “participatory” and “contributory” expertise, and analyze appropriate compositions of the two, regardless of the formal status of those who claim them. The following instances describe the implications for studying data science more clearly.

Medical expertise

Studies of medical doctors provided the basis for the modern study of professional experts (Becker et al. 1976, Freidson 1973). They also provide the basis on which Eyal (2013) develops expertise framework as an alternative to the earlier formal understanding of expert work. The basis for this claim comes from an extensive analysis of the rise of the autism epidemic in the US over the twentieth century (Eyal et al. 2010). Eyal et al.’s challenge to a formal understanding of expert work can be most immediately seen in the association of autism diagnoses with the deinstitutionalization of mental health in

the 1970s. In that older system children with certain behavioral patterns would be diagnosed with mental retardation. Although that classification and the system to take care of them changed, children with such behaviors did not. Parents were left alone to deal with these problems. Medical practitioners began working together with parents and developed a systematic understanding of their children's disorders they were observing and reporting. For instance, Eyal et al. describe the case of Bernard Rimland in particular detail, analyzing how Rimland worked with parents such that they jointly constructed new expertise through the combination of their experiences with autistic children and his arcane knowledge.

Although medical doctors still participate in this process in significant roles (Liu, King, and Bearman 2010), the ways in which they do so challenge a formal understanding of expert work in several ways. On one level, there is the institutionalized system that is not doing the formal work anymore, leaving parents with much uncertainty. On a subtler level, the medical practitioners who are still part of that system, operate in ways that reject an understanding of them drawing on authoritative abstract knowledge. This undermines the core idea of the study of professions. Eyal (2013) therefore suggests to not think of experts and their work, but instead of a more comprehensive understanding of expertise that does not assume formally designated experts as central actors. The differences between the two approaches can be seen along several dimensions, ranging from the increased scope of those who participate, such as the parents as opposed to just experts in the instance above, to a focus on the work involved in a problem, not the protection of it, and even including the objects involved in this process, not just the people.

The premises of this new approach apply to other cases as well. I have already pointed out law, but we can take public administration as well. The European Union, for instance, relies heavily on the informal connections in Brussels and a few other administrative centers (Mudge and Vauchez 2012). In other words, relevant expertise remains heavily contextualized rather than abstract and generally shared. Although these cases have demonstrated the promise of an expertise framing with particularity, its utility can be seen in technological settings that resemble data science more closely as well.

Technology and engineering

We can once again consider some of the key ideas of this approach in the context of specific cases. I consider one instance that unfolds entirely in an academic setting. I focus on sciences below, but

some of its lessons are useful here as well because it has taken an empirical focus on the collection and interpretation of data with respect to changing technologies, which is of course also relevant for modern data science. I then show that central processes that research revealed also facilitate applied work in the technology industry, which resemble the institutional setting we find data science in more closely.

As part of his comprehensive study of an academic research field that specializes in the detection of gravitational waves, Collins (1998) documents variation within this relatively small community of researchers with respect to how different groups interpret the results from the data they collect, and the conclusions they allow for.[25] An important finding in this research, and one I stress more below, is that the interpretation of the data was associated with the cultural context the different groups were embedded in. This interpretation was also associated with the technological setup. Gravitational waves physics relies on expensive and advanced technological setups in order to measure the empirical existence of their abstract claims. Substantive arguments are associated with the functionalities of the measurement technologies. Thinking once again about abstract knowledge, the core basis for professional autonomy, we find limitations in as far as the abstract ideas about how gravitational waves might work become encoded in technological designs of measurement devices, preliminary findings that may support one design over another, funding sources that support their ongoing development, and so on. Law in some sense struggled with a technology as well, specifically with that of bureaucratic administration. Because bureaucracies developed outside of the legal field, and to such great extent that they pertain to legal work in a relatively invariant form, they are not so salient as a feature of legal expertise. Devices for measuring gravitational waves are central to the debate and index variation within expertise of the field.

We can find technological effects closer to the institutional setting of data science in modern technology. Studying engineers and technicians, Bechky (2003b) has found significant informal activity along the process of designing and building machines. While the struggles over different responsibilities also invoke arguments of jurisdictional competition, it became clear that informal communication played a much more important role for completing the projects. This is consistent with other research on software development projects, which has found how experience with coordinating tasks trumps substantive contributions toward influencing further memberships and technical contributions (Ferraro and O'Mahony

[25] See also Collins (2004) and Collins and Evans (2007).

2012). Bechky also documents ways in which engineers identify with the objects they design, claiming ownership even once technicians take over the work (Bechky 2003a). This kind of identification resonates with the work of autonomous data nerds, and the informal relationship they build with their work and community (Coleman 2013). Instead of prolonged collaboration as in Bechky's case, Coleman observed the effect of passionate communication on the level of comparatively much more fleeting digital communication. At the same time, we have to recall that a distinction between bureaucratic and proprietary ownership of technology forms a sharp division with "hobbyist" understandings of technology work (Kelty 2008, Weber 2004). The focus on expertise beyond individual experts thus directs our attention toward the specific moments in which experts apply their esoteric knowledge, and how they apply it, even in the same substantive area of problems.

In the context of modern communication technologies, such informal coordination patterns more easily find their way into formal organizations with bureaucratic divisions of work. Stark (2009) documents in particular detail how the collaboration in the modern technology organization he studied differed from more familiar project work. Direct communication unfolded between employees of that firm with so different backgrounds that they could speak about the same problem with very different definitions in mind respectively. These different understandings, or evaluative principles, easily lead to conflicts. Unlike conventional understandings, these kinds of conflicts are productive because they uncover novel solutions. Consistent with Collins's analysis of technology, Stark predicts that "[c]ollaborative organization will continue to coevolve with interactive technologies" (Stark 2009, 117). Compared to the traditional hierarchical forms of organizing, which Stark's analysis is situated in, these processes differ to such an extent that we more appropriately think of them as "heterarchical" organizational forms.

With these more specific instances of the expertise research, we can now consider implications for approaching data science. For data science, this side of the debate suggests to focus on the interactions between leading data scientists themselves, as well as the informal strategies followed by members of the group more broadly, as they engage with their organizational environment. From this perspective, data science's salience as an expert group might not result from the formation of professional community defined by the formal knowledge its members share, but through the specific ideas of key actors who take advantage of the changing technological landscape. Moreover, as these broader changes toward more

powerful data processing technologies get applied in varying practical context, we should expect significant marks from these specialized experiences on the way data science experts define their expertise.

2.2.3 Sciences

The two sides of this literature on expert work reveal an opposition in their respective emphasis on formal and informal processes. The division between the professions view and the expertise view extends to the sciences and scientists. Whereas in the context of applied work the cases each side studies differ in a way that complicates direct comparisons, science provides a more coherent setting. I therefore consider them together in this section. Moreover, as the basic positions remain unchanged, here I describe the respective positions on the level of scientific work, instead of reiterating the broader views.

To recall, the main thesis of the institutionalist view holds that we can understand scientists if we understand the formal boundaries and rules of scientific work. One early and longstanding idea hypothesized the cumulative advantage of scientific work, suggesting that highly recognized scholars receive more recognition whereas less recognized scholars would not improve their status. In his argument, Merton (1968a) analyzes the role of highly productive scientists with a focus on the additional effect of formal recognition through a Nobel Prize. This recognition as implications for their collaborators in as far as the wider scientific community will be more likely to attribute the idea with the high-status author. Merton also analyzes the role of prominent scholars from a functional perspective. Here we can also note a critical difference to the other position I consider next. Merton attributes the significance of these scientists to the way they share norms and values, which can be routinized and institutionalized. He does not consider “techniques, methods, information and theories,” as distinct objects that are part of this process, as some of the arguments emphasize below (Merton 1968a, 159).

Science varies with the different roles scientists have in different countries (Ben-David 1971). We can see this, for instance, in the emergence of the statistics discipline. Universities in continental Europe required for each faculty to represent an entire subfield, which statistics was initially inadequate for. The less compartmentalized American departmental system allowed for increasingly promising applied mathematics to enter the fields benefiting from quantitative analyses. More recent work focuses specifically on scientists and finds, for instance, that scientists do follow institutional incentives, but that

different kinds of incentives divert their strategies in different ways (Evans 2010, Foster, Rzhetsky, and Evans 2015). This research has complicated the early ideas. The specific disciplines investigated under this view still reveal shared patterns of scientific work, for example that career concerns lead to more conservative research agendas, although fame follows from more creative tactics, and that commercial orientations lead to the evocative finding that scientists "know less about more." In other words, as the expertise view emphasizes as well, the institutionalist view also sees the boundaries blur between science and other fields. Findings, such as those from the biotech sector, nonetheless continue to support arguments that rest on formal features of the different settings that begin to overlap.

This largely formal view of sciences has an informal component as well. Scholars are accountable to their disciplines, not their formal organizations, the universities. Although they may be formally employed by them, their peers in the scientific field are more relevant for their careers (Abbott 2001). As the often formal scientific associations, this basis of their work is formal still, but also reveals a conflict with other forms of formal organization.

The other camp also focuses on the activities of the scientists, but doubts that all primarily try to find a shared truth and follow systematic patterns. Instead, this view holds that groups of scientists and other involved parties define the truth together and collaboratively devise research strategies that can legitimately reveal it. Whereas for the formalist side the scope of the sciences is clear, because it is formally defined, this is where the informal focus begins. Gieryn (1983), for instance, points out that scientists need to define the boundary between relevant work and irrelevant work in the first place.

The contextual and other factors differentiate between contributory and participatory expertise. The more formalized and institutionalist views assume that scientific fields grow independent of the content of their work. Collins' studies of how scientists interact about their research, even within the same specialty, shows, on the contrary, that it is not so clear whether they consider each other members of the same group, even if they are able to discuss some central problems (Collins and Evans 2007). In other words, formal status and affiliation have little bearing if the scientists are unable to the work without extensive collaboration. Another finding supporting this challenge to the formally-oriented sociology of sciences points out cultural differences in how laboratories interpret evidence of research in the same field (Collins 1998). These cultures overlap with the cultural differences of the countries these laboratories are located

in. While national differences are in principle consistent with the formal view of sciences, it would dismiss the relevance of these informal ways of discussing research results among direct colleagues. To consider one final example, we can turn to Latour's work on processes unfolding within a laboratory (Latour and Woolgar 1986). Here Latour shows how the production of facts, arguably the core efforts of science, unfolds in the histories and micro interactions between researchers in the laboratory setting. In other words, these findings stand in stark opposition to those arguments that emphasize inceptives and orientations that shape research across a field, considered above.

The opposition of these two camps has prevailed for decades (Freudenthal 1991), though recent work has begun to productively integrate some core ideas from each side (Shwed and Bearman 2010).[26] This work has begun to transcend the division by focusing on "issues," which are informal, rather than formal disciplinary boundaries, and yet measuring activities of scholars working on these issues formally and comparatively, rather than considering the informal features of collaborations. The data science case might benefit just from the possibility of such integration. It creates the additional challenge of pertaining both to the scientific and the applied context. I therefore consider a conceptual basis for such integration next.

There is no clear basis to expect more leverage from one side, compared to the other. Data science relies on mathematics, statistics, computer science and addresses problems of the state, firms, education, and so on. In other words, it is very likely that institutional forces shape data science work. At the same time, the promise of a distinct group of experts suggests that they deviate from those existing institutional patterns. The expertise literature has uncovered informal mechanisms to this effect.

2.2.4 Synthesis

The two sides of this literature on expert work define the opposition in their respective emphases on formal and informal processes. The inconsistent cases through which both camps develop their respective arguments render the contours of this division less clear. Data science moreover emerges in a setting neither side considers directly and thereby only allows us to drive hypotheses of underlying processes on the basis of analogies (Stinchcombe 1978). Indeed, those choices reflect the different times

[26] Kuhn's (Kuhn 1970) seminal work on the sciences can be interpreted as another intermediary view. Kuhn also focused on the construction of facts, instead of assuming general truth. Instead of scrutinizing informal interactions, however, Kuhn analyzed larger and more fleeting debates.

of writing.[27] They even invoke one of the findings in research on expert occupations itself, in that it "becomes apparent that the ability to obtain and maintain professional status is closely related to both concrete occupational strategies, as well as wider social forces and arrangements of power" (Klegon 1978, 259).[28] As with the vast digitization and related technological changes the "wider social forces" have shifted once again, it seems appropriate to consider approaches that are not strictly developed on the basis of the context that is changing here, which I do next. Meanwhile, and for that purpose, their respective contributions clearly offer possible ideas for understanding data science that operate independent of those varying contexts.

We have seen a transition here from formally bounded communities of experts to formal groups without significant community features and then toward expert groups without formal boundaries. Across these varying positions we could also see a consensus emerge on the importance of abstract knowledge. This consensus lacks a clear indication how to specify abstract knowledge, however, except for the basic opposition between centralized or specialized knowledge and knowledge that addresses heterogeneous problems without falling back on a central idea or theory. Finally, even the varying formal and informal processes consistently pertain to groups whose members control their own work. Data science's empirical context indicates no tendency toward any of these existing sides. Its computational basis resembles engineering and the informal coordination found in software projects, whereas the quantitative basis of its work is consistent with the ubiquitous spread across institutional and substantive areas, despite their origins in a small community of scholars.

[27] The great irony of this literature is that it studies other versions of itself, and sometimes also itself directly (e.g., Moody 2004, Abbott 2001, Stinchcombe 1997). From this perspective, it is helpful to define the gap which data science reveals in the literature with respect to the context in which it has remained open. The professions view is the oldest. One of its central premises is to distinguish professions from occupations (Sandefur and Gorman 2011). Some of the ideas associated with occupations could be relevant here to the extent to which data science is relevant for large organizations, which often dominate occupations. More generally, however the research on occupations often relies on consensus data for addressing questions of occupational effects on individual attainment (e.g., Weeden 2002). This sort of data clearly cannot address the learned professions (Freidson 1986), let alone the emergent side of the problem. Aside from this thriving branch between sociology and economics, the classic professions view lost traction for several empirical reasons, such as the lack of new professions and theoretical closure on the basis of persuasive arguments. One hopeful idea in this literature has been to turn to experts more broadly (e.g., Seabrooke 2014). The group I have discussed here with its expertise focus had also begun to pursue such a direction, mostly with the studies of scientific expertise, which I consider in the following section. It has also moved toward studying problems with direct effects on the lay public. Again considering context, this work seems to respond to times in which jurisdictional claims were settled and changing problems were addressed within this context. In other words, although the two sides often define an opposition, they address different empirical problems of different eras. The thought community view, which I introduce next, brackets the data science problem. It helps to define the puzzle around data science's salience as well as to shift the focus toward recognizing thought communities. Perhaps because it more directly relies on European ideas, or because it lacks a method or theory, it is relatively disconnected, overlooked or irrelevant (Zerubavel 1997, 46) from the professions literature as well as the expertise literature. I suggest to combine this literature with the work on professions.

[28] This reading of the different sides is further supported when considering that similar class to focus on expertise (Eyal 2013) have been made repeatedly, even in studies of more formal aspects of abstract knowledge (Klegon 1978).

Both sections considered so far, of quantification and of experts and their knowledge, leave a clear gap. Research on quantitative problems was seen to fall short of outlining the basis for an integrated data science role. Studies of expert groups have consistently revealed forms of abstract knowledge connect heterogeneous areas. Meanwhile it has shown disagreement on the social processes that coordinate the construction of such knowledge. This disagreement indicates that we need to consider the struggle underlying community formation directly and more broadly than the solidly institutionalized academic and professional settings considered here allow.

2.3 Thought communities and the public

Just like the many other instance in which research shows that we need to understand cases differently from how we did before, the sociology of expert groups has revisited the same set of cases. This first led to arguments for an understanding of functional and values-based groups, then structural and competition-based processes and most recently relational and informal processes. Data science resembles these cases as it defines arcane knowledge as its own. At the same time, it has given no indication of features expert groups usually have, such as formal boundaries or associations, or direct forms of coordinating. Moreover, while research has revealed significant variation in the problems expert groups treat, there has been much less change in the problem areas this research considers. Such a lack of variation with respect to empirical settings reduces opportunities for discerning patterns by which processes within them unfold. Finally, even where expert groups define new problem areas, not all expert groups gain public salience. All these limitations suggest that even enlarging the scope from the data and quantitative analysis jurisdiction toward expert groups more broadly might by itself not reveal sufficient variation as to identify process that might account for data science's salience.

At the same time, many instances of community formation on the basis of shared knowledge that have little to do with expert work have become obvious today. Here we can just think of kinship systems (Lévi-Strauss 1963), Catholicism (Brown 1982, 1992, Brown 2000, Brown 1981), capitalism (Weber 1988), or states (Anderson 1983, Wimmer 2013). While these cases outsize data science, they also widen the scope of processes that could contribute to its salience. I therefore also consider instance of this kind, although smaller ones, in addition to work on professional groups specifically.

2.3.1 Thought communities

Professions and experts are often considered exceptional for their esoteric knowledge. This focus offers analytical leverage relative to mundane occupations. At the same time, it constrains our view in other ways. Upon stepping back from the intricacies of whether expert work is governed by formal or informal processes, we begin to find community formation instances in many other settings that look arcane to outsiders as well and make processes salient that are not so easy to see among expert groups. Such a shift of perspective goes as far as juxtaposing surgeons with "Austrians and Indonesians, Mormons and Muslims … [and] … college graduates and high-school dropouts" (Zerubavel 1997, 11). Instead of considering membership in them as attributes of those members, as is often done, this approach focuses on the processes and features that define these groups in the first place. More generally, this enlarged scope defines communities as collectives of individuals who participate in several communities that do not entirely overlap. These multiple memberships define their individual perspective even compared to others that are also part of one specific community, but not another one. It does not assume that these community memberships have equal influence. Unlike the long dismissed definition of professions as value-based communities oriented toward service provision, this literature defines communities on a much richer and more diverse basis of shared experiences and imagination, sensitivities or cognition.

Having shared views is sometimes straightforward. This is clear, for example, in conservatives and liberals who have different connotations when they look a newspaper, sports or scientific arguments (DellaPosta, Shi, and Macy 2015). It is more difficult to see in other instances, such as for lawyers whose view of a contract is different form the piece of paper clients rely on (Howarth 2013). Both instances result from a collective process of assigning meaning to objects (Zerubavel 1997). We can consider data science from this perspective as well. This is evident for example in the connotations the word "coding" invokes among qualitative researchers, who understand from it to summarize specific observations into larger themes, whereas programmers understand writing scripts of computer commands. Once we pay attention to such differences systematically they begin to index the boundaries of the thought community that defines them. The process defining these shared meanings, such as memories, which preserve the continuity so striking of lawyers, Christians, families and many other groups, operate in subtler ways.

Memories make for subtle markers of groups because it is difficult to imagine an alternative version of what one knows. They are also so ubiquitous that once we begin to look for them, they quickly illustrate the binding effect in professional and lay contexts. The power of such conventions lies in their externalities, as we can see in the more broadly relevant instance of counting time. Christians begin counting time with the year of Christ's birth, over 2000 years ago, whereas Muslims define the beginning of their history 600 later with the Prophet Mohammed's flight from Mecca (Zerubavel 1997, 85). Similarly, sociologists retrace their beginning to the first half of the nineteenth century when Auguste Comte defined the term. Sociologists use classics since Comte in various ways without critically engaging with their ideas (Stinchcombe 1982) and we think about countless mundane events when we recall a given calendar year, although each time we also implicitly commemorate Christ's birth. For sociologists, moreover, systematic thinking about social problems of course predates all these events. These memories are therefore less consequential for their original facts and more through their rhetorical capacity to coordinate social life.[29] Whereas data science may assign coherent meaning to modern practices such as "coding" and "regression," relevant memories are not so obvious anymore but need to be considered as part of a systematic analysis of data science's formation.

Several more mundane instances underscore this point as well. They also demonstrate how memories of particular events and rhetoric can be strategically used to define templates and hence the basis for novel thought communities. This begins with greetings. Focusing on a particularly consequential instance, Allert (2009) demonstrates the implications of the Hitler salute for Nazi Germany as he recalls the transformation of a diverse landscape of greetings toward one that, imposed through force, made Hitler central throughout society in everyday interactions. Generalizing a similar point, Wimmer (2013) recovers the definition of ethnic communities as a strategic means in the creation of the basis of state power, the consequences of which are indexed in the wars they engage in. Even the Catholic Church has as part of its effort to design a set of distinct memories, such as our understanding of time as considered above, redefined existing ones for its own purposes despite their lack of any theological significance (Brown 2000, Zerubavel 1982). While widely salient, these instances also emerge from a combination of

[29] Carruthers and Espeland (1991) make a similar point for double-entry bookkeeping. To the degree to which its introduction benefited from rhetorical rather than rational effects, we can think of the trading enterprises bookkeeping made possible as thought communities.

factors that could invite to dismiss the importance of shared memories as part of this process. Such dismissal seems less intuitive in other instance. Meadow's (2011) discovery of the transgender child, for example, delineates processes of defining this category through shared experiences and memories to overcome ancient cultural ignorance of this possibility.

These different instances take us further away from our initial concern with data science, technology nerds, and even the enlarged scope of professions, sciences and forms of expertise. Yet, all the memories we have just considered live on through several processes. Oral transmission of defining memories is of course important locally, but writing, art and physical structures all contribute to their durability and anonymize their transmission. Lawyers, doctors, AIDS activists and parents of autistic children share these processes as well, and perhaps even data nerds. Before I specify how the idea of thought communities converts the variation of cases and approaches in the existing literature into analytical leverage for understanding data science, I have to clarify the question of how processes the literature has identified thus far might facilitate lay salience.

2.3.2 The public

So far we have mostly considered how members of these communities come to recognize each other as peers, but not how others see them as a community. For instance, we just considered how sociology defines its boundaries on the basis of remembering a shared set of ancestors, and the gravitational wave research from the previous section recalled the struggle for funding of previous projects. Neither group is salient to the lay public.

In order to gain broader salience, a thought community needs to provide a framework that integrates different moments in which it becomes relevant for others as a class or category of related instances (Dewey 1954). For example, sociologists get recognized for their comments on social problems but not systematically for their solutions thereof, for which they offer advice to policy makers and others. For data science to form such an integrated category we can expect significant complication given the deep historical divides between quantification processes and problems.[30] They include individual

[30] Expert groups are often considered in their institutional form, such as professional associations, scientific labs or expert moments. Although rare, this attempt to consider them as emergent categories is not entirely new, as we can see in DiMaggio (1991, 272), who suggests that "institutional theory has neglected the contradictory tendency of successful institutionalization projects to legitimize not just new organizational forms, but also new categories of authorized actors whose interests diverge from those of the groups controlling the organizations, and new resources such actors can use in their efforts to effect organizational change."

economic activities of interest for the state for taxation and the risks that are part of this, which are encoded in insurance systems, and as distant problems as counting assets in businesses, or stars in astronomy (Porter 1986). While statistics is relevant for all these problems, it has remained an arcane academic effort. Moreover, computer programming is seen to be part of software engineering and thus separate from quantitative analysis. To be sure, the data science context does not lack abstract categories themselves. Counting, analyzing and doing all that with modern technology summarize many specific activities. They are also seen independent of one another. These associations can change over time, as is evident in the emergence of autism as distinct collection of disorders, or considering mental distress as a psychological and no longer a primarily religious problem, as described above. Such change rarely comes easy, however.

We can observe the struggle that is part of such redefinition processes in the early attempts of counting labor—not even expert labor—itself. As part of her historical account of that process, Conk (1980) describes the evocative instance in which a census fieldworker asks a respondent about his occupation. The respondent then indicates the employer, and upon being asked for more specificity notes tasks involved in being of service to that employer, but not the details of the work itself. It took the census bureau several iterations to devise the occupational scheme used so widely today (Conk 1980), as is reflected in the formation of an increasingly refined administrative apparatus. This included, for example, the first organization in a permanent bureau in 1902, over half a century after the first attempt to collect such data, to systematically improve how it gathers these information (Conk 1978). Although this may seem long ago, it indicates the depth of this specialization and hence the distance to other areas of counting. While modern technology facilitates some of these tasks today in ways that make specialization less important, the persistent divide between engineers and practically no less able or critically involved technicians reminds of the social obstacles of such redefinition (Bechky 2003b). For a salient community, we therefore expect that experts recognize similarities in each other's experiences independent of specialized applications, such as projects or organizations, and in ways that the public finds relevant. Similar to the ways in which experts assign shared meaning to skills and problems, discussed previously, their appearance needs to provide the basis for joint recognition. Without formal coordination, the basis for such a shared appearance is not so clear.

Having overcome so many of these difficulties, society has routinized and institutionalized assistance for many types of problems. It helps us see unreturned property as one of those problems to seek help from the police for, and unreturned love as another type, which friends are better for helping with. Society cannot help us categorize arcane problems by definition.

With no one responsible and no process in place, it follows that in order for data science to gain lay salience, it must define and articulate consequential activities in a way that is consistent across substantive or institutional areas and relevant for their clients and others. Lay consequences are often seen to invoke state action and hence much research on professions focuses on their state recognition (Zhou 1993, Abbott 1993). The state also takes time to recognize their relevance. It follows that in order to understand data science's salience, we need to shift focus on its consequences independent of their state recognition. Stepping back and retracing the foundation of the state, Dewey points out that "[t]hose who are affected form a public," and that only based on that the state follows (Dewey 1954, 28). In other words, although the state is important for responses and initiating consequences, the significant process begins to unfold much earlier.

This relationship is still evident in the canonical legal case, despite its tight connection to the state. Courts of appeal, for instance, respond to the problem of decisions once made in a specific context as they come to affect different problems (Stinchcombe 2001). In other words, law integrates problems into its stock of knowledge beyond the scope of a specific problem it has addressed at one point in the past. Similarly, the role of accountants helps their clients get the bookkeeping right and furthermore provide a basis for others to trust the credibility of those clients they immediately serve (Abbott 2011). Contrary, engineering consequences, which are no less significant, are often seen as in association with the organization employing the engineers and not engineering as a distinct group (Vaughan 1996).[31] If we consider these instances in Dewey's perspective together with the first part of this section, lay salience results from experts' ability to form a thought community on the basis of defining a framework for the consequences of their arcane knowledge.

[31] This distinction is not pure, of course. This can be seen as the legal products that provided the basis activities leading up to the financial crisis were less associated with the legal profession and more the banks that implemented them (Howarth 2013).

Overview

The three sections on quantification problems, expert groups and thought communities complement each other. They jointly outline ideas for how data science is able to define a community on the basis of a distinct stock of arcane knowledge, and render it publicly salient. From the first section we could understand the vast spread of quantitative data analysis and processing without finding significant cohesion. In the second section, we saw that while abstract stocks of knowledge are related to whether groups become publicly salient, there is significant disagreement on the processes by which groups form and define such knowledge as theirs. The third section, which has considered such community formation independent of the expert appearance and salience, contributed a range of processes by which members recognize each other. We thus also found analytical opportunities. Experts create a public as they relate heterogeneous skills, and their consequences, to one another. In other words, data nerds create salience themselves. We are then left with asking how these components form a coherent idea in order to use them as guidance for analyzing data science empirically.

2.4 Struggling with uncertainty

2.4.1 Summary

Data science for the first time allows that we observe an expert group emerge. It also reveals a paradoxical process in which arcane knowledge gains public salience. The empirical context in which this process unfolds is not unfamiliar in sociology; the way data science defines it is. Addressing this dual challenge has led us to consider a diverse set of rich debates pertaining to data science's substantive setting, processes other expert groups have undergone, and of community formation more broadly. In spite of this material's heterogeneity we can nevertheless distill a concise set of implications for understanding data science. Such a view must consider expert activity and struggle to define shared ideas directly, and consider that these underlying principles are not limited to institutionalized settings of science and professional work.

This perspective allows us to adjust the explanandum and explanans such that they reconcile the existing debates, as well as the empirical specificities of data science. The basis for comparison shifts from state recognition for services professionals earn their living with, toward practices that generate a public. It follows that the context of work no longer constrains the processes we focus on in order to

explain the problem of data science's salience. Instead we can focus on the definition of stocks of knowledge directly. The division of labor, particularly among experts, remains the relevant context for data science itself, but this detour enables us to consider a broader range of processes that analysts have found to define shared knowledge elsewhere. This shift of focus away from the kind of formal claims and informal collaborations that have been associated with expert work in the past assigns a greater role to activities of individual experts as they define problems data science knowledge addresses. It thereby offers a set of mechanisms that account for ways of coordinating on a scale previously associated with institutionalized mechanisms. Several instances illustrate such an approach, as we see next.

2.4.2 Resolving uncertainty through formal representations of concrete problems

This conceptual apparatus initially rests on the idea that formalization works where it represents concrete problems and where its application leaves room for experimenting and improvising (Piore 2011, Stinchcombe 2001). Blueprints coordinate the work on a construction site because workers who implement them understand their abstract notations, and laws because appellate courts revise and adjust them. Law remains relevant as it is adjusted in ways that accommodate new problems. These adjustments follow from the work of individual architects and lawyers (Stinchcombe 2001). For them the relevant body of knowledge is well known whereas for data science it is not.

Architects and construction workers, and engineers and technicians interact through drawings, which constitute formal representations of the substantive problems. Because these formal representations of the substantive problems they aim to address represent the problems only imperfectly by definition, the experts have to improvise as they apply their abstract knowledge in order to encode the problems appropriately. Data science introduces the additional complication of still struggling to define problems to begin with. Abstraction processes are therefore not institutionalized in courts or blueprints, which are both directly related to the substantive problem area of the respective professional group.

Practical implications follow and we can benefit from the thought community framework and consider a case from a non-expert context. As considered above, for understanding Christianity we need to consider the origins of its calendar although they are irrelevant for its meaning and function today (Zerubavel 1982). Following this example, we need to approach data science with a sufficiently inclusive view as to allow for those backgrounds to enter the scope of our analysis that would later be considered

irrelevant for their professional purposes. In addition to helping to understand data science's salience, such a design leads to several opportunities.

Empirical tests of these directions require rich contexts. Modern technology and the internet spread globally. It is common today that cell phones spread more quickly than landlines and that internet access is more often possible through smart phones than traditional personal computers. Even more than spreading geographically, these technologies penetrate economic sectors and industries, public administration, education, health, science and many other areas of social activity. The design of an empirical strategy that aims to capture data science and the processes by which it gains salience as an expert community needs to reconcile breadth with sufficient detail as to specify the precise claims by which data scientists integrate specific problems and abstract stocks of knowledge in public. Data science introduces the additional challenge of constituting a field that is still emerging. Just like the literature offered limited conceptual guidance for analyzing data science under these circumstances, the directions it can offer toward effective designs remain insufficient as well. The next chapter recovers from it appropriate strategies for today's context in addition to introducing methodological approaches developed with broader problems in mind.

Before we consider these technical details, let me briefly summarize the contribution this conceptual framework I have proposed here makes to the existing literature.

2.4.3 Contribution to the literature

This shift of perspective with respect to expert problems mirrors similar shifts in organizational (Piore 2011) and economic problems (Whitford 2002). Whereas in those two contexts the analytical problem is clearly defined—to identify alternative mechanisms to bureaucratic or market principles—, the heterogeneous institutional contexts experts integrate each come with their own principles potentially shaping expert knowledge. For instance, the concept of 'street-level bureaucracy' (Kaufman 1986, Piore 2011) is designed to account for the role of individual flexibility in bureaucratic organizations. The challenge that is part of explaining the kind of expert work data science engages in is complementary to street-level bureaucracy in that it requires us to explain coordination on the individual level where no formally defined set of rules applies. The same holds for economic problems where the idea of pragmatist means and ends explanations offer an alternative to the prominent yet insufficient rational choice

framework (Whitford 2002). In the context of expert groups, this approach centered on individual activities helps to overcome the divide between formal associations and informal collaborations.

The promise of changing perspective in this way that departs from experts directly is not limited to data science. Data science defines just one instance within a technological transformation that unfolds quickly and widely. It simultaneously creates areas for experts to define novel problems in, as well as an infrastructure that facilitates ways of coordinating solutions around those problems that were not possible before. Data problems have gained salience as one type of these problems. Others may follow, if experts find tactics to make them salient. The processes by which data science has gained salience in the context of the novel coordinating mechanisms offer a useful basis for analyzing and understating those other attempts, successful or not. As mobile phones become relevant for legal and medical work, and take over many other functions in our private and professional lives, it is not so easy anymore to identify the appropriate settings to find the experts that shape these modern applications. Especially as many will not gain the salience data science has as an expert group, understanding the kinds of categories and memories the data science thought community defines its expertise through can be expected to help us understand other cases more clearly. We therefore need a systematic way to watch data science directly.

3 Methods and empirical design

3.1 Established methodologies and novel directions

The diverse debates we just considered for conceptual directions to approach the data science case also come with great methodological heterogeneity. Approaches vary within and across the three fields of professions, expertise and thought communities. Studies of professions often take a historical perspective (Abbott 1988b, Carr-Saunders and Wilson 1933, DiMaggio 1991), many take an ethnographic or interview based one (Menchik 2014, Freidson 1961, Phillips, Turco, and Zuckerman 2013), and some use quantitative analyses (Zhou 1993, Zhou 2005, Heinz and O 1982).[32] Studies of expertise are similar except that they tend to avoid quantitative analyses more strictly. This is because the problems this camp is interested in require analysts to dissect arguments experts make (Collins 1998, 2004, Collins and Evans 2007), as well as to take into account the settings of these claims and the processes leading up to them (Wynne 1992, Eyal et al. 2010, Epstein 1996). Thought communities have also been studied through historical and other qualitative approaches. Compared to the expertise literature, this approach focuses in particular on the distribution and variability of ideas, memories and so on, rather than arcane arguments (Brekhus 2007). In spite of their diversity, all three camps are relatively consistent in that they largely avoid quantitative analyses, at least conventional ones.[33] Although this avoidance could guide research designs itself, considering its reasons provides the basis for devising strategies that might utilize novel data sources and analytical tools.

3.1.1 Existing strategies and shortcomings

The non-quantitative approaches are quickly summarized. Scholars from all three frameworks the previous section laid out, and from all sides of the debates, have observed experts at work. The studies that define our understanding of professions until today were grounded in observations of hospitals and interactions with medical practitioners (Becker et al. 1976, Freidson 1961). These sites continue to provide the ground for addressing question of medical knowledge (e.g., Menchik 2014). Others have interviewed lawyers (Lazega 2001, Phillips, Turco, and Zuckerman 2013, Espeland and Sauder 2007),

[32] The attainment literature that takes into account occupational effects relies exclusively on quantitative methods (e.g., Weeden 2002). As I argued before, however, surveys allow just limited perspectives on the formation of a novel profession that is not yet encoded as survey categories, let alone census data (Freidson 1986).

[33] There are exceptions, of course. Abbott deploys a simple descriptive quantitative design in his seminal book on professions for the legal case (Abbott 1988) and has developed a more advanced quantitative program (Abbott 1991), which I consider below. Although central to the literature, these designs are unconventional.

economists (Fourcade 2006, 2009), public administrators (Seabrooke 2014), as well as scientists from various areas (Collins 1998, Owen-Smith and Powell 2004). Historical studies have relied on documentary evidence in order to retrace the processes, events and turning points that have shaped professions and expert groups (DiMaggio 1991, Abbott 1988, MacKenzie 2011). Research on thought communities has also often studied historical instances from different periods (Zerubavel 1982, 1992). Contemporary studies in this camp have had to define relevant sites more carefully than expert settings often require, such as private households (Nippert-Eng 1996), suburban neighborhoods (Brekhus 2003) or a series of subtle sites that jointly represent instance of moral or value decisions (Cerulo 1998). Overall, these literatures insist on no clear procedures that can be considered standard approaches. In addition to the formal variability across professions, which complicates quantitative analyses, another problem might just be that professionals and other experts often have scarce free time or lack of other inclinations to attend to social scientists studying them. Whatever the reason, the diverse approaches across these accounts suggest that the main methodological challenge might be to gain access to a site that reveals instances in which experts define and apply their arcane knowledge. We can perhaps best described the consensus this literature implicitly reaches as the task to analyze the basis and consequences of expertise and distinct sets of ideas amid organizational, institutional and other contextual effects.

Quantitative studies have relied on formal indicators that surveys could encode as far as possible, such as licensing status, the presence of codes of ethics, or the existence of specialized institutions for formal training (Zhou 1993, Zhou 2005, Heinz and Laumann 1982). These studies treat professions and occupations as formal entities and find associations with externally attributed prestige and status. This idea of a definite set of categories falls short of accommodating an emergent group such as data science. Specifically, the indicators offer limited access to the arcane knowledge experts define as theirs, however, which in many cases only emerges from the expert discourse itself.

A key methodological challenge this literature faces is an absence of systematic strategies for indexing forms of expertise. This problem has been addressed in separate instances. Quantitative analyses of the legal case, for which its professional status is undisputed, have persuasively revealed the variability and heterogeneity of knowledge that is nevertheless consistently associated with the legal

profession (Heinz and Laumann 1982, Abbott 1988). Law's institutionalized status facilitates these approaches in that it offers both clear boundaries and formal definitions and traces of its professional activities. Measuring heterogeneity in institutionalized entities such as the legal professions requires significant engagement with the case itself. As a consequence, analysts have been prevented from devising empirical strategies that could test comparatively the widely made conceptual argument that different forms of knowledge are associated with professional autonomy. Research in other settings has offered indicators and found systematic variability within institutionalized entities.

3.1.2 Empirical categorization

Considering those strategies requires that we step outside the context of expert work. The key achievement of these efforts has been to transform categorical diversity into substantive terms. We find, for instance, that ethnicity is often considered directly salient and measurable. Several recent studies have challenged this interpretation from different directions. One important debate focuses on the role of skin color as a marker of ethnicity, and how individuals and others understand it (Monk 2015). Taking a historical and global perspective, Wimmer (2013) demonstrates strategic changes introduced into ethnic narratives for political purposes. These studies have found different ways that reveal the complexity associated with underlying concepts that seem obvious to many, be it ethnicity or occupational affiliation.

This shows that effective quantitative methods can be designed to the extent that entities have a substantive basis in common. In order to see this, we had to consider applications beyond the professional and expert context. Deriving implications for data science then requires that we understand the underlying principles in more general terms. The studies just summarized have considered formal entities through the perspective of the constituting parts. The specific types of observations depend on the respective setting. These methodological strategies constitute the "empirical categorization" framework (Abbott 1992). The chief argument that undergirds this methodological orientation emphasizes events as complex combinations of variables (Abbott 1992). Whereas quantitative approaches that impose indicators remain unable to capture the variability that is part of abstract stocks of professional knowledge or of thought communities, this set of tools formally takes into account how these dimensions

come together in the context of events and event sequences.[34] In other words, this view focuses on patterns that emerge from indicators appropriate in the respective contexts, instead of imposing standard measures across them. The main task for considering such a design for the data science case then remains to clarify the basis of relevant dimensions or appropriate indicators.

We know how law defines its specialties, for institutionalizing professions we know of the role of events such as starting dedicated schools, and we know of the relevance of skin color for ethnicity, as well as shared histories. We can use this intuition in order to understand cases with less clear boundaries and to recover those contours empirically. In one of the rare applications to expert work, this approach has recovered the significance of temporal sequence of formal events such as schools, associations and so on. The challenge with data science is not primarily one of temporal unfolding, as the advantage of the overall design is already that we are able to observe it within one of the central moments that the other approaches aim to identify historically. In other words, the choice of the moment to study the case addresses the most significant temporal concerns and prior shortcomings. We cannot interpret the data science institutes that have emerged over the last few years as part of an event sequence, as this sequence is still unfolding. In addition to the conceptual challenge of understanding thought communities, which the empirical categorization strategy addresses, it follows that there remains a practical problem of accommodating the various moments in which they express themselves.

These abstract steps require concrete observations. The challenges associated with this approach arise from the multiple types of data through which we can observe the formation of data science, including textual and network data. Analytical tools have been developed along these lines now for a couple of decades (Lazer et al. 2009), with major advancements in the last few years that were able to exploit large-scale textual datasets in order to show the varying fates of intellectual debates in different cultural and institutional contexts of England and France (Marshall 2013), for instance, or the diversity and amount of concrete information collaborators exchange (Aral and Van Alstyne 2011). While these specific contributions are not immediately relevant for data science, questions of contextual effects of intellectual debates and diversity of information both are. These methods for textual data therefore offer guidance for identifying patterns within the traces that data science leaves behind as it gains salience.

[34] This methodological focus also has strong roots in social network analysis, beginning with the blockmodeling framework developed by Harrison White (see White, Boorman, and Breiger 1976 for the original definition of this approach).

Although the existing studies of experts and professions have not taken advantage of these developments, they seem promising.

This review has led to suggest a combination of strategies that jointly lead to a comprehensive analytical design for capturing the emergence of data science and the processes by which its arcane knowledge base could gain broad salience. Qualitative approaches offer significant leverage for specifying the expertise thought communities draw on. At least of the second kind of such approaches help to map out the boundaries and divisions of these communities formally, and thereby formally specify relevant characteristics to cases whose expert status is well known. This allows for comparative designs. We have also seen, however, that the concrete implementation of these strategies has to accommodate the details of a respective case, which I consider next.

3.2 The data science case

3.2.1 Existing designs in the data science context

The description of these tools itself does not clarify how they benefit our understanding of the lay salience of arcane data science expertise. It has provided initial guidance by demanding that we specify dimensions, or indicators, of data science expertise, of which we can discern the patterns in social activities. The public discourse on data science outlines at least some of its applications. The initial claims of its utility in digital companies make them an obvious context for considering data science, though reports on data in public policy immediately suggest a broader scope. We also find its background associated with professional training programs, which various universities have designed. All these instances could seem like obvious places to begin considering the salience of data science expertise. This abundance of opportunities requires some disciplining.

Without any existing description of the field, the first step must be to systematically outline its contours with respect to its expertise and lay application. We could envision survey approaches of the kind used in the legal profession (Heinz and Laumann 1982). Here one could target organizations that claim to have data scientists work for them or seek their assistance. Proposals for academic projects that pursue this strategy have been solicited already. Even as these efforts offer some insights, several questions remain with respect to the association of stocks of knowledge and lay salience. First, data science itself still works on its definition. It follows that imposing survey questions undermines the

variability often associated with professions and makes it impossible to capture where the profession itself is still uncertain over its competencies. Second, whereas law provides a relatively clear population to sample from, data science roles are not yet formally defined. The significant limitations of a survey design with respect to capturing the basis of their expertise result from the unresolved question of identifying the problems data scientists are seen to solve.

Moreover, interview based designs, also used for studying law (Phillips, Turco, and Zuckerman 2013) and other formally defined professions (Gill 2009), encounter some of the same problems and have a harder time of capturing the overall scope. For instance, while interviews respond more appropriately to the incomplete definition of data science expertise, they also require a scope to be defined initially. This cannot be imposed for data science, where one of the key questions concerns the process of defining relevant knowledge itself. One alternative would assume that data scientists are sufficiently connected such that the interviewer could pursue a snowball design. This would leave us unclear over the alternative definitions and framings of data problems that such a strategy might have systematically overlooked. It follows that a qualitative design can offer important perspectives, if it identifies a site at which data scientists and the public face each other for the purpose of defining relevant expertise. Such an approach circumvents biases resulting from the ongoing emergence of data science.

For finding the appropriate design we can consider other cases as we did for devising a conceptual framework. Here it is telling that studies of the medical profession have more readily relied on ethnographic designs than studies of the legal profession have. Indeed, observing interactions among students in medical schools (Becker et al. 1976, Menchik 2014) or teams of medical professionals of various specializations and ranks in hospitals or private practices (Zerubavel 1979), promises much more variation than the solitary activity of drafting legal documents or the scripted interactions in court. Data science by its own account involves significant quantitative analysis and software coding. From this practice-oriented perspective, it therefore resembles legal work more than medical work. In other words, watching data scientists at work promises few observations that could reveal the distinct stock of knowledge underlying their expertise.

The puzzle of data science's lay salience itself points to a solution. Because law's institutionalized status, there is no need for lawyers to prominently articulate their distinct expertise anymore.[35] Contrary, data science, which is just gaining lay salience, is still seen in need of articulating its distinct role. This kind of effort needs to be taken into account as part of the empirical design as it offers a much richer context for observing data scientists defining their work and expertise. As much as this rare moment we find data science in promises novel observations, it also comes with complications. Most critically, it offers no clear opportunities to consider the features of abstract data science expertise relative to those of established professions or expert groups and their training institutions, associations, and so on. I therefore consider novel techniques that are capable of exploiting more minute, informal instances of expert work that occur in the practices of the emergent community of data nerds as well as in those of the established occupational groups.

In order to interpret results from the novel techniques requires that we gain a robust understanding of data science. To this end, I propose in part one a qualitative design that is responsive to the challenges outlined above. While a detailed description of the design follows below, let me just briefly note that this analysis relies on observations from the data technology scene in New York City, which I have collected over a three-year period. I have focused on events at which data nerds articulate their expertise and applications to the public and to each other. This rich analysis provides a robust basis for novel, comparative designs, as I describe below.

3.2.2 Designing empirical strategies for the data science context

The conceptual framework of empirical categorization from the previous section, together with today's ubiquitous recording of behavioral traces and modern tools for processing them, offer leverage to address this problem. The key challenges in the data science formation that traditional methods seem to address less well pertain to the scope and shape of the stock of knowledge data science defines as its own, relative to those stocks of knowledge of other professions, expert groups or thought communities. And with new opportunities the more basic question arises of what might constitute appropriate observations. The implementation of these tools and methods requires types of events that are both

[35] Though see the fierce arguments in the past (Stevens 1983, Abbott 1988).

incidents of skill, knowledge and expertise, and consistent across the heterogeneous contexts in which groups, communities and professions emerge and operate.

The most basic activities of expert thought communities involve developing and applying arcane knowledge by engaging in research and arguments over its status and advising clients on its basis. This happens for most expert groups most of the time in some form of organizational context.[36] They offer a basic scope of comparable patterns by which expert groups can be observed and thus provide an opportunity for the analyst to define the expert communities or categories empirically (Shwed and Bearman 2010). Studies of the salient and institutionalized professions like medicine and law have long emphasized skills as a central mechanism through which experts apply their arcane knowledge (Freidson 2001).[37] While those accounts were able to turn to the respective expert groups for definitions of those skills, data science offers no clear profile. Moreover, the existing literature offers no indication on how skill arrangements index abstract knowledge comparatively across expert groups. This combination of clear conceptual direction with a vague operational guidance requires an approach that pragmatically identifies an area that describes expectations of expert work and that pertains to a range of cases for which we are certain of their professional status, such as law, and its absence, as in many instances including engineering occupations. The implementation of such an approach can benefit from modern capabilities to analyze textual data at scale, which therefore offers an approach to address this central question.

Chapter nine in part two leverages these novel capabilities. It analyzes a dataset of job descriptions, which are both central and significant to expert work, and requires from this novel approach that it translates the substantive significance into analytical leverage. In response to the existing literature, law servers as the main comparative case for data science. As counter factual cases, I also consider three areas of expertise that have not gained lay salience as autonomous expert groups. I analyze financial advisors, who have lost their autonomy quickly to banks as well as software engineers and risk analysts, two groups substantively close to data science but also without significant autonomous salience.

[36] The presence of organizational embeddedness should not be read to index the degree of autonomy, as its implication can vary across organizational roles and functions for different occupations. Hospitals for instance facilitate much of the work of medical doctors, which has only more recently fallen under greater dominance of insurance firms (Gorman and Sandefur 2011).
[37] Conk (1980, 41) documents the deep roots of this problem by pointing out that the "1910 Census thus institutionalized a methodological problem. Without providing an adequate definition of or criterion for determining 'skill,' the United States Census began to classify occupations according to 'skill.'"

The problem of vague operational guidance also translates into the second feature of expert groups, their autonomy over knowledge production. Here much research has focused on citation practices as indicators of knowledge production (Evans and Foster 2011). Some of these approaches are interested in the origin and role of individual contributions. The question of the emerging data science thought community reflects a collective process. Here the framework of empirical categorization implies that we compare groups of scholars on the basis of how their contributions engage with available stocks of knowledge (Shwed and Bearman 2010, Moody 2004, Barabási and Albert 1999). This framing is consistent with the focus of the thought community literature on the anonymous character of groups who encoded their shared experiences and perspectives in knowledge. The key question following from the conceptual review then is how they depart from institutionalized practices of knowledge construction, beyond personal communication. Once again, novel tools and methods facilitate that we can address the classical concerns of the field in the modern data science context.

Chapter ten in part two analyzes data science in its academic context, with a focus on patterns of scientific contributions of scholars who train data scientists. The patterns by which these contributions build on older ideas in ways that are similar or heterogeneous directly index stocks of knowledge. I compare the shapes of these stocks of knowledge of the data science instructors to those of legal instructors and laboratory directors training graduate students in systems biology. As before, and consistent with the existing literature, law, although substantively very different, offers the main baseline as law school faculty produce in academic scholarship as well. I consider systems biology as another recent formation of a thought community that has remained arcane as a field of expert work.

I discuss the empirical, methodological and analytical details in the respective chapters.

Overview

From the many ways in which others have studied specific professions and expert groups, here we were able to identify three main themes that promise to account for data science's distinctive content and the scale at which it emerges. These themes address the question of how data science defines its distinct expertise, how its skills are seen to apply to practical problems and the principles by which its stock of knowledge integrates distinct arcane roots in scientific expertise.

This translation of existing strategies and conceptual directions into the data science context also allows us to consider the competing explanations from existing research. To recall, they have focused on the effect of the technological background of a field of expertise, their organizational and institutional context, and the formal and informal ways of coordinating. A qualitative design that focuses on accounts around data science's unsettled definition of distinct expertise offers a sufficiently wide scope to allow those processes the literature has specified to enter the analysis. We have also devised novel comparative approaches, which generate data science and familiar expert groups from the constituting features, by matching classical conceptual directions with modern technical tools in the organizational and academic setting, the two main contexts of existing studies. This combination allows us to directly specify the features of data science and other expert groups that gain lay salience among each other by comparison to those that remain irrelevant to the lay public. All these considerations leave the advantage the emergent status of the data science case offers largely underappreciated. Defining a new field in a context without formal control also implies a struggle of resolving the significant uncertainty that is inherent in that status for the analyst, a problem that I consider next.

3.3 Data science of data science

The designs I have just described entail quantitative strategies that resemble those in data science more than conventional quantitative approaches in social science.[38] They constitute a sort of data science of the data science case. Here I argue that this setup offers analytical leverage from the reflexivity it introduces, in addition to the results these methods reveal of data science on the basis of the data they help analyze.

The quantitative approach leads to the paradoxical effect of also improving our qualitative understanding of data science. This payoff comes in two steps. On a high level it provides technical expertise. This is common in studies of expert groups, which follow the claims and arguments by which experts articulate their distinct knowledge (Collins 1998, MacKenzie 1978, MacKenzie 1981). For instance, this method enabled Collins (1998) to discover varying "evidentiary cultures" within the same field of experts as a result of his familiarity with the underlying questions. This "participatory expertise"

[38] This will become apparent from the details of their technical implementation, which I leave to the respective chapters.

allows the analyst to inquire into the background of the claims experts make.[39] We can even see this effect in an almost universally salient context. Weber (1988) quotes a Benjamin Franklin speech as a means of introducing the continuity of protestant ethics from their religious origins all the way into political and economic ideas at a different time and on a different continent. Franklin's prominence, and the importance of the country he represents, renders the meaning of the arcane theological and historical backgrounds of economic activity tangible for us today. This strategy of learning the ideas of a specific group allows us to recognize similar arguments in different terms, as well as different arguments in the same arcane language. While these details are critical for understanding the basis of data science's expertise, they fall short of revealing the struggle data scientists engage in as they seek to address problems for which no correct solution has been defined, or is at least not accessible to them. In other words, they assume sufficient agreement on the problem as to allow arguments over appropriate solutions.

The second aspect of implementing data science strategies addresses this point. The interpretive understanding that more readily surfaces through intimate familiarity with a setting builds on the researcher's deeper understanding of the kinds of arguments that are being made.[40] This level is not so often made salient, either because the analyst and her audience are familiar with the terms already, just not in the respective compositions the analysis reveals or because the terms have been formally defined. While the former mostly applies to contexts from everyday life, the latter also holds for instances as arcane as the gravitational waves community just considered, which defines its knowledge formally. The analyst can study those formal definitions.

Importantly, the community comfortable with using and applying the terms also understands those that are not as formally defined as they could be. Zerubavel's reflection on the struggle with understanding the difference between jam and jelly or trash and garbage upon joining the American "thought community" as a non-native English speaker indicates this point for the lay context (1997). We can also see it in Evans-Pritchard's analysis of the Azande in his frequent remarks on how the community

[39] All this is to say is that acquiring some expertise in a given area is necessary for the analyst to establish the same conditions one encounters in lay contexts, which analysts are largely familiar with anyway. MacKenzie (1978, 1981) took advantage of his own training in statistics to a similar effect.

[40] The many contexts in which this familiarity has contributed to our sociological understanding include having growing up on a farm (Bourdieu 2008), taking boxing lessons (Wacquant 2004), attending elite boarding schools (Khan 2011) or experiencing professional events as a spouse (Fourcade 2009).

he studies uses terms differently from the way they would translate into the western context. Understanding a word from their language that translates into poison in the formal definition would provide a false account of the way Azande use it and its consequential role in their decision-making (Evans-Pritchard 1978, 212). It took both analysts several years of encounters with the problems they present in their native setting. In other words, it takes experience with using the items and terms of a thought community, if it does not define the terms formally.

This point can be briefly illustrated in data science itself, even before entering the analysis. Data is important for data science, yet its definition is not always clear. Aside from the popular debate on "big data," which the first substantive chapter addresses from the data science perspective, the Wikipedia definition pointed out that the availability of "unstructured" data has been widely associated with data science competencies as well. When I began with the project I was well acquainted with different sets of structured data, which organize the observations in rows and columns. I therefore also knew, at least judging from the negation, that unstructured data would not have those. I could not imagine, however, how it might look instead, let alone how to analyze it. Then it came that I collected data on expert tasks (discussed above and in more detail in the first chapter of part two). They were available as "JSON" files, a format I was unfamiliar with. Unlike the structured datasets I had used before, these JSON files did not have rows and columns. Yet, I found some simple tools that made them easy to query and navigate. Moreover, this design facilitated updating existing datasets as I collected more observations as a consequence of data science's ongoing emergence. Whereas classical spreadsheets would have required that I indicate a complete set of variables, or add missing information for all existing observations when a new observation with another piece of information enters the dataset, here I was able to include all information efficiently during the collection process. While some data scientists learn about JSON files as part of their formal training, many do not. In other words, my experience (in this instance and many others of this sort) likely resembles those of practicing data scientists with respect to struggling to address problems (my question of how expert tasks describe data science skills) with new technological capabilities (unstructured data in JSON formats). I experienced the process, and struggle, of finding technical solutions to substantive problems.

In short, consistent with much research on expert groups, this design of using data science techniques provides a certain amount of literacy that facilitates how I read the discourse in the data nerd community. Moreover, it provides me with the experience of navigating the uncertain context in which neither problems nor solutions have clear definitions. This is partly necessary to achieve the appropriate literacy in a context where few formal resources exist to study it. It also turns the observations Coleman (2013) and others have made regarding the attachment to technological work into analytical leverage. Whereas those accounts found that personal attachment to the technology work matters, this design allows for specifying the precise ways in which it does.

3.4 Organizational arrangements in analytical and practical terms

In order to persevere the distinct qualities of the positivistic analysis and this interpretive design, and thus to leverage their integration, I divide the analysis into a qualitative and a quantitative part. The overall argument follows from both the analysis of what I observe (qualitatively and quantitatively) as well as the experience of designing the quantitative analysis. Quantitative analyses can be accused of their inability to capture the meaning of the relationships they find, whereas qualitative accounts are often accused of missing patterns beyond their local scope. Here both inform each other. In other words, learning from the field teaches us about the field.

The conceptual and methodological apparatus we have just devised provides analytically useful abstraction. This leverage comes at the expense of obscuring the substantive problems around experts interfering with matters as significant as family privacy and emotional wellbeing, which we have seen in association with data science. In combination, however, the two sides provide the basis for a robust and relevant argument. In order to reconcile both sides, I map the familiar cast of technology nerds from the introduction onto the arcane concepts around professions and thought communities from the theory and methodology sections. For this purpose, I focus on the principles of organizing work they respectively represent and describe. In other words, I suggest that we can understand data science and its salience through the roles promoted by C. Wright Mills and those embodied by Bill Gates, Aaron Swartz and Linus Torvalds. These characters serve as representatives of different types of arcane thought communities and the respective organizational processes unfolding within them around formal boundaries and informal

coordination.[41] Here I summarize the steps we have taken so far in order to demonstrate how they complement each other.

3.4.1 Analytical design

The existing literature revealed conceptual models of the different ways in which expert groups define their stocks of knowledge. Professions are seen to establish their status formally whereas expertise movements pursue their agendas often informally. Both constitute instances of thought communities. This integrated view introduces specificity into our understanding of thought communities by describing different types of these communities on the basis of the principles we find in professions compared to those we find in expertise groups. Conceiving of them in this way has the advantage of relaxing the formal understanding of professions and of scaling the local and informal definition of expertise movements. This step alone falls short of addressing the question of data science's salience. We need to operationalize different principles by which professions and expertise movements define their stocks of knowledge empirically in order to investigate their relationship with those respective outcomes.

Building on the conceptual framework of thought communities, we have also designed a series of analytical and methodological steps in order to index their principles empirically. Data science's emergent status requires that we begin with qualitatively mapping out the contours of the stock of knowledge data scientists articulate their public relevance with. While this strategy is pertinent for revealing contours of a distinct thought community, it falls short of clarifying with sufficient certainty the principles by which such a thought community gains lay salience. I have therefore designed a comparative analysis as well.

Data science, just like many other expert groups, lacks formal markers. It follows that we need to recover its contours from concrete instances in which experts define and apply their expertise. We have found comprehensive conceptual guidance for designing such protocols, but primarily for a lack of technical capabilities until recently no existing implementations. We could combine these existing ideas as a basis for operationalizing distinctive knowledge production and application, and for recovering

[41] Relating them to one another thus provides a rhetorical device for how the argument works. Other contexts offer this utility directly. States have heads, firms have executives and churches have their respective leaders. Even where we are interested in the role of unnamed masses, the citizens, employees and believers, we have a relatively tangible image of some form of actors that helps guiding an argument. The same holds for some expert groups, where analysts articulate their argument with references to Sigmund Freud in psychiatry, Albert Einstein in the sciences, or several other less famous individuals. The division of labor Mills describes lacks those visible actors or roles. There is no direct association between these individuals and data science, which underscores this interpretation here as models of practice, and not a source of influence.

abstraction as the integrative basis. Moreover, modern computational methods allow us to take advantage of unstructured data available for cases from different institutional and substantive contexts. As a result of this we are able to recover the precise patterns of analytically useful cases that help us specify the principles by which arcane knowledge gains lay salience.

This formal strategy complements the models of work we derive from descriptions of the substantively relevant types of technology nerds. In other words, the analytical setup behooves the initial problem in the division of labor C. Wright Mills helped us recognize in its historical trend, and which we could again see through Bill Gates, Linus Torvalds and Aaron Swartz in the modern technology context.

3.4.2 Organizational arrangements

We can now view the dual substantive and analytical problem data science confronts us with through this setup. Here the question translates into asking whether, on what basis, and in which combinations data science resembles the tech nerds and the more anonymous learned occupations of lawyers and medical doctors, for whom we have considered Mills as a representative. One way to start addressing this question would consider the individual data scientists who are credited with having defined the field. Chapter eight, the last chapter in part one, considers their role in detail. Beginning with this group would ignore the collective basis of the initial definition. It would also make it more difficult for us to recognize Mills's features of a more anonymous yet integrated expert role. Instead we need to allow for role differentiation to emerge from the community directly.

This strategy generates the following mapping of characters. Bill Gates provides a face for bureaucratically specialized work. In the literature, we have seen the focus on this kind of work begin with the classic research on professions, which grounds itself on the distinction of learned and autonomous occupations to those whose tasks follow from their organizational setting. Among many clerical occupations, engineers offer a case that has long puzzled the literature for not gaining autonomous salience. The empirical design reveals support for these processes to the degree to which we observe patterns of data nerd work overlap with bureaucratic or other formal task and work definitions and settings.

Literature on expertise movements has recovered more of the complexity in work arrangements, as we could see particularly in autism and other medical problems. In the substantive technology context,

Aaron Swartz and Linus Torvalds represent two types for technology expertise movements. Similar to the technology nerd defined by organizational boundaries, the type of hacker Torvalds stands for emerges from patterns that fold into specialized expertise. Unlike the bureaucratic definitions, however, specialization stems from informal components of their expertise similar to the groups that defined medical conditions that we have found in the literature. In the organizational terms of modern technology work, we have found this type of expertise to resemble heterarchies. Whereas heterarchies are often seen to emerge as employees communicate and coordinate across formal specializations, here we have seen Linus Torvalds design bureaucratic principles on top of informal and specialized projects. We can therefore think of these arrangements as "inverted heterarchies." Because of the open structure of Torvalds's Linux movement, and those like it, evidence can emerge quite explicitly with direct engagement in it, or in the form of accounts that emphasize the significance or intensity of communication or collaboration as a way of organizing data work.

Hackers like Aaron Swartz coordinate their activities on an informal basis as well. Contrary to the group Torvalds represents, though, they work without mobilizing and maintaining a larger community. There is therefore not much organizing that is part of it. Yet some of the expertise research has found how analogous experiences lead to shared knowledge others could not imitate (Collins and Evans 2007).[42] We have begun to see indicates of such coordination in the technology contexts as the hackers Coleman (2013) studied often relied on just minimal interaction. We can also consider a practical instance in which a talent agency that usually finds gigs for musicians has begun taking these nerds under contract as well illustrate part of the nature of this work (Widdicombe 2014). Unlike rock stars, who need to meet the taste of an audience, technology nerds have to fit into the project specifications they are being hired for addressing. Without prolonged engagement, which this model circumvents, they need to improvise. Evidence of such processes in data science would emerge in the methodological design through patterns that break with otherwise shared principles altogether.

Finally, the learned professions are widely salient but remain largely anonymous. Because Mills's central point emphasizes the utility of this anonymity as a general role model for guiding occupational mobility, we can consider Mills as a representative for this group for analytical purposes. Although

[42] Collins and Evans (2007) studied these processes in arcane contexts but also in common ones, such as color blindness.

professions are often observed through their formal association, the members and hence the principles through which they operate, work and unfold in the context of their clients. It follows that we cannot expect evidence in the form of overlap with formal boundaries when we consider patterns of expert work and knowledge. Instead, here we expect that their abstract knowledge systematically transcends the formal boundaries and specific client problems.

Let me illustrate this arcane conceptual design of differentiating between types of work on the basis of how their tasks unfold. Andrew Abbott has, as part of his seminal project of analyzing expert groups from the perspective of how they claim distinctive tasks as theirs, developed a methodological framework of sequence analysis. As I described with more specificity in the methods section, this method aims to delineate and characterize groups on the basis of the development steps they undergo. Abbott illustrated the sequence methods through a study of traditional dances. This analogy clarifies the other organizational arrangements as well. In his original analysis, Abbott focused on the sequences of morris dances in historic villages of the Cotswold region of the UK (Abbott and Forrest 1986). Like Mills's understanding of the autonomous professional, all participants of a traditional dance follow the same steps. This prevents chaos even as no one centrally coordinates them. Take ballet as the opposite, where choreography is critical. Like in a bureaucracy of Gates's proprietary work, each dancer follows a carefully planned pattern. Then take informal expertise seen in the specializations of heterarchical arrangements of Linus Torvalds, or in more vivid terms of breakdance. Groups collaborate in order to put on a coherent show, but follow no formally planned steps of a ballet, nor the routines of a formal dance. And lastly, in order to consider hackers of the Aaron Swartz style think of Steve Paxton's contact improvisation idea from the 1970s. Although it constitutes a larger movement in dance, the specific performance requires neither preparation nor intensive communication beyond slight and fleeting contacts. There is clearly no plan nor a uniform pattern and not even the informal coordination of a breakdance crew.

Overview

Taken together, we can think of the different organizational arrangements of work both in substantive and in analytical terms. In substantive terms, we have Bill Gates, Linus Torvalds, Aaron Swartz and C. Wright Mills. I introduced them and their characteristics in section one. In analytical terms, we have bureaucratic specialization, inverted heterarchies, contact improvisation, and abstract

knowledge. I have defined these principles in section two on existing literature. The substantive terms are designed for better keeping track of the empirical observations in part one on types of technology work. The analytical terms are designed for generalizing those observations, as described in section three, and as a basis for both the qualitative part one and the formal design in part two.

I. Data science in New York City

Introduction to Part I

Data science is widely discussed in public and professional discourse. Yet, unlike our relationships with accountants, medical doctors and other professionals, few have ever sought or will seek advice from their data scientist. As data scientists are rarely encountered in person, it follows that their ubiquitous presence in modern technology conversations, sometimes giving them a concerned tone, becomes hard to make sense of. And it seems possible that this just exposes them as an unsubstantiated role and misleading marketing tag line. Data science's recognition across different industries, academia and the public sector render such a conclusion unlikely, but by themselves these observations still fail to resolve the puzzle of its salience. We have just seen in the previous chapter several directions for gaining a systematic understanding of this question. These directions suggest considering the grounds of collaboration and competition as well as more abstract forms of coordination and community building, which have also occurred in other types of technology nerds. Analyzing the degree to which these processes and perspectives pertain to modern data nerds requires that we watch data science's salience unfold directly.

The following chapters set out to discover data science in the larger technology scene and in a setting that allows considering the directions the literature suggests. This requires an empirical design. At least by some accounts originally from San Francisco, data scientists have entered established organizations and emerging startups across cities in the US and globally. This background might suggest San Francisco as the ideal site to find and observe them. Its long technology legacy however easily confounds our understanding of data science's salience.

New York, by several measures another prominent place in data science today, has failed to establish strong technology sectors in the past. Observing data science there thus has the advantage of seeing it unfold independent of a general technology legacy. A qualitative design is most suitable for taking into account the large and often relatively unfamiliar features of modern data technology at least in a preliminary fashion. It encounters another challenge. As I discuss in the following section in more detail, data scientists distinguish themselves little in their daily activities, revolving around computer screens, from other professionals or experts. Most unusual for corporate offices are perhaps their stereotypical hoodies, which at the same time they have in common with the tech community more broadly where they

have become the modern white-collar. We therefore need to observe data scientists in moments that reveal their distinctive work.

Both the technology scene and New York City are big. Briefly recalling the main directions of the research that we have considered above helps to find a more specific setting. The processes those studies have described for other professional and expert movements range from competitions lawyers have engaged in to monopolize legal work to coalition building of medical experts and patients around the recognition of as significant issues as autism and HIV. For the local technology scene these studies imply to avoid looking for data scientists directly, as that strategy might easily miss other groups competing for similar, data-related problems. Moreover, the activity at relevant sites needs to be sufficiently rich to capture the relations data science or its members might foster as part of defining and promoting their expertise around data problems.

A second set of directions focuses on group formation more broadly. They particularly address such instances where more subtle definitions of thought communities replace overt mobilization on the basis of shared memories, understandings and ideas. These processes pertain to many non-professional aspects of social life, such as religious groups, ethnic communities and even nation states. In each instance, these studies reveal the impetus of subtle characteristics such as cultural conventions, historical narratives and other identity markers as part of informal group formation. In data problems, several occupations and academic fields contribute to the technical basis data science draws on. This overlap does not contradict a distinct role of data science by definition, but for data science to define a distinct jurisdiction, it requires discussions with sufficient depth as to move beyond those concerns all these groups have in common. In more practical terms, we need to find a setting that accommodates rich interactions. I describe such a setting next.

We have also begun to conceive of these arcane conceptual principles in more common terms. For this purpose, we have turned to the set of publicly debated technology nerds as representatives of different arcane ways of organizing work in the technology setting. Accounts of software development have for instance emphasized a division between proprietary projects and open projects. Two prominent characters in this area are Bill Gates of Microsoft and Linus Torvalds, creator of the open Linux operating system, respectively. Both types of projects require significant specialization and integration of

components into the larger system. In the type of work Bill Gates represents, these specializations follow from the bureaucratic division of labor, whereas in the case of Torvalds, they emerge from the community and follow protocols for coming together as a comprehensive software product. In our arcane terms, the type of work Gates represents maps onto those occupations that receive directions from bureaucratic hierarchies and thus differ from the learned occupations or professions. Torvalds, although also representing specialized work, maps instead onto the expertise literature that emphasizes informal movements seen so far most clearly in the medical context. For another emphasis of this literature, we have also considered the hacker Aaron Swartz. Swartz is known for a series of projects, which have no technical relation to one another. In other words, here we see more heterogeneous orientations than specialized work. Finally, it is not so clear who might represent the learned professions, such as lawyers and medical doctors and the arcane process unfolding in their context, precisely because of their anonymity. We have therefore considered C. Wright Mills as representative for the professional organization of work because it was his account that put particular emphasis on the significance of anonymous salience, which also concerns us with data science. This brings us back to the two main problems data science confronts us with, and which we begin to address in this setting.

Besides recovering the problem of salient and obscure forms of work, data science has also caused public concern for its impact on everyday life. This was evident in the reports on quantitative data analyses that inferred and acted upon private circumstances and emotional states. At first glance, the responsibility for these activities is immediately clear, as it can be associated with the grocery store chain and the social networking site that implemented them. Such a conclusion pays too little attention to the systematic association of the activities through the data science community. These associations become all the more significant because of the different contexts that they invoke beyond these two instances. It follows that we need to study the processes underlying the design and implementation of these applications in order confidently conclude how they undercut boundaries and connect different contexts.

This focus leads to the question of what defines the relevant expertise and controls its application. It also pertains to Mills's point from early on that emphasizes the consequential decline of autonomous work and role models of free entrepreneurs and the bureaucratization of learned professions. In other words, understanding the principles that coordinate data science expertise is significant because its

salience may serve as a novel instance to guide younger generations into positions that are much less subject to bureaucratic oversight than has been common for many occupations over the last decades.

Setting

This research is concerned with the public salience of arcane data science expertise. It follows that we need to consider how data science engages with the public directly. Especially the digital nature of much data science work facilitates significant indirect connections, as the public uses data science products. We have considered summaries of media reports on many such instances above. One could glean important details regarding data science expertise through case studies of how they construct these applications. At the same time, data scientists work with computers when they code, analyze spreadsheets or prepare reports and presentations. They also meet to discuss and coordinate their work and the problems they encounter along the way and casually bounce ideas off each other as they do these things. They resemble many professionals in those activities, albeit in varying compositions. Even more detailed observations of the decisions they make regarding quantitative models and strategies to build applications that shape social life raise a series of additional questions of how it is that the expert group gains salience itself, separate from the organizations that benefit from those implementations. Understanding the characteristic features of their expertise would therefore benefit from observations in which data scientists face the lay public directly.

Data scientists, as many professional programmers today, actively endorse and contribute to the open-source movement, in which the code underlying applications is publicly made available for further improvements and adaption. They also participate in less technical public meetings of the technology community, as part of which they organize data science themed events and conferences. Instead of strictly focusing on making code available, these meetings typically involve presentations and interviews on stage and in front of audiences and focus on the work more generally. These events constitute a direct process by which data science becomes publicly salient.

We have to view them in their dual role as simultaneously scripted and intrinsic performances. The accounts I consider often look like prepared statements or interview answers. This is because they are often observed as part of public presentations and sometimes directly result from Q&A interactions. In both instances, however, they address the public or respond to questions from event attendees, not the

researcher, and thus constitute intrinsic professional activities. They therefore address challenges related to the emergent status of data science, which I discussed above.

The activities I observe are part of regular gatherings open to the public and taking place after working hours. They include presentations and discussions on topics around quantitative analysis, data storage and applications. They thus directly capture the moments in which experts try to make their arcane knowledge broadly salient. Observing these events and following these interactions resembles an ethnographical more than interview-based research design, although the accounts I present below may occasionally suggest otherwise. At the same time the scripted program and style of these events, following conventions of modern professional presentations, undercut most analytical leverage aside from the substance and rhetoric of the presentation and interactions between speakers and the audience. The standard features are quickly summarized.

The majority of groups I have joined meetings of have coordinated their events through the web service meetup.com. The service was founded in the aftermath of the September 11 terrorist attacks and as a response to the widely reported decline of American associations, aiming to enhance civic life in general. Whereas the service mainly consists of an online interface which users can sign up with and that offers an infrastructure to organize activities, groups meet in person. These activities range from leisure themes, such as book clubs, hiking groups, and wine tastings, to the professional groups with technical interests I have attended. The online interface offers email lists, calendars, meeting-specific message boards and upload options for photos, notes or slides such that groups can organize regular meetings with. Activities of these groups vary in size and frequency. I have joined about a dozen groups for the purposes of this research. I have attended at least one event of most of them, but focused on attending meetings of three or four groups regularly, ranging from a focus on data problems in general, to groups specifically focusing on statistics and machine learning. Data science has regularly emerged as an explicit or implicit theme at these events.

Each group has its own administrative responsibilities. In the groups I have participated in they have remained minimal, however, and were left to designated group organizers who schedule events and find hosts. Different, more or less prominent organizations in New York City have hosted (though rarely organized) the meetings I have attended as part of this research. Bloomberg and Facebook as well as

New York University and Columbia University are among the most well-known, but also Spotify, Stackexchange, and many other startups, co-working spaces and incubators host these events. These hosts sometimes took the opportunity to introduce themselves and give a short pitch, but never for more than a couple of minutes. Also, with just one or two exceptions, none of the presentations primarily seemed like sales or marketing events. Some academics presented their research, some for-profits showcased their products, and sometimes programmers demonstrated new open source tools they have been working on. Presenters were always asked to frame their presentation for an audience that is interested in the problem area broadly and to offer their experiences in it, but not to pitch their product or findings.

All events I have visited follow a similar schedule. They begin at between six and seven in the evening, officially with some time for networking, but more importantly for attendees to arrive. Indeed, many just wait on their own until the main part begins, although others catch up with friends or other regulars, or meet new people. Then the main part follows, with typically one, sometimes two, but in one regular event up to five speakers. These presentations cover some topic related to data, such as storage or analysis, with varying degrees of technical details, ranging from oral histories of how projects were started to scientific discussions of different ways of scoring statistical model performances. Live demos are often part of the presentation, and, at the slightly more technical events, presenters do live coding. These substantive presentations or discussions last for around an hour to over two hours, with opportunities for some more networking afterwards. Just like during the initial gathering, many attendees leave immediately, even rush out, while others stay around for a bit and chat.

Data science's developing status made it impossible to define a strict scope for the groups I participate in. For instance, one group emerged in the first year of this research explicitly titled data science (although many indicate their affiliation with the theme in a longer list of topical keywords), and I attended some of its meetings. The advantage of not limiting my scope to this group was that I could observe many occasions at which experts from the broader technology field talked about data science or their experience with data scientists, and not only data scientists promoting themselves. While the mere observation of these references leads to no deeper conclusion beyond indicating its salience, the nuances around these different perspectives offer a rich basis for beginning to understand the sources of

data science's broad presence through the ways in which data nerds define their expertise in public settings.

Even without drawing boundaries in advance, over time I reached some saturation with respect to both temporal and substantive scope. Some instances illustrate this process. One group began each meeting with giving away free T-shirts based on a random draw.[43] I have won twice. I have also seen the same speakers at different events as well as a number of participants. I have seen speakers present at multiple groups, or repeatedly at the same, and participants attending events of different groups. This is not to suggest, however, that this community is small, or settled, as novel topics, solutions, and speakers still dominate.

These events are also meaningful to the participants. Participants not only chat in small groups in the beginning, indicating at least some cohesion, but more importantly also regularly ask questions and themselves take the stage at the beginning for brief announcements, such as regarding job openings or new projects. One speaker recounted when he first came to the event, as a participant then, working for a large company but met his co-founder and now came as a speaker and founder of a small company. Organizers, in turn, also work on ensuring that the meetings remain meaningful for participants by disciplining presenters to address audience interests over their own and by keeping out recruiters all together.

I have no systematic profile of the participants. While a couple of them surely attend for free food and drinks, where this is provided, many come for the substance. This became clear in both direct conversations as well as indirectly through their questions to the speakers during and after the presentations. Some of these interactions are part of the following analysis. The groups also reflected some diversity, both within and across events, with respect to participants' backgrounds and interests. Moreover, that significant attendance holds even for the most technical and mathematical talks also speaks to their significance. I have once again no quantitative data on them, but it was customary for speakers to ask for a show of hands regarding the compositions of the audience (technical and non-technical participants) or their familiarity with a programming language, software package or analytical method that was relevant for the talk. Based on those observations, at not strictly technical events,

[43] The function which downloaded the list of participants and chose one was of course programmed by the organizer

technical participants were in the minority, but in general there were always participants ready to challenge speakers on all kinds of topics. I did see complaints about insufficient technical specificity once in email comments on the meeting thread following the presentation of a marketing person. I have also seen a speaker apologize to a professor in the audience for the lack of technicality of the presentation (the faculty member was still interested enough to ask a question afterwards).

All this suggests that these events and specifically the presentations they feature provide a rich empirical context to study the formation of data science. Similar to an interview design, these events provide a framework that organizes the substantive focus: data. Like in an ethnography, it is not the analyst who imposes that framework; it emerges from the site itself.

I have kept detailed field notes during all these events, typed on my laptop computer. While this would be awkward in many settings, here it was common for attendees to have their laptop open before and during events, often with computer code on their screens. For the analysis, I coded the notes on a simple scheme, which I developed on the basis of notes I have taken during a semester-long data science class. This class, while only representing one view of many that are possible, significantly shaped my own perspective before I went into the field where I made the main observations. The class attempted to provide a one-semester introduction, and combined both technical skills and exercises with applied problems directly introduced through external speakers from the industry. It also offered two sets of tutorials for the mathematical foundations of data science and their application through statistical computer programming as well as a two-day workshop as a general introduction to computer coding. After the three years in the field that followed that experience, I recall it as a comprehensive introduction and thus a useful basis for a coding scheme of the field notes. For the purpose of developing the scheme, I revisited the notes on the course and extracted themes ranging from mathematical and analytical skills to substantive concerns and applications in various domains. I then turned to the main corpus of field notes and coded them on this basis. I used this step also to update my coding scheme on the basis my observations from the field and coded them a second time on the basis of the updated coding scheme. For the purpose of presenting the results of this analysis, I went back and extracted the specific accounts I had taken notes on during my time in the field from videos that were recorded at the time. I have furthermore participated in meetings remotely after finishing my time in the field directly. As part of this, I

have watched video coverage of meetings I had not participated in at the time when they took place, but that were organized by the groups I have joined regularly.

Next I articulate the analytical concerns from chapters one and two in the context of this setting.

Analytical strategy

3.4.3 Conceptual framework

We can study data science in this setting with the analytical setup we have designed above. To recall, we have reconciled substantively rich but analytically impoverished tech nerds from the introduction with a precise conceptual and methodological apparatus in order to take into account their respective ways of organizing work. Together they offer models that allow us to focus on the ways in which data science resembles the tech nerds or the more anonymous learned occupations of lawyers and medical doctors without ignoring the concrete ways of organizing knowledge they represent. The specific association of principles of expert practice and substantive tech nerd characters begins with Bill Gates, who represents the organizational control of expert work. Informal expert movements have revealed more variability and complication, which we had also seen among technology hackers in substantive terms. Here Aaron Swartz and Linus Torvalds respectively represent two facets for technology expertise movements. Finally, the learned professions are widely salient but remain largely anonymous. Because Mills's central argument emphasizes the utility of this anonymity as a general role model for guiding occupational mobility, we consider Mills as a guardian and hence representative of this group. With these characters in mind we can more clearly anticipate the different specific principles associated with formal and informal expert groups in New York City's data technology scene.

Although these tech nerd characters do not enter this setting themselves, the types of expert practices they stand for unfold here as well. We can begin with the now dominant image of bureaucratically defined tasks in firms and other organizations. In the technology context, we have seen this through the role of Bill Gates and his letter to hobbyists. In the public accounts that undergird the following analysis, we would expect that speakers articulate the problems they are concerned with through the capabilities their organizations offer or the needs they face. In the bureaucratic division of labor that would entail specialization or standardization of tasks as part of a more comprehensive project. For other accounts that represent the expertise of the movement around Linus Torvald's Linux

community, we also expect accounts reflecting specialized tasks and problems. Contrary to the corporate accounts, here we expect references to independent identification of those specialized tasks from the side of the experts working on them, not the bureaucracy. In other words, while these tasks also address specialized problems in the larger Linux project, the contributors identify them directly, not the leaders of the community.

We expect other accounts to also indicate liberty in deciding on the tasks they embark on, though without integrating it into a larger project. This practice reflects the kind of projects Aaron Swartz initiated or contributed to. By this I mean projects that address a problem, such as designing code to comprehensively organize ubiquitous online news, or free access to legally public information. These projects are not necessarily limited to the efforts of a single hacker, but they neither rely on nor contribute to the specialized coordination of the work of a larger community, regardless whether it is open Linux or proprietary Microsoft.

Finally, it is less clear what kind of accounts we could expect that are indicative of the autonomous role model in today's technology context, which C. Wright Mills found lost long ago in learned occupations. It is clear that it should signal autonomous task selection independent of the organizational contexts they apply their skills in. We recognize these arrangements with heterogeneity of tasks familiar from Swartz coupled with more of the integration seen in both Torvalds and Gates.

3.4.4 Measurements

How does arcane data science knowledge gain lay salience? In order to answer this question we need to consider two problems. First, we need to understand how data science relates to both the mundane and practical contexts it is applied through and the arcane expertise it draws on. Such an analysis reveals the degree to which data science constitutes a set of distinct practices or must be seen as an extension of different kinds of existing ones. Second, we also need to consider the experience of defining data science in order to understand the processes underlying the structural positioning, whichever it turns out to be. This focus raises questions of how speakers articulate the novelty that is seen so widely as part of their work.

I position data science with respect to five substantive contexts, discussed in five chapters. Each of these moments considers both the structural position of the data nerds and the processes of defining the respective tasks and necessary expertise.

On one level I consider data science accounts across different contexts that we know impact expert work. Here I focus on identification mechanisms of nerds with this expertise. I begin with accounts that pertain to appropriate technology and organizational applications, then continue with a focus on project implementations and skill requirements, and end with views of community construction. The first two chapters consider data science in the jurisdiction of data problems and the boundaries they draw around them with respect to data technology and data in organizations. These questions constitute, in one view, a foundational problem of professional legitimacy. Especially chapter five, the second in part one, also begins to address another view by focusing on the formal and informal relationships shaping data science. Chapter six complements this analytical focus. It scrutinizes data science projects and skills in order to consider the degree to which data science expertise is shaped by substantive specialization and institutional contexts, both with respect to organizational and technical problems. While all factors discussed across the three chapters are associated with variation in data science expertise, consistent patterns emerge as well. In this respect we will particularly note the integration of analytical tactics and quantitative methods. Finally, chapters seven and eight shift our focus from data science relative to its social and technical contexts to data science itself. They consider variation across paths into data science and relationships of data science to existing disciplines and knowledge. This set of questions positions data science into the organizational context of modern technology and its applications. This positioning emerges from the accounts directly.

The qualitative design gives access to an interpretive analytical lever as well, which I refer to as “contours.” From that perspective, these accounts of data science across contexts reveal different rhetorical scripts for different purposes, or features that constitute some key contours of the data science thought community. As data science emerges, there is no good basis for expectations regarding its organizational role, except perhaps technical arguments. Chapter four, the first of part one, reveals instead analogies to substantive settings, including some very unlikely ones relative to data science, and in chapter five we find attempts that emphasize organizational complementarities more than the conflicts

one might expect for jurisdictional struggle with other functions that could provide similar services. In chapter six, we learn about improvising, instead of deliberate operation and adhering to authoritative guidance. Data nerds follow such tactics even in the technical areas they are more familiar with, and not only in the context of the many substantive problems they encounter without training that could have prepared them. Even when considering the group-level, in chapters seven and eight, where one might expect efforts of collective mobilizing and mutual recognition, we find emphases on chance and surprise over data science's salience, individually centered narratives and definitions of practice. This set of discoveries reveals a surprising rhetoric for an emerging expert group as the accounts come together through specific experiences rather than shared aims. I discuss these processes at the end of each chapter with all views laid out.

Overview

Pertaining to both public problems and individual opportunities, data science gives good reason to seek a better understanding of it. Because data science is both novel and large, analyzing it systematically requires overcoming significant uncertainty, and hence a conceptual and analytical apparatus as inclusive as the one I have presented here. In order to establish some clarity as to how the conceptual frame, empirical measures, and subsequent conclusions come together, let me just briefly anticipate some of the main processes we will see unfold in the empirical setting.

Among these different aspects, the organizational and conflict-based explanations, or Bill Gates and C. Wright Mills in more common terms, show little bearing on data science definitions, although they frequently enter accounts in initial attempts to shape them. Data science task definitions also remain consistent beyond direct relations among the leaders of the community, who are significant in Linus Torvalds's Linux project. In other words, while data science does rely on specific technological developments and organizational environments and recognizes elite members of the group that define its direction and purpose, the way data scientists apply their knowledge seems to remain largely independent of all those aspects. This barebones description recalls Aaron Swartz's projects and data science directly invoke his style of defining work. Settling on this role alone would fall short of acknowledging that data science also shows signs of the other technology nerds considered before.

This leads to the core argument that data nerds identify with the data science community through a rhetoric that is shaped by distinct technological applications and leads to distinct implications that others have no means to expect, and hence react to in the sense of circumventing the data science community. This argument explains how data science can simultaneously draw on arcane knowledge and address the public, thus gaining lay recognition. The technological details underlying data science ensure its autonomy of technical and organizational contexts as well as its independence of direct coordination among leading figures, while the community that shares this skill set jointly defines rhetorical tactics to articulate their utility to lay clients. This general summary also begins to address the concerns data science has initially confronted us with. It begins to indicate that it is necessary to understand data science independent of its organizational and institutional context in order to undercut its harmful invasion of public privacy and behavior, as well as for leveraging the individual opportunities in modern technology work it indeed seems to offer.

Identifying this argument from the substantive material and capturing the rhetoric requires to sort out some terminological confusions.

3.5 Jargon

All social groups and thought communities have their own stories and memories, which shape specific terms and references that make no sense to outsiders (Zerubavel 1997). Arcane knowledge of expert and professional groups, which we are concerned with here, consists of technical jargon by definition. It follows that this analysis of New York City's data science scene will entail some surprising encounters as well, both unexpected and unheard of. Because most accounts aim to address broader audiences anyway, these instances remain sufficiently rare as to leave the narrative flow unharmed. Where they remain, these instances also bear analytical leverage. As some terms slip into the speakers' accounts amid the public context they may not only confuse here but also puzzle the audiences at those events. These moments constitute essential observations in an analysis of the processes by which arcane knowledge gains lay salience. I nonetheless explain technical terms after data nerds introduce them, either in the text or in footnotes, in order to provide clarification in ways audience members could have asked for while readers cannot.

3.5.1 General terms

It is nevertheless helpful to establish some common ground. Some terms might see familiar, at first. Whenever speakers discuss “Python,” they refer to a programming language, not a snake. This language is comparable to “R,” which is both important here and more familiar to social scientists as alternative to the statistical software packages Stata or SPSS. To continue with animals, “Hive,” usually home to bees, in this context refers to another programming language specifically designed for handling large amounts of data with structures that deviate from the familiar tabular format of rows and columns. Adding to the confusion, Hive’s bee logo, which is plausible given the names common meaning, has the head of an elephant. This frightening combination of sting and trunk comes from its relation to “Hadoop.” Hadoop refers to a software framework for storing the data, which we can then access through Hive (and analyze in Python, or R). The elephant reference comes from the developer’s son’s toy elephant, named Hadoop. This personal story is known in the community. Finally, a third class of terms that neither invokes animals or invented words directly denotes technical meanings. One of the most frequently cited of this kind is “MapReduce.” It refers to a language that is composed of the two steps, map and reduce, with a framework for distributing analytical tasks across a cluster of machines. These terms, and others, will be defined as they appear in the following discussions of data problems.

3.5.2 Algorithms

There is a technical understanding of algorithms, and there is a way in which it is used. Sometimes the two overlap. Because we need to understand and analyze the accounts, and in order to remain consistent with the sociological literature (e.g., MacKenzie 2014, 2016), I focus here on what data nerds mean when they talk about algorithms. I have already introduced algorithms as complicated objects in the modern technology context. They enter this discussion, or the data nerd accounts, as simultaneously specific and general instances in which the above programming languages matter for practical outcomes. Technically, algorithms are independent of computer code. For instance, a Hadoop algorithm may perform a specific data extraction task, a Python algorithm may restructure data such that a more specialized algorithm can analyze it in R. Depending on the specific task, however, it may be easier to transform the data in R or do the analysis in Python. In other words, algorithms do not necessarily capture

the idea of a problem someone tries to solve with them (although in very standard instances, such as basic linear models in statistics, they do).

Let us consider this instance from the field. This account is from a representative of a service that provides price estimates of real estate on a large-scale basis. It recounts how they used to develop a statistical model in R, the statistical programming language. R is useful because many academics and other data analysts and programmers have contributed to its library of statistical models encoded in computer algorithms. Because of technological limitations, R cannot fit those models to the entire dataset of real estate observations this services included in reasonable time. Therefore, in the past, they would then hand that model over to software engineers for them to implement in a programming language with faster algorithms. This took several days, and increased the potential for mistakes in the processes of translating computer code from one language into another because the same commands are not available, rules differ, and so on. This processes changed, however, once computational power became less expensive. At that point, the entire translation process became unnecessary because faster machines could process the entire dataset even with the slower R algorithms. New technologies made questions of faster algorithms irrelevant. Although they become practically less important, they remain ubiquitous.

3.5.3 Science and data science

The Wikipedia definition of data science largely built on its relation to existing academic fields. Data science has practical consequences, however, and we accordingly moved the discussion quickly toward problems associated with work seen as data science. We ended the introductory discussion with different arguments for how to understand work, with a particular focus on applied professional work and expertise as well as on scientific work. We found that the literature is deeply divided on how science works. From which perspective should we then understand these accounts? One way of understanding science is as an institutional system. Another way is to see it as a series of interactions over arcane questions.

We primarily think of it in the terms data nerds describe it in. Sociologists who define these views are used to thinking about people and how they interact. Data scientists may not, so that considering what they see in science, and science in them, offers analytical leverage because it provides a basis for comparisons of identification mechanisms. This dual role of science as a novel case and familiar empirical

context of work introduces complications because it can be confusing which science we talk about, the one data nerds mean or the one we know. At the same time, it is useful because it challenges to pose precisely this question of whether the way data nerds use it differs from other understandings, how it differs, and to what effect.

This approach promises benefits for both sides. For data science, it seems important to understand how the ideas associated with this label relates to the ideas and activities that have defined this label originally. For the sciences, it should be relevant to consider the ideas and activities gaining so much salience with the same label that leads to so little for most others.

4 Technology

In order to discern the sources of the data scientist's salience, we need to understand the world she operates in. A clearer sense of quantitative data processing in the context of advanced computational methods and large-scale applications provides such a basis in terms of the degree to which technology defines the professional role of the data scientist. This basis allows us to analyze in further steps social factors shaping it and finally, data nerds' own efforts.

Modern data technologies are obscure. This chapter aims to render the technological context of applied data analysis in similarly tangible terms as we rely on to understand other instances of social activity in worksites. For law, for instance, most of us know that litigation constitutes one of the central problems lawyers address. Studies can therefore focus their contribution on specific problems in this area without introducing the court system more broadly (e.g., Sandefur 2015),[44] just like analyses of nuances in medical knowledge get by without extensive descriptions of hospitals or medical training (Menchik 2014). Other cases require more context and description. For instance, we need to consider the division of responsibilities at NASA, including specific engineering arrangements, in order to understand the organizational failure leading to the Challenger space shuttle explosion (Vaughan 1986), and the office arrangement of new media startups in order to understand how different orders of worth organize web innovation (Stark 2009). These instances involve complications that the average reader or observer would not immediately have knowledge of. They can still be made sufficiently clear such that they lead to a more comprehensive and robust understanding of the underlying social dynamics.

Accordingly, this chapter begins with considering the problems which data science is seen to resolve, in technological terms. That way we can gain an understanding of these problems prior to considering data scientists' arguments with respect to other aspects of their work, skills and expertise.

This exercise begins to address the main concerns data science evokes. We have seen many technologies rise and put individuals or society at risk. The most prominent instances are those related to large engineering projects as well as food and health (Wynne 1992, Shwed and Bearman 2010). These instances and their harmful consequences often result from collective organization, not the technology

[44] Not so many know that only a small fraction of lawyers works in litigation (Howarth 2013). But this is unimportant for studies of litigation as they can still build on common knowledge.

alone. At the same time, we also know that technology shapes how social interactions unfold, again with disastrous consequences (Weick 1990). Given such different effects, it is necessary to study directly the degree to which the technology data nerds work with shapes their tasks, and hence their public impact.

The technological context also pertains to the question of individual opportunities for similar reasons. Technological change and the automation it facilitates threaten opportunities for human labor. Data science poses no such threat and this demographic question is not my focus here. The accounts from the introduction have instead also suggested qualitative variation in how data nerds coordinate their work. The main question that emerges in this respect is concerned with the utility of existing knowledge amid the changing context of modern data technologies. Herein, we need to focus not so much on how much work computers take away from people, but rather how nerds use technology in different ways for different applications. Considering the empirical context helps to think of these implications in concrete terms.

What do we need to pay attention to at these data and technology events in New York City? To recall, this specific focus on the technological context contributes to our analytical strategy of identifying the contours of thought communities in data problems and how experts construct them as they articulate their distinctive contributions in public. As we are just setting out here in the data jurisdiction, we have little basis for formulating specific expectations. For instance, we can recall the existing literature, which suggested that data experts in the past have remained deeply embedded in the institutional context generating the data they analyzed. This image is at odds, however, even with the most impressionistic observations of how data science has been recognized to address modern data problems across a range of substantive areas. As I have argued above, we gain more leverage from other groups of modern technology nerds and experts from the introduction because of the organizational principles they represent. While they differ from data science in their substantive concerns and appearance, they provide some direction for the kind of observations that are relevant in this substantive setting. At this point, they lead to several different, but nonetheless specific expectations as to how data nerds may define their expertise.

The technical context of modern data processing has some familiar features that likely impact expertise definition. Historical origins lead us to take into account that data processing has long

constituted the center of canonical corporations. Here we can think of IBM's work for instance, which Mills has noted himself (Mills 1951). It follows that even if alternative processes have contributed to shaping this context more recently, it seems unlikely that the resources and competencies IBM and other corporations have built up historically remain without effect today. In other words, we can expect accounts that resonate with this bureaucratic definition of work, which we have started to consider Bill Gates as a representative of for this analytical purpose.

Moreover, the technological basis for data science requires sufficient financial resources to fund servers, processors, and so on. There seem to be no strong grounds for the independent projects Linus Torvalds stands for, in which a large community of computer nerds coordinates relatively informally around alternative solutions to data technology.[45] When it comes to work, financial relations have the power to script coordination formally. That power holds to some degree at least for the data technology context. What is more, this context offers similarly few reasons for expecting evidence of the anonymous professional Mills described. This is because the specialization that is often required in a large-scale effort as modern data technologies. These specializations suppress and anonymize positions they entail.

Finally, and contrary, the interplay of corporations and hackers, like Aaron Swartz in the technology setting, has shown the constant opportunity for erratic challenges to the corporate hegemony. We can therefore expect them here as well.

I begin with considering the salience of these arrangements to data nerds in the imagery of "big data," a term that has increasingly gained acceptance in spite of wide criticism for and acknowledgement of its technical irrelevance. We see how those involved in problems concerned with large-scale data on a technical level share that discomfort, but at least for the purpose of public discourse, continue to engage with the term. I next consider the data jurisdiction from the perspective of the technologies designed to address problems of storage and logistics, referred to as "the stack," and from the perspective of "tools" necessary to utilize the data. Finally, the various problems are systematically seen to require "miracles," or the role of the data scientist.

[45] Although there is some as soon as the hardware is in place.

4.1 Big Data

Experts who work with large-scale data respond to the ubiquity of “big data” in public discourse with explaining how they relate to it instead of directly explaining the importance of the problem they address in its own terms. They often take an additional step and explain big data with respect to well-known and obvious problems, which are sometimes not technical at all, however. The technical field of modern data processing, in turn, is made tangible to their audiences through these rhetorical devices, as can be seen in the following quotes.

A good place to start may be the role of President Obama. Here in Steven’s presentation he serves as a baseline to illustrate the importance of big data:

> But first, the obligatory comparison over time of search results. This is comparing “big data” to “Barack Obama” over the last year. And you know, a year ago big data and Barack Obama were roughly the same popularity. Big data is growing in popularity, Obama is staying at roughly the same popularity. And despite it being a presidential election year, big data is growing in popularity and Barack Obama is staying at roughly the same popularity. Big data is very important right now. And looking at massive amounts of data is very important, …

Stephen anticipates skepticism. He tries to persuade the audience to recognize big data as a general phenomenon, for which he refers to evidence from “obligatory” search results. It is not clear why they are obligatory, but the connotation that they are familiar seems plausible in an age of Internet search. Although it is a technological problem, Stephen presents it here with reference to political discourse. While that demonstrates its public importance, that by itself does not warrant professional attention, at least as the immediately following justification suggests:

> … and one of our customers is actually the Barack Obama campaign, through the startup “optimizely”—how many of you have heard of optimizely?—so, for those of you who haven’t, they allow people to do A/B testing, in the cloud, so, attach their JavaScript bug to your webpage, go to their interface, and you can tweak a web page, and it changes for 10 percent of the viewers of the page. They are using [our tool] to figure out exactly the effects of those changes.

It’s only in this second step that Stephen invokes his own project. Stephen repeats the same reference he previously used for positioning big data as a larger phenomenon. The Obama campaign is important enough to serve as a reference point, but his project is really part of big data, something that is more important still as the audience can see in a graph on the slide that supplements the comments above. At the same time, the importance of big data as a problem Stephen’s tool helps solve has not been central to their strategy in the past, as a note by one of their investors reveals at a much later event. That investor recalls how Stephen's project only embraced the idea as it gained popularity.

Interesting and important to note is the level of specificity with which Stephen describes the project that connects his own to the Obama campaign. A rhetorical move like invoking the president gets supplemented with technicality. Stephen's own specialized project, on this basis invoking bureaucratic definitions of tasks, requires familiarity with some technical details of another specialized project. And this is just one application for this tool. Just imagine a specialized case, as would be familiar in the bureaucratic context. For instance, Stephen's strategy translates into explaining to someone a word processing software package with reference to the role of a speechwriter who works for President Obama. While just Obama might be useful marketing, the speechwriter detail is unnecessary for seeing the utility of word processing. It would be much easier to express it in general terms of writing. Perhaps Stephen could also explain this software in general terms, but he chooses the detour over explaining an entirely different project. For Stephen, Obama, whom everyone knows, through his campaign and its specific technology strategies illustrate a problem almost no one would be aware of.

Aside from the suspicious prominence and lack of technical precision of big data, which Obama here helped to overshadow, the following quote indicates the perception of a different threat:

> We decided to not focus on building the new cool application. We decided to do something much less sexy, but just as important, because we recognize that everyone who was going to do this cool stuff will need data, and that obtaining data would not be the easiest thing in the world, but it would be a very necessary thing.

Big data refers to the larger trend. Here we see a defense against a potential hype around the practice of data analysis specifically. This can be seen in the reference to "sexy" applications. Although unrelated to technology, popular media frequently describes with this attribute the novel type of professional expertise. In other words, the odd use of the word sexy here signals a community that has added a connotation to "sexy" such that it helps this speaker explain his project.

Thus, being unsexy should not imply unimportant, as one might think based on those other descriptions of data applications, as Chris, this speaker, quickly continues to explain:

> I think the key thing to know about who is using our data, I describe them as they are oceanographers, not fly fishers. These are business that are spending 100s of thousands or often times millions of dollars, building business insights and analytics on top of this data, they need this data to be full coverage, they need it to be supported, they need it to be reliable, they need it redundant, their needs are very different form the needs of a consumer who is pulling stuff off the [inaudible] for example.

Let's accept the implicit assumption that everybody in the audience knows fly fishers in order to discern Chris's strategy. Doing so, this framing translates into a comparison of nerds specialized on

hidden corners of the “water jurisdiction” to oceanographers, who by definition deal with the problems concerning almost everyone, or a comparison of hobbyists to professionals.

Consistent with the others, Chris foregoes a specific example. Chris goes one step further than Stephen by not just relying on a more widely shared redefinition of the word sexy. He instead redefines vocabulary from scratch on the basis of an abstract group. This way he can leverage the implications associated with new terminology for his otherwise arcane problem.

Without introducing these analogies, the distinction between individual consumers and corporate customers would not surprise anyone. This move thus suggests either a novice in marketing, who was more surprised about the different needs than others would be, or implications regarding the product. Since specific examples of what the larger customers require follow, Chris seems to leverage the analogy for specific project features, and not the overall distinction of clients. In other words, although everybody can see social data, there are problems with it that only emerge to large-scale users, who therefore require all the services Chris enumerates. We thus see yet again the design of a rhetorical framework for articulating a problem that is too arcane to see directly, and would make solutions therefore seem irrelevant.

Once again, the mere presence of a larger trend that we see this way, in the division of labor between professionals and hobbyists here, is not considered to warrant the audience’s attention. Chris therefore continues:

> So, the interesting story behind this is that the world kind of changed in terms of social data about a year and a half ago because Twitter recognized that there was this world of oceanographers that needed data access that was different form the consumer access needed from the public API. And they decided to actually launch commercial, I call it commercial grade products that had these attributes, and make it available for business. Now it is a side note in history that they chose [our project] to be the first partner to bring this stuff to market, important side note for us. … But the point tonight is that it is not just about Twitter. And sometimes when you think about social data and interesting things that are happening in social data, people go to Twitter, and I would argue that is because they have had for a long time the best access to data, …

In the words of his own comparison, Chris’ project makes the ocean accessible to oceanographers in the first place. That seems quite significant, more so than a technical explanation that involves APIs, data streams, cleaning, and other obscure specificities, may indicate. We can extend his comparison to see this effect more clearly. From a Western-centric perspective, Chris tells the audience also that Twitter is really just a small part of the ocean we supposedly should have thought of it as because of its sheer

size a common observer sees in it. This additional discovery follows from Chris' note on moving beyond Twitter in his project and at as part of this presentation.

Deciphering these accounts leads us into interesting territories. We know that lawyers also make unintuitive comparisons, but those remain between new cases and older court decisions (Stinchcombe 2001). Expertise often focuses on knowledge in small groups, which limit the possible scope as well. Here we see indication of both. The unintuitive comparisons neither rely on formal rules, as law would, nor on esoteric references that require prolonged collaborations among peers. If such processes seem less pertinent here, should we then understand these descriptions as artifacts of the technological objects they refer to?

Both speakers' arguments were based on convenient references with Obama and Twitter directly relating to their respective projects. Though Mike, an investor, makes a similar argument from a broader perspective:

> Sure, let's take a step back and take a look at a trend that's happening right now and that I think is really interesting. So think pre 2000, or look at 1900 to 2000, and every technology that came out, I've had a debate with people on this but, every technology that came out, even if it was invented by someone else, it was operationalized by the military or some government entity. So think airplanes, think space travel with computers, radar, the internet with DARPA net, everything that was interesting that came out was really operationalized—the transistor, was invented at Bell labs, but then the space program drove the operationalization of transistors before it went into consumer space. But then came 2003, 2004 and Google came out with their, you know, their Bigtable MapReduce papers. And what we have seen since then, with prism, is that the government is now doing the opposite, the government is now taking technology that was built for Google, Facebook and Yahoo!, and they're pulling that out and they're applying that to the NSA and the CIA.

If "big data" and its "sexy" applications might feel like a suspicious hype, this is not because they are unsubstantiated, but because it is indeed a recent phenomenon. This argument may help to contextualize what the audience experiences and to demonstrate the significance of this trend, but Mike anticipates new concerns and asks his audience to …

> … forget what you think about what that means as a citizen and whether or not you're comfortable with it, and look at this from a technology trend perspective, we have never seen this before in the modern technology era, and I think this is really really interesting.

To be sure, the same point can be made with economic instead of political reference:

> [Y]ou look at the last 15 years, and I mentioned Google and Facebook already, but they kind of created $100 billion of equity value respective out of thin air, right, historically creating $100 billion in, of equity value is pretty hard to do, right, if you build a company and you did an incredibly good job over a very long period of time, you created a $100 billion company, right, Microsoft was not a $100 billion company at IPO, ahm, but Facebook and Google were, and what fundamentally those two companies were doing that was different from everybody else, and I think it comes back to the data they were able to monetize. …

The redefinition of terms we have seen in the other accounts is missing in this one. Mike instead focuses on a rock solid historical reconstruction.[46] We can also look at these accounts from a different perspective, where we consider the role it plays beyond providing context. We then note that this effort provides arguably distinctive meaning for the data jurisdiction itself instead of just specific tasks in it. Mike thereby extends the previous accounts on the basis of efforts to render data problems more meaningful, indicating a more conventional side. It is also non-technical and our thought community focus gains traction here. It helps us to recognize the emphasis on economic utility, technology and government, as the speakers' different grounds for distinguishing modern data technology from earlier technologies. Instead of teaching the technological basis of these relevant problems, speakers translate them into terms the public is comfortable with already.

Considering these accounts jointly, the public debate around big data creates problems and opportunities for those engaging with large-scale data professionally. Many become aware of the jurisdiction but could easily doubt that these problems are indeed of legitimate professional concern. Amid the various rhetorical strategies invoking the Obama campaign and other problems with no direct relation to data we begin to see a world that is relatively elusive to most, both because of its recency, but also because it requires technology that the public rarely interfaces with directly. Through large-scale data this arcane technological world becomes directly relevant, as large-scale data is relevant to the public as the basis for many modern problems ranging from gossip to political engagement.

Other presenters suspect that technically minded members of the audience may not need this level of illustration. That some are uneasy with the label just reflects their familiarity with large-scale data before it gained attention under the new term. Even this group may not be familiar with all of its facets, however, which we can see as speakers provide more detail on the technical specifications of solutions to data-related problems, which I consider next.

4.2 The Stack

The holistic trends several speakers just vividly described to the audience break down into specific challenges after all, each with its distinct implications. With respect to our interest in what defines data

[46] The historical significance of DARPA in US innovations is also documented elsewhere (Weiss 2014). The overall accuracy of this recollection is unimportant for our purposes, however.

science work, we need to consider that specialization and compartmentalization of work might more readily emerge at this higher level of technical granularity.

This begins with the data points themselves, the observations that as a result of their abundance aggregate into big data, as Michael explains:

> There is this digital mirror universe that is all around us, of tweets, and check-ins and point-of-sale devices and transit events, and all these things that are happening, this internet of things is pulsing events into the cloud, and for a long time many of those events were invisible …

As we could have suspected from the interesting territory we were led to above, we have in fact entered a different universe, in their eyes. We have heard about Twitter before with respect to its overall relevance in the data context. Here we learn about specific activities that generate the ocean we were introduced to in the previous section, or the cloud in this case. And the cloud is a less idiosyncratic description of the aggregate storage of information on servers accessed through the internet, instead of hard drives physically near to the producer of the information or data. Michael introduces us to the additional level of specificity through the imagery of the "pulse," well known as a biological process but usually irrelevant for technological devices. In short, we have another instance of redefining a familiar word, as "sexy" or "oceanographers" before.

On this basis we see the additional problem arising in the attempts to make the cloud more tangible in the minds of the audience:

> … I think what we are seeing happening increasingly is this digital world of events, this pulsating digital nervous system of the planet, is actually being made visible, and there are new technologies that can help us create value out of those streams.

All the time this speaker shows a slide with a shape resembling North America as we know it from maps, except here it's contours emerge from plotted points representing geo-coded activities, and without the formal boundaries a map typically has. The digital traces Americans emit become directly visible, though not in all of the forms a naive observer could have expected. The image Michael shows the audience neither resembles a cloud nor a pulse, though it does perhaps invoke ideas of a nervous system.

Here we find what other instances of sexiness and oceanographers have led us to suspect; these descriptions and explanations do not aim for technical accuracy, even in these narrower questions. Whether this renders them useless or enhances their salience we cannot tell here. More importantly for understanding the contours of an emerging thought community is whether we see these practices across

presenters, and in how far they resemble each other. In this respect we can consider that they are technically literate and thus, step outside of their comfort zone by engaging in these figurative arguments.

Meanwhile, getting to the point of engaging with the digital pulse is the real problem, as another speaker no less evocatively reminds us of:

> … talking to an audience of data analysts and data scientists and data professionals, you guys all know that the analysis, the science, that's the fun part, but the data acquisition, that's the painful part. There is no shortage of data out there, right, there is millions of data sets on the internet, but the problem is it is so fragmented, it takes hours of your time finding it, cleaning it, downloading it, merging it, synchronizing it, and then do it all over again the next time you have to run some data. And you know, 80 percent of your time and probably 99 percent of the blood, sweat and tears is just getting good quality data into your system.

So, what is the system, and how come it creates so many tedious and complicated steps?

> [data] munging is the coal mining of the information age, it is the dirty work that has to get done before we get data that can fuel our analytics. So, right now the current data stack is kind of broken, it's sort of a Frankenstack, you've got Hadoop at the bottom, not even sure we need Hadoop, if we are doing continuous processing maybe we can just do it in memory and get rid of that.

The problems are so serious that even the "bottom," the foundation of the technological infrastructure system that has been created over the last few years, is not settled. And it goes on from here:

> Storage, you know, storage is broken, that's we have all these legacy databases just don't scale. So we have the rise of NoSQL and other storage options. Finally, on the analytics layer it is still very custom, you have some things out of the box. But let's face it, if you're doing analytics at your company you are rarely just buying SAS enterprise miner and plugging it in.

The problems change as we move to higher levels of the stack, and with that the narrative. The bottom seems set, albeit in an unsatisfactory manner. The analysis of the data stored in the bottom, on the other hand, requires customary work.

The coal mining takes place in the Frankenstack. We also see that these ideas with no relation to one another in the worlds they are taken from, the common stock of knowledge, can take on a very specific meaning here. The work is dirty, as coal mining, because much is available, while not buried underground, hidden online in places such that one has to dig to find it. Once we have it, we process it with an arrangement of tools, such as Hadoop and NoSQL, which like Frankenstein do come jointly by definition but are nevertheless made to hold together as an integrated object. Although this account introduces very specific tools, it also invokes a common reference to Frankenstein.

Such references would certainly not find their way into purely technical discussions. The combination nevertheless indicates some uncertainty over the degree of appropriate technicality and

widely accessible illustrations, in particular with respect to the combination of specific tools. To be sure, problems with the stack not only affect experts familiar with the details, because …

> … visualization, and this is where a lot of focus has been, because you can see it, right, and there is lots of startups that are creating cool visualizations, you know, visualizations on the web. Unfortunately that's not enough, it is very superficial. … Your visualization can only, your experience can only be as fast as the engine that is powering it, right. You can have a Ferrari chassis, but if you don't have a Ferrari engine, it's really not much of a Ferrari.

Now things make sense. Fixing the "Frankenstack" addresses the visualization problem; a better stack ensures that the digital pulse of the world becomes visible to, and benefits users and consumers. But this does not solve the problem of custom analyses, the importance of which we see next.

And just to take stock before moving on, the focus on the technical problems surrounding data science has led us to learn about a series of familiar problems, but seemingly unrelated to data expertise, and even one another. Interestingly, we see these attempts to connect data problems to components of other, presumably very familiar, narratives independent of the problem's proximity to the lay user. Instead of merely describing an arcane problem in common terms, such as sexy technology applications, or data streams and the pulse, the idea of a chassis and an engine emphasizes the relationship between two arcane objects of different degrees of relevance. Although technical concerns vary with their relevance to users and clients, this distinction is irrelevant for the nerds themselves. Their expertise seems systematically applicable.

Similar to experts who work out problems together with lay clients, here we see experts articulating their competence and utility in a supposedly broadly relevant way. Instead of working with clients on their specialized problems, however, these nerds rely on common ideas. Consistent with this strategy, we also see the generally applicable technical basis in the formal tools, but no evidence of a formal framework for applying them as procedural law and blueprints would offer for litigators and for architects.

The solution is not just technical. Even after fixing the stack, at least part of it, experts struggle with anticipating the type and extent of analytical problems and opportunities, as Bobby's experiences illustrate:

> [Existing logging systems] were all kind of, somewhat of broken, and they weren't scaling very fast. It would actually have been a lot easier to fix them all, you know, it would have been a couple of weeks of work. But instead what we decided to embark on to build something general … Within a few months of launching this, we had over a hundred categories being logged in this. People just came out of the woodwork. All kinds of things we never expected, we'd have never dreamed of, that people found.

Here we learn that fixing the stack not only falls short of addressing problems with custom data analyses, it enhances them. By this I mean that aside from the technological problems, the mere possession of data emerges as a problem itself. It is impossible for lay people to imagine the complexity of building a new logging system in just a couple weeks. Having people came out of the "woodwork" with many good ideas, on the other hand, begins to index its significance in Bobby's experience. This framing becomes all the more relevant with respect to our question of what provides the basis for these connections between technical problems and common references. We learn here about a problem in a bureaucratic context where it is clear from the language that this was not the result of careful planning, if we once again leave aside the technical accuracy and specificity.

We can also see why speakers have such a hard time articulating solutions in specific terms. Just to continue with some of the evocative images speakers have provided us with, more data not only creates opportunities for "fly fishers" to find new fish hiding places, it addresses central problems ensuring the balance of the oceans, as Bobby explains further that:

> … there's always sporadic complaints [from users] about log-ins, you know, somebody getting logged out on the [social network] site, and nobody could reproduce it and nobody actually believed that like this was really broken, because this is like, again, this is core to our, you know, of course this is right. And then one day there was an engineer who got logged out, and he was like 'I know that was not okay, that was a bug' and so he wrote this ridiculously verbose log of everything that ever had to do with log events and with this huge amount of context and sat there for like two weeks pouring and he found no less than 12 bugs in the log-in system.

We can begin to imagine the complication here when even with an intact stack it takes an engineer two weeks to at least map out this custom problem. Bobby built the initial tool, which the engineer then built an application on. The rhetoric around responses Bobby did not anticipate rules out a formal, bureaucratic task description for this problem. The narrative on how the engineer had the idea that there might be a problem at all supports this as well. It also rather rejects than supports the significance of close collaboration between them as part of this development. The tool with which the engineer found bugs would not need to become a specialized component of Bobby's tool, or the larger software, in the way Linus Torvalds integrates applications into the Linux system. The way this project unfolded maps onto the organizational arrangement of contact improvisation.

Specifically because we see few indicators of familiar forms of coordination, these accounts invite to focus on community identification mechanisms, without a standard process in mind. Speakers outline the tasks of a jurisdiction of systematically collecting data and their utility in terms of using that to detect

bugs. These tasks involve specialization around a system that is useful for others, although on an abstract level that merely provides the foundation for a specific application. The idea of contributing to a community is important not on the basis of agreement over certain tools and problems the community has. Self-identification here involves practices that circumvent the absence of such standards.

For considering the utility of such a conclusion, we need to remember that the complexity of working with the data just constitutes one side. The challenges begin on a much more basic level. It is uncertain how many more problems like this are hidden because even engineers have trouble estimating how much data the digital pulse they are responsible for facilitating creates when made visible, or recorded as data:

> So the other thing that happened, once we started the logging system some people freaked out, you know, all this data, this is like, you know, how can you do this, there is all this data, and I was like this is awesome, there is all this data, ahm, but it was a, sort of a constant battle with people who were afraid that we, ah, this sort of slippery slope fear that like once you are going to make it possible for people to log, they are just going to log stuff. … And we found an interesting thing, that you couldn't just go an ask people how much, you know, you'd ask them 'oh do you know how much data this is going to be and I'm going to plan for that' or you say 'do you think that's a lot,' like nobody knows, even if it is like, you know, you are logging 1.21 jigabytes, or something, and I was like, they don't know what those numbers mean, they don't know what it costs, they don't know what the value is ….

It is easy enough to imagine that such uncertainty indeed constitutes a significant problem, especially in a formal organization. That people "freaked out" indicates their responsibility for a process they had no good understanding of in this respect, while that those who participated "don't know" to what extent indicates the limitations also of informal communication for this problem. We learn about this uncertainty here without the kind of references from before, to common if unrelated imagery and ideas. Instead we are told anecdotes. Yet, just like Frankenstein and Ferraris constitute common knowledge, so is the idea of technological disruption of orderly organizational arrangements. In other words, although this account resorts to a different presentation style, the rhetoric resembles that of earlier accounts in that problems associated with modern data are expressed in common knowledge and experiences. This moment of overcoming the specific tension between technology and planning reveals the kind of contour line we have seen emerge here in modern data problems more clearly as it rejects a familiar way of addressing uncertainty.

This account also reveals a deeper, and potentially systematic basis of the uncertainty we have seen repeatedly. The general surprise about the unpredictable size of data generated by social behavior seems curious from some perspectives. Despite common intuition that may suggest otherwise, the

complexity of social behavior has by now been broadly accepted, seen in the ongoing existence of the sociological discipline. It follows that since our understanding of social behavior is incomplete, our understanding of data of social behavior might be incomplete as well. Machines, on the other hand, are designed by humans, and, as a result of that, one might expect that data on their behavior is easier to anticipate. In the last account by an engineer, however, we see quite clearly that there was no understanding of the scale or scope of machine-recorded data as well. While this might be of little surprise to sociologists of technology, it was worth recounting for an engineer. It follows that much of the uncertainty seems to be intimately associated with the data, independent from technology.

We have already seen that big data, or large-scale data, is related to modern technology as it follows from storage infrastructures that have become available at lower costs and with greater capacity than they were just a decade ago. It has an independent quality as well, as solving remaining technology problems around ensuring the availability of that data does not also address problems pertaining to its utility. This can be seen in the struggle around data analysis throughout various instances discussing them. If fixing the stack cannot solve data analysis, what can?

4.3 Toolbox

Ubiquitous data recording, particularly of digital traces, entails that the data comes in great diversity and with many problems. While the stack ensures the collection, availability and processing of data, it does not clean, shape and analyze it. Here we follow the steps required for analyzing data, beginning with structuring it in ways that reveal its utility, taking stock of its content and then more abstract analytical tasks. This shift in focus leads us to consider in yet another technical setting the kinds of strategies data nerds describe their expertise with.

4.3.1 Structuring

We can begin to see the minute detail of problems with utilizing data in the following presentation, a live demo, of a tool to address them:

> So I looked at this data, I said it's got three columns, right, it's got, what has it got in these columns, we've got business ID, date and description. And so, first of all let's strip off the quotes. Well, you can write a script to that, that'd be kind of annoying. Why don't I just highlight a quote, and hopefully someone will get the idea that I want to get rid of quotes. So the first suggestion it gives is '`replace in this column at the beginning and the end, get rid of the quotes and replace them with nothing`,' the second suggestion is how about '`replacing the quotes in all the columns,`' and I like that one so I'll take it. So very much like Google there is a ranked list of probable things you want to do based on your

> interaction with the data. And you can preview them and see which ones you want. So you can read this command, which is in a reasonable language, or just look at the output and see what you get.

Joe here describes his tool for working with data with reference to the idea behind Google's commonly familiar predictive search—where Google suggests search queries as soon as a user begins typing. Just like Google suggests "definition" when I type "sociologist" or "search" when I begin with "lawyer,"[47] Joe's tool completes "quotes" with "replace" and "all columns." While perhaps appearing pedantic, this account and the detail in it reveals some of the substantively trivial yet technically intricate features of data structures, such as something as obvious as quotation marks around textual entries.

Moreover, the reference to Google returns us once again to the style in which other speakers already associated arcane data problems with common knowledge. While theirs were mostly illustrative references, here we find a case that implements the widely familiar intuition of modern Internet search for arcane data analysis. Solutions even to the technical data problems are uncertain for those with the necessary expertise. The common references with which they describe their utility are equally useful for them to navigate those uncertain problems. For considering the background of these references, and thus the coordination patterns they reflect, we can consider that it suffices to know of an idea to implement it, such as Google's search intuition. There could be a direct relation between Joe and Google, but since that would have added status to his project, it seems unlikely to not mention it in such a setting.

I should also consider that Joe is one of the very few academics speaking at these events about proprietary projects. While we have discussed the mechanisms defining specialized problems in academia in chapter two, this entirely applied purpose here suggests that they have little bearing on this idea as well. In short, Google's salience helps Joe design his own application without direct coordination.

This strategy addresses some problems with data, but not others. For instance, such specific problems can be anticipated for these generalized solutions. Quotation marks in columns are sufficiently frequent and problematic for enough users to benefit from a tool that anticipates them. As soon as we consider that this example worked off of spreadsheet data we must also recall that at this level of detail such a solution does not address the range of problems associated with varying data formats, to name just one other mundane and easily visible problem here. Along these lines let's specifically recall how at the end of the previous section we have seen that much important data is "logged" in computer systems.

[47] These search returns may be specific to my Google profile, and hence look different for others who try them, or myself in a different time and place.

This means it is not available in clean columns and this tool would not help so easily. Joe's perspective indicates that it is considered significant enough to serve as demo topic. This provides the alternative to the "Obama strategy" from the first section. Stephen's tool addresses problems few have, hence he cites Obama, whom everyone knows. Joe addresses a problem many have; hence he just focuses on it. As we have shifted focus since then, these different approaches reflect different specializations within the data jurisdiction.

Many recognize these specialized problems as important and propose integrated solutions. Yet, others have a radically different perspective and insist on tools that, while not resolving a specific set of problems completely, are capable of making progress on a broad range of problems. Where Joe turns to Google for inspiration, the other perspective considers the command line. The command line refers to an interface for working with a computer through commands, not unlike Stata for data in its design, but more general in its applications. The command line is directly and intimately related to the history of modern computing. Until the invention of the mouse and graphical user interfaces for navigating software, it was common to type the commands directly. This skill and style of work has been lost for common users, most significantly with Bill Gates's windows system. It has prevailed, however, amongst the groups of the two other tech nerd characters from the introduction. Hackers of both Linus Torvalds's and Aaron Swartz's kind heavily rely on the command line until today. In other words, while its use suggests the prevalence of autonomous expertise over planned bureaucracies, it does not by itself predict the role of Torvalds's heterarchical organization or Swartz's contact improvisation. Let us turn to specific accounts instead.

Jeroen, who presents his book project on these kinds of strategies, argues for the utility of this approach in the following way:

> Everything you do on a command line can be automated, can be scripted. This is very different from a GUI, a graphical user interface, where you have to drag and drop and navigate through menus and click buttons yourself, this is very manual. Linux, or UNIX in general, on the other hand, allow you to automate everything, and I think that is a big plus, especially when you want to repeat things, right, if you have lots of data to work with and you want to apply a certain action a lot of times then this is a good thing.

The command line is different from a GUI; hence it directly opposes Joe's solution. This is manual work, without drop-down menus to navigate, or search prediction that anticipate subsequent steps. Jeroen draws boundaries around the data analysis problem differently, as the following quote shows.

> HTML, as you know, is a form of XML and the tool `xml2json`, well as the name implies, converts xml data to JSON data. And, well, the output is here on this page. "`jq`" is a very interesting command, we are now here

> just using it to display our data in a nicer … way, it is a relatively new tool, and if you work with JSON data, I can really recommend it, it's a wonderful—it is like grep for JSON.

Jeroen mirrors Joe's rhetoric, but opposes his practical approach. He compares the tool he proposes to re-organize data from one format into another to an analogous one, called "`grep`." The utility of this comparison rests on the familiarity of the audience with the "grep" tool, which others might know from text processing applications. In other words, `grep` is to the technical audience what Google is to the broader audience. It is almost impossible for a lay audience to recognize this meaning. But this is a technical audience, and Jeroen invokes an idea they might be familiar with independent of modern data analysis, text processing, in the unstructured data context they might not know well. Contours of modern data problems emerge therefore consistently in common as well as in arcane terms, although in different shapes.

Indeed, just like Joe articulated the utility of his tool in the context of a common example (of San Francisco restaurant violations), so Jeroen continues with a specific, though arcane case as well:

> And what you see here is, well, we give it one parameter, which is sort of an expression, given to `jq` on how to transform the JSON data that we have into a different format. Basically all that we do is that we are specifying four new—variables—in our JSON dictionary per row, namely country, border, surface and ratio, and what we are then saying is that it should have the values of these `tds`, `tds` are the cells in our rows, and we can index them with one, two, three and four. It is a bit cumbersome, but it really, as you can see now, we really have that table into a structured format that we can work with.

This interaction with data differs significantly from the earlier one. The arguably more sophisticated tool Joe presented shows the distribution of each variable in histogram form and reflects changes, for example when ignoring outlying data points. The command line, on the contrary, shows nothing without instructions, no distribution, no variable names, no selection of the type of values they may contain. It resembles Stata in some way, except that it is not even designed to primarily deal with data and thus does not narrow the range of commands to this context. Engaging with data therefore requires one to anticipate properties of the data that other tools anticipate for one.

These two perspectives interpret the problems related to data differently, seen in the respective tools they propose. Joe's tool helps to correct different problems in a specific data structure, whereas Jeroen's tool helps transform different types of data into a specific structure. One promotes the type of highly specialized tools Bill Gates represents for our analytical purposes, and the other a tool resembling Aaron Swartz's projects in its applicability to many unrelated problems, although it could also constitute a specialized instance within Linus Torvalds's Linux heterarchy. The specific settings help to contextualize

these perspectives further. Joe addresses an audience generally concerned with data, Jeroen speaks to technical data nerds. Although they describe the problems differently, both are data problems. Joe indeed points out that his tool is compatible with JSON. What is a side note for Joe is Jeroen‘s main focus. Data is thus organized in different specializations, which nevertheless overlap, if marginally. On a subtler account, we can see that they articulate their opposing views in similar ways by invoking at least supposedly more familiar references. I turn to this strategy at the end of the chapter.

We also see here another instance of a distinction that already emerged in the previous section. Specific data problems contrast to variable data structures. In the previous section problems of clean stack designs contrasted to problems requiring custom data analyses. This pattern testifies to the importance of this opposition. The two, however, are not necessarily linked in obvious ways; those nerds concerned with neat stack designs may reject graphical interfaces for data analysis. The division reveals an instance in which technological concerns generate lines of conflict among the groups aiming to address them, as we have seen in the historical opposition between Bill Gates’s bureaucratic definition of work and heterarchical and improvised definitions of hackers and hobbyists. As both perspectives can be seen here to pertain to data problems, this division itself informs our understanding of data science expertise.

4.3.2 Counting

Although clean structures are important, the range of problems that complicate data analysis include many more. This entails as fundamental ones as counting. Some principles carry on. Jeroen’s description indicated how scripting data structures into computer code requires significant abstraction in the form of representing specific information with general indicators of how and where that information is stored. This kind of problem translates in other operations on data that allow turning it into formats that reveal patterns otherwise easily overlooked.

A widely applicable such instance can be seen in a presentation that considered minute geographical information. It is easy to imagine the context of neighborhoods and how they are equally relevant for a series of observations falling within them. Sociology thinks of them as “neighborhood effects” (Sampson 2011). The following summary of a data science talk illustrates such a problem in the context of taxicab transportation patterns in New York City, and the technical challenges coming with it.

This presentation demonstrates just in a side note how the command line allows passing the data from one tool to another. Here this technique helps to analyze large-scale data of which processing every single observation would take too long, and be unnecessary. What is so difficult here? It might seem easy to locate a point on a map and see which neighborhood it falls in. That is not so easy for millions of observations, of course, which New York Cities iconic yellow cabs produce. In the digital context, where Google's maps application is ubiquitous, we still have to imagine the complexity of encoding a simple map. Although digital maps often change the information they provide intuitively, zooming into a specific street corner, or looking at the cities as a whole, or the country, entails specific layers of information, data, on each level. At each layer we need the information of corners around countries, counties, cities, neighborhoods, and so on. Depending on the resolution, the number of their corners quickly grows. Processing all the transportation observations requires considering for each one to also generate the appropriate map and see which area it falls in.

Because they are customizable, command line tools are useful for conducting intermediary analyses that create a new dataset out of the original one. In one such application, for instance, this presentation proposed a series of "preprocessing" steps for a data set of around 100 gigabytes (i.e., although a lot, still much less than many datasets in industrial-scale production such as social media, sensor, or other log data). In his analysis of yellow cab and Uber rides,[48] the speaker used the command line to first aggregate trip destinations to the neighborhood-level. This strategy began with taking samples and then reduced the number of single observations, before identifying general patterns. The multi-dimensionality of geographic information just described complicates this intuitive solution such that it benefits from the command line utilities. As part of the presentation the speaker walks the audience through the details of attributing coordinates, stored in one file, with the circumference of neighborhoods, stored in another. Through a command line script, he passes the two pieces of information to a tool specifically designed for determining the larger geographical areas specific points fall into.

Here the Aaron Swartz strategy beats one that would fit Bill Gates's definition of work, or contact improvisation trumps bureaucratic specialization. To be sure, Joe's specialized tool, from the data structuring section above, could be designed to recognize geographic coordinates and match them to a

[48] *The New York Times* published the results of this project a few days afterwards (Schneider 2016).

respective layer on a map. While quotation marks in spreadsheet columns seems like a more prominent problem than geographic coordinates, they still make for a relatively general problem as well. Contrary, the following account offers a less common case and thereby demonstrates the utility of more basic tools once again.

This talk illustrated a series of data processing tools and strategies. The overall process, which involved steps that took several days of laptop processing, would not sustain production-level applications, for two reasons. Sampling, such as taking a subset of the millions of trips in this case, is useful for understanding patterns, but not for facilitating interaction with each individual user. Similarly, a historical dataset reveals underlying patterns, but undermines ongoing interaction with the result of data analysis. There are other strategies to address these problems.

The MapReduce modeling framework, which we have already considered above, facilitates large-scale data processing in different ways. The most basic problem it addresses is counting frequencies of observations, which makes for surprising session titles in advanced quantitative computation courses. It also accommodates more complex applications, as the following interaction between Sameena and her audience illustrates:

> … one of the tasks that makes such a goal [including natural language in financial performance analysis] hard is the fact that SEC filings are not small, they are about 3 terabytes in the compressed format and to create any quantitative models you have to have a significant amount of back history for the models to be created on. So this is certainly not, ahm, tiny data. And running this when we started working on it we actually realized that it would take us a few months for all the processing to end, so we needed to look at some other scaling up methods.
>
> So one thing I just wanted to say is, if you have any questions feel free to stop me in between, I'd rather have this be a discussion.

We once again see some discomfort in speaking about data size as Sameena avoids the non-technical and overused reference to "big" data in an equally non-technical but cleverer negation of "tiny" data. Instead of resorting to comparisons, here the speaker quantifies the amount of data directly. More relevant here, however, is the dialogue that unfolded immediately following the presenter's invitation to ask questions:

> Q. What were you processing it on?
> A. On a Hadoop cluster.
> Q. Yeah, and in saying that it would take three months?
> A. Oh, that was just on a simple machine.
> Q. Right, but what were you running?
> A. What we were running? The actual process?
> Q. I mean, ahm, if it's gonna take three months, are you running Python, running R, or are you running …
> A. Oh yeah, so we were running Python, with that estimate, that estimate was based on Python and some parsing tools that we have written, but we actually, ahm, I mean, I don't think if we had migrated it to some

> other language it would have sped it up much. So our thinking was let's put it on Hadoop and try it out that way.

Python, of course, offers no graphical interface to delete quotes and such. This was the problem Joe considered above. The interaction signals a certain comfort, of the technical community at least, with writing scripts instead of clicking boxes. While it was relatively easy to imagine the geo-coding implementation in Joe's Google-style data application, analyzing entire texts with a specialized focus in mind introduces much more idiosyncrasy. A recommendation algorithm, like Joe's, would have to anticipate all the steps Sameena describes above. At the same time, the Python programming language applies here as well. Whereas the previous presenter worked on the taxi trip project by himself, Sameena led a team effort, which implies significant informal activity as part of the implementation. Not unlike the previous JSON observation, around the common programming language we see different modes of organizing once again. There a bureaucratically specialized approach met an approach combining contact improvisation with possible inverted heterarchies. Here, a combination of bureaucratic specialization and informal collaboration overlap with contact improvisation.

In order to understand the organizational arrangement of this expertise, we also need to take into account the audience interest in the technical detail of data processing. The prolonged back and forth between the presenter and the audience indicates the lack of a commonly shared understanding to articulate this interest. This lack of technical language makes the invention of narratives that use commonly understood knowledge plausible. Once they figured this out, R, Python, and Hadoop are all common tools.

Let us also return to the challenges of textual data compared to geographic taxi data in order to better see the scope of this expertise. The previous instance of the taxi trip analysis drew samples to solve a similar problem. But samples are less useful if even a few observations require to be extracted from large amounts of textual data, hence the speaker passed the data on into an additional set of tools:

> So here the actual problem was not so hard for it to translate to a MapReduce job, and hence my point on that many jobs can be easily migrated to a MapReduce and the challenge for massive data sets is not in migrating it to a MapReduce job, the challenge is just how you think the challenge may lie, and where you start looking for value. So we wrote a simple MapReduce job and it processed it quite fast. All eight years of filing got processed and we could create models in under 30 minutes, which was good enough for us.

The ease of migrating tasks to MapReduce is disputed, as we will see below in more detail. Yet this speaker considers the non-technical problem of defining a question as the major challenge. It provides a more arcane solution to the taxi problem. At the same time, their respective questions and resources

differed as well. In short, one basis of the uncertainty data nerds navigate stems from the problem that multiple solutions are possible and settings differ too much as to evaluate optimal ones, thereby inhibiting specialization.

The accounts describing the technical implementation of counting problems have not resorted to articulating them in some unrelated but commonly known references. Time, another non-technical aspect, has been salient here as well. Time is a common reference, similar to the organizations from the discussions of the stack or the vivid ideas associated with big data, if less evocative compared to a pulse or Obama. Even in these much more subtle explanations, we therefore see the emergence of abstract relationships between technical capabilities through common references.

These accounts have only brought us through the data preparation phase. Often these processing steps indeed aim to facilitate conventional analytical strategies. But not always, as we see next.

4.3.3 Estimating

The MapReduce framework distributes tasks across multiple machines in order to facilitate counting problems, such as of words and phrases in financial statements. Statistical estimation, on the contrary, is often conceived of in matrix format. For this information must be combined, not distributed. In applications that require timely processing of large-scale data, these problems hence require special frameworks. Moreover, it is not always just about scale, but in specific problems also timeliness. Instances in which users expect results immediately complicate the problem, especially if the underlying data continuously records behavior or other information. Many online services, for example, are built on the promise of instant recommendations in combination with the utility of statistical estimation.

Because the underlying mathematical methods stem from a time when continuous estimation was not yet a problem, or opportunity, they require significant adjustment.

> And so the claim I am going to have is that most people now accept that the state of the art is now called stochastic gradient descent, or SGD, and that this is for instance what Vowpal Wabbit does, if you ever use Vowpal Wabbit, but many other tools are using this. And it's unbelievably simple intellectually. And this is what is awesome about it, this is a tool that will just go forever on. You leave today, and you have complicated models that you want to fit at huge web scale, you can just use SGD and you are sort of fine.

Here we learn several things. There is the impression on consensus over stochastic gradient descent, or SGD, that this can be seen because it is part of a tool with the name of "Vowpal Wabbit," and that this is useful for modeling large-scale data. An explanation of this tool as the result of a project at

Yahoo! Research seems unnecessary in this audience. And nobody speaks up (while on another point someone did). John, the presenter assumes an audience that is comfortable with the terminology he offers, and interested in understanding the underlying process.

This extends our understanding of the community so far. John offers no common references. Unlike the problem of processing textual SEC filings, this context of "huge web-scale" itself is commonly known, if not from the technical angle John takes. Moreover, the idea of statistical estimation is among these groups much more familiar than modern MapReduce techniques. This makes it less surprising that although the speaker cites a specialized tool, he focuses mostly on the technicalities of a formal method without broader introduction. In other words, while the gap speakers tried to bridge with respect to big data technology seemed ambitious, in this analytical context it seems relatively small. This consensus and the distribution both characterize the data nerd community.

This can be seen as despite this assumed consensus here, uncertainties remain even from the speaker's perspective, as we see in this comment:

> Let's talk a little about what goes wrong. So this already came up, so the fact that you have to choose this constant set size eta, that's bad, the fact that you have to tune this, I mean anything that's tunable is always just bad. Things that are tunable are likely to go wrong. Good batch algorithms, as I said, use this line search algorithms, which will set eta for you. We don't really have a great theory of how to do that when you are in the online setting. We are starting to have it, and Brad Allesandro, when he heard that I was doing this talk, wrote to me and said 'we've been using recently one of the things John Langford has published a paper on, that works very well.' So people are starting to come up with algorithms for setting eta, but it's something that is definitely an active topic of research.

From Vowpal Wabbit we have moved quickly into mathematical notations and interrelated algorithms. Solving the remaining problems and relationships between them constitutes community effort, as the email exchange indicates. Community, of course, is a non-technical idea, just like organizations and time. Thus, even the most technical accounts invoke common references in order to articulate the details of implementing arcane data expertise. Thus considering the tools used for data analysis has revealed to us the anonymous opposition between utilizing graphical user interfaces and the command line for data processing, or sampling and MapReduce for data analysis, and finally the collaborative work involved in translating traditional techniques into modern contexts. The consistent strategy of making common references connects the variables substance underlying those reference.

Tools synthesis

As part of the analytical repertoire we have just briefly considered three "tools": cleaning and structuring, preprocessing and estimating. Aside from providing some technical context of data science, we have also discovered a distinct cultural signature that draws heavily on existing tools and frameworks, while preserving sufficient technical purity in order to combine them such that they accommodate specific problems. We could see this strategy contrasted to ways interacting with data that resemble lay uses of software services. Because some of the problems are so obvious and the solutions pertain to lay users directly, it is easy to overlook the expertise that is required for making arcane technical tools directly useful. We see this next.

4.4 Magic

This chapter has so far considered modern data tasks as a jurisdiction of problems beginning with the activities generating data, technological infrastructure recording and storing it, and programming tools to analyze it. This has revealed strategies of speakers that connect arcane technical problems to common ideas and references across different tasks of data analysis. The kinds of references vary with respect to the knowledge speakers and audiences share. The data jurisdiction involves a series of expert tasks in which data nerds often play just an implicit role. Those solving data problems sometimes justify their efforts with simplifying the work of a data scientist.

When explicitly mentioned, the data scientist is described in interesting ways, as we see here:

> Every time I read one of these articles [on the slide, e.g., Big Data: A revolution that will transform how we live, work, and think], I felt a little bit like … this is where the magic happens.

In this context, we once again move between specific and evocative language, such as log data infrastructure and sexy applications, and stochastic gradient descend and now miracles and magic. Resorting to language of miracles and even a novel term like data science itself therefore signals not just a lack of understanding of the technicalities of the underlying expertise, as also the reasoning that follows the previous comment suggests:

> But again as a statistician, right, as coming out of academia, somebody who I thought I had the tools to do magic, there wasn't really anything that I saw in the stack that allowed me to do magic. And so what I ended up believing was that the miracle occurs here, right, the miracle occurs at the data science layer, but it is still not very clear how to do data science in a modern way, right.

Indeed, the type of practice and work of different roles in the shared context of the data jurisdiction remains deeply ambiguous, even within the academic community that has sophisticated technical

understanding. On the other hand, just the few accounts in this chapter have covered some aspects of modern data science. They are observations from public events, and not formalized in textbooks, at least in this combination. From a formal perspective it is not clear what modern data science is. Much knowledge is available, however, informally.

Guidance from one arcane context toward another comes from common references, as we see here as well:

> So the way I think miracles occur is basically the picture here [picture of actors Jonah Hill playing a baseball statistics analyst and Brad Pitt a team manager in the movie 'Moneyball']. You need both, and I actually love this picture because I think it does capture data science perfectly, so you need both a data scientist who is kind of the nerd, the quant, the guy who understands the statistics, the guy or the girl who understands the statistics, and you also need somebody who can do, who can actually act and make decisions, and change a company based on those decisions. And it is the partnership between those two that can actually make magic happen.

The rhetoric around miracles and magic also makes for potentially effective self-marketing. These accounts reveal more than this rhetoric, however. What we have seen expressed before in wild comparisons of analogies between substantive and technical problems, here resonates with the combination of managerial and technical roles. And just in case Obama and oceanographers were not prominent enough, Hollywood helps us understand all this. It provides none of the details of SGD, to be sure. Nonetheless, data nerds appear everywhere and their competencies are commonly applicable.

Just by itself, a perfectly functioning stack will not solve custom data analysis problems, and predictive queries will not integrate novel data streams. These challenges nevertheless seem to operate under a different mode of practice than the purely technical difficulties related to the infrastructure, a mode that sustains the consistent label of data science. The reference to magic here indicates that this practice is not well defined. We have seen in so many instances a highly technical and sophisticated basis. Its application seems magical, because there is no technical language for the combinations of those technical components in applications. The nerds instead resort to common knowledge as references, and to images of popular press and cinema for the role itself. These practices bear sociological relevance because they help us understand community identification in practical instead of nominal terms. Nerds find sufficient technical tools, such as the command line, but also GUIs, SGD and others absent technical language to articulate their relationships to one another.

Chapter overview

How do data nerds organize their expertise? Whatever the technical components of which the stack is put together, a pattern emerges with respect to the relevance and underlying principles of how the details relate to one another. Held together through common references, they apply across a number of contexts, including, but by far not limited to, politics, on the federal and local level, social media and engineering. Recalling this diversity is important because it points to a critical direction of professional work. Consistent with many well-known professions, data nerds face the challenge of finding little guidance in substantive or technological constraints, let alone otherwise ubiquitous organizational planning. Yet, established professions still show evidence of specializations. Although law appears as a coherent and integrated profession, research is well aware of significant specialization in the legal field (Heinz and Laumann 1982, Espeland and Sauder 2007, Phillips, Turco, and Zuckerman 2013). Specialization is even more obvious for medical doctors (Menchik 2014, Freidson 1988). While they all heal our bodies, they focus on different parts of it. The specialization clearly seems relevant, although it is no meaningful indicator of professional promise by itself. Engineers are well known to specialize, and also for lacking professional autonomy. In the place where they rely on abstract stocks of knowledge, we have just found references to common images and ideas in the data science accounts. Meanwhile, a common core began to emerge around analytical tasks. A number of questions therefore arise with respect to how data nerds apply this broad technical basis to specific problems, and how that process contributes to its salience.

This way of organizing expertise has consequences. The diversity suggests that data technologies are not harmful in the way many industrial technologies are, even if the rhetoric of coal mining draws those comparisons with respect to the scale at which data has become available today. The critical difference lies in the relatively cheap processing and manipulation of the data that result from modern capabilities of sharing them in an anonymous fashion.[49] Appropriate expertise of utilizing these capabilities constitutes the more pertinent problem. Importantly, similar kinds of problems appear across substantive areas, dispersing their overall basis for concern. To be sure, we saw some systemic features as well, such as those related to widely shared data sources. Overall, however, these accounts seem to

[49] Cloud computing has allowed small actors to draw on powerful computational resources before making long-term financial commitments.

suggest that the sources for public concern are widely distributed, and not directly associated with familiar organizational forms.

For the individual opportunities, these results indicate that relevant expertise is not exclusively tied to modern technologies. Here, data science seems to differ from biotechnology, another recent case in which new technologies have brought with them individual opportunities. In those instances, careers were tightly linked to that new sector and the universities where the underlying technologies and ideas originated in (Owen-Smith and Powell 2004). We have seen few such references in these accounts, and none that seemed to have shaped the work beyond constituting aspects of it.

Before focusing on these abstract community characteristics prematurely, we can think of these organizational arrangements through the cast of characters from the introduction. In those terms, we have seen here much of the bureaucratically defined approaches, which Bill Gates promoted in his software development projects. We could find them in several corporate efforts, aiming to facilitate big data operations and the work of data nerds. While data matters here, its speakers have articulated its application with respect to the specialized purposes large corporations and smaller startups offer.
We have also seen evidence of different ways of specifying similar or at least closely related tasks. Contrary to the bureaucratic definition and anticipation of work, the availability of data has disturbed the orderly life at technology companies where the sheer amount of data has created uncertain challenges. Similarly, existing proprietary software is considered ill-suited to leverage the data in the first place. All these observations indicate moments of unexpected challenges and new conditions of the kind Aaron Swartz's contact improvisation induced as well, although these projects here remain politically far removed from his activism.

This discrepancy of bureaucratic and insurgent definitions of tasks in the same context becomes even clearer in other moments. We saw, for instance, references to open yet direct digital ties to specific actors and their data streams. These actors can be seen to gain central positions in the data context as their data is relevant for groups so general as to be thought of as oceanographers. Likewise, we have encountered the communities of Linux developers with their specialized tools and insider jokes, which directly invokes Linus Torvalds's definition of tasks. These kinds of projects therefore suggest a division

of labor that is specialized around leveraging these data sources, although the respective expertise is not necessarily tied to the companies themselves.

Finally, some accounts also begin to reveal direct attempts to define distinctiveness and broad visibility. We could see evidence of this through those speakers who emphasize the reversal of the private and government applications of significant technological innovations, and those who speak of magic, which is of course technically vague but commonly understood. In other words, we have seen integration of technology, the economy, and various substantive applications. It is not clear how they overlap, and hence on what basis the different specializations relate to one another. Contrary to the other nerd perspectives, here we begin to see ways of articulating the broader significance of modern data independent of specific actors associated with it. This reminds of Mills's emphasis on the simultaneously visible and anonymous role in the organization of work.

The main question is then whether data science has ways of integrating these diverse areas with more mundane practices which "magic" suggests, or if data nerds resort to rely on guidance from older organizational foundations with more institutionalized knowledge. If data science is able to define a body of arcane knowledge that robustly applies across these contexts, data scientists gain flexibility and hence autonomy. Since all are instances of data, the answer could seem obvious. After all, statistics has developed tools for data analysis for over a century. But the literature on this development also finds that data analysis has remained constrained to institutional contexts in the past, such as the census, the insurance industry and risk modeling (Porter 1986, 1995, Desrosiáeres 1998). The eclectic references that speakers have made here suggest that today's technology is better able to make data accessible beyond such boundaries. We have seen ways here in which data nerds relate these obscure technologies to the social stock of knowledge. We cannot tell yet, however, how data nerds integrate these novel capabilities with respect to one another. This question defines the task for the following chapters.

Among the disarray of stack layers and technical tools for structuring, processing and estimating, we were able to discover a subtle yet strikingly consistent use of some very substantively distant analogies. This could be dismissed as irrelevant and interpreted as signaling technical incompetence of speakers and audiences. Doing so would ignore an interesting curiosity, however, and leave much

analytical leverage underutilized. We have seen references to common ideas and knowledge throughout the technically arcane accounts. They have implicitly suggested comparable characteristics of the common ideas and arcane ideas, such that the former help us understand the latter. In other words, we can see them as applications of analogical reasoning.

Analogies offer specific instances of comparisons and for making claims. It reminds of findings of Carruthers and Espeland (1991) that show the significance of rhetorical effects over rational effects of double entry bookkeeping with respect to facilitating commercial interactions. Their historical study reveals the continuity of the narratives business accounts started with all the way to modern practices in purely numerical rows and columns. Data science's emergent status offers too little basis for testing the relevance of Carruthers and Espeland's finding in detail. They nevertheless support the recognition of rhetorical effects that come with rational methods, and thus encourage to investigate how such a process unfolds in data science specifically. This context enhances variation because underlying problems spread out along the stack and across substantive issues that arguably require a broader and more complex set of decisions than those associated with quantifying economic units, as complicated as that may be. In other words, we can acknowledge that the nervous system and Ferraris make for different narratives than the fellow merchants that appear in early narratives of how much they owe their business partner, and by when. Before concluding the former to be too far-fetched, we need to consider its role in the overall set of relationships that define the contours of a thought community.

Contours: Illustration

In order for problems, be it obscure diseases, convoluted laws or complex technology, to become visible, they need to be integrated into the social stock of knowledge (Berger and Luckmann 1966). Here we have seen data nerds and technology experts struggle with such integration as they were speaking at public events. The struggle is evident when we consider the engineers Bobby mentioned early on in this chapter, who could not estimate how much data the processes they oversee generate. How can the audiences Bobby and the others address know such things, or why they are important? Rather than focusing on the technological details, we learn about the Obama campaign, oceanographers, coal mining, Frankenstein, Ferraris, and the global nervous system. All of the references carry some public meaning,

while none reminds of the modern data industry. Data nerds connect them for us.[50] This effect indicates autonomy from the technical infrastructure. Nevertheless, a question remains of whether nerds, who address problems with so little public relevance as to lead them to construct such narratives, can still preserve their expert status. In other words, can we specify the fine line between illustrating common relevance and remaining consistent with arcane knowledge?

Aside from providing vivid images, the analogies these illustrations imply indicate a technical device for making unfamiliar problems salient. Marketing research investigates such strategies systematically and finds that people tend to construct analogies in order to understand novel products (Hoeffler 2003, Gregan-Paxton and John 1997). Although marketing is not central here, and discouraged, the content of these presentations often resembles novel products at least in rough terms and of an analytic kind. It is therefore possible that some of the analogies we have seen here directly benefitted from such research as speakers may have prepared their talks with marketing advice. Either way, just recognizing the possible role of marketing here suggests considering the data jurisdiction in broader terms than purely reflecting the underlying technology and substantive problems. On a more technical level, the different illustrations ranging from oceanographers to the global pulse and nervous system have no substantive relationship initially. In these presentations, many functioned as analogies to obscure technological problems, leading to indirect connections between them through their shared association with technology. Marketing could offer one explanation for the usage of analogies, although a similar result could follow from the common strategy of people marketing research draws conclusions from. While the following considerations suggest that it is more of the latter, the subsequent analysis shows that this distinction is less significant for the question of discerning contours of the data science community than their common distinction to more technical strategies for making analogies.

Not all analogies and illustrations work well. As I mentioned before, there were few clear promotion instances. Sometimes technical and non-technical colleagues presented together. This instance without a technological person part of the presentation signals the importance of technical competency. In one instance, a speaker tried to illustrate the benefits of data accessible through his project with the idea of "unleashing" the data. Together with introducing the term he quoted one of his engineering colleagues

[50] Although law is often thought of in terms of its highly technical and abstract knowledge, Berger and Luckman (1966, 77) remind us of how judges have to bear in mind the concrete word of the defendants facing them as well.

whom he had consulted with describing this as a marketing idea, not an engineering term, and recommending not to use it. The speaker nevertheless presented it to the audience just to meet much skepticism, as could be seen by the subsequent comments. In short, although to an observer many illustrations look equally unlikely or technically unsubstantiated, there seem to be important nuances in which they are acceptable.

Aside from indicating marketing's coordinating effect, the findings of its research may explain our observations as well. Here we expect personal experiences in association with technical problems. Some analogies were outright unrealistic or erroneous from a technical perspective. Here we can recall how we learned about sweat, blood and tears, jigabytes (which don't exist), and woodwork at technology companies. Their purpose is most surprising in a technology context that is ripe for formalization. At the same time, it seems intuitively clear that this figurative language aims to illustrate the significance of certain moments in the process of doing data science work. The references connect very common experiences with technical contexts. They effectively avoid potentially lengthy descriptions of the underlying processes, but also accept significant risk for their own status. Even as technically literate presenters articulate arcane problems through illustrative analogies, in addition to addressing each other as well as a broader audience, they also provide templates for others to go on and talk about these problems without the same reference in mind. This can easily result in instances in which their peers resort to analogies that may indeed reflect their incompetence with respect to the technical details and raise skepticism over problems defined in this area more generally. Applied in such ways, analogies could give rise to friction between groups within the technical community.

On the other hand, we should not dismiss analogies as idiosyncratic narratives. Unlike marketing's primarily substantive concern, sociological research design leverages analogies for their analytical utility. Considering its interpretation would suggest slightly different implications for data science. Although as a scientific discipline much sociological research follows theoretical guidance to discover relevant problems, some methodological guidance relies on analogies. This begins on the basic level of defining "units of analysis," such as in the case of assessing effects of class positions where the assumption is that all individuals as part of a class are analogous on the basis of their relation to the means of production

(Stinchcombe 2005, 152-7). Considering analogies in this context reveals their role as a shared concern on a more general basis. Here they thus also contribute to defining a thought community.

Returning to the accounts of this chapter, we find less idiosyncratic references than we have considered so far as well. The idea of “the stack” can be seen as an analogy, as well as the idea of “data pipelines,” which is sometimes used to describe a sequence of steps of accessing, manipulating and analyzing data. Both appear systematically. We have also seen evidence of generally relevant meanings in very specific analogies. We learned, for instance, about grep for JSON, that work of three months can be done in thirty minutes with a different processing setup, and that these thirty minutes were “good enough” in one case, whereas another expressed content with three days processing time. Unlike previous analogies, these comparisons are qualitatively relatively similar, each comparing one technical aspect to another. This illustrates a way of discussing problems where the relationships between them remain arcane otherwise. As a result, analogies define boundaries of a set of problems. These kinds of analogical illustrations of the data processes, which resembles their sociological use independent of specific applications, indicate some reflexivity of the community in as far as they articulate the systematic patterns underlying their work.

With this in mind, we can reconsider the problematic reference to “unleash” data as a way of considering the different implications of the respective strategies for illustrating problems analogically. Like many other illustrations we have considered here, “unleash” is commonly known. It has no direct relationship to data, but that was unproblematic for other illustrations. Considering the technical details of appropriate analogies, we note that unleashing a dynamic with its own momentum, like a dog can run on its own once let off the leash. None of the others has this connotation. It has no intimate connection to the problems this thought community claims responsibility for.

Some evidence neither focuses on the marketing interpretation that emphasizes familiarity nor on the sociological interpretation that emphasizes technical similarity. The ambiguity that comes with this strategy of defining problems can be seen most clearly on the level of the data science label itself. Chapter eight discusses the origin of the term and its conjectures in more detail. Meanwhile, just recalling the composition of data and science from the introduction reminds of room for confusion. Sciences rely on data, without making such references in their label. From that perspective the label has no utility.

Moreover, with statistics the label directly ignores a discipline dedicated to study quantitative data. At the same time, the idea of science provides the most complicated but also clarifying illustration of data science. Like in academic science, the idea of data "science" invokes an image of a bounded group. Although less so than Ferrari chassis and non-Ferrari engines have absolutely nothing to do with the data stack and visualization, data science and academic sciences do have different roots as well. On an abstract level they still resemble each other on the basis of describing a community of people with overlapping arcane knowledge, a relationship that I revisit throughout the following chapters, and particularly in chapter eight.

This discussion of the mechanisms of the rhetorical strategies has revealed some distinct contours of the data nerd thought community. Most consistently, all these efforts indicate that data nerds do not let technology speak for itself. The second finding is inconclusive and mostly points out directions for further inquiry. The different kinds of analogies and illustrations, ranging from misleading to technically useful but also signaling some ignorance, show a heterogeneity that, while not surprising in the context of an emerging group, undermines explanations that would account for this emergence on the basis of a cohesive group. They operate to some degree independent of the experiences shared in those accounts with respect to the technical considerations underlying modern data processing, which the main part of this chapter has discussed. Two questions follow. More immediately we need to ask how data nerds persuade others of their utility as far as the illustrations have limited utility as well. Second, and more systematically, we need to ask what integrates this community if neither technologies nor cohesive groups do so.

5 Organizations

With the data jurisdiction sketched out in basic technical terms and rhetorical strategies that began to reveal initial contours of a distinct thought community, the question emerges how nerds apply the technology, which we have seen being described so vividly, to practical problems. Here I examine how data scientists position themselves relative to organizations and clients whose problems they address, and other expert groups who they interact with as part of this.

This question pertains to public concerns with data science's impact on everyday life as well as to individual concerns with the work opportunities it defines. With respect to the former, this section considers the role of organizations in defining data science tasks, and by extension their consequences. While we have already found ground to question the significance of bureaucratic specifications in data science, we have not yet systematically considered alternative processes of how nerds interact with organizations as part of implementing data applications. This perspective leads to a more comprehensive account as it begins to untangle the range of ways in which data science's consequences unfold by considering applications independent of the wider attention they have attracted, including those that may not pertain to the public directly. Although the main puzzle results from data science's broad salience, these more subtle effects could shape the expertise definitions it finds recognition for.

We can implicitly see here also how data science's role in organizations is relevant for individual concerns with work opportunities. A new role may operate in similar ways as many familiar ones, tightly embedded in the organizational structure. This seems once again unlikely for data nerds as the activities data science has gained recognition for were seen to some degree independent of the formal organizations around them. If that is not applicable here, we need to understand on which grounds data science differs, and ask how it articulates those differences as to justify that status in a context of those other functions not sharing it. We can consider different the organizational arrangements we found of the four technology nerds in, and on which we have focused as models before.

The previous chapter has begun to outline in basic terms how the technology nerd characters map onto today's data problems, or, in more arcane terms, the organizational principles data scientists coordinate their expertise with. We have seen a number of hierarchically induced as well as heterarchically integrated specializations, and some improvised and broadly relevant initiatives. We have

also discovered grounds for subsequent questions. It has remained unclear how these partly contradictory forms of organizing work come together in specific moments where nerds apply their data expertise. In other words, we were not yet able to specify how practices as different as those seen in Bill Gates's views and Aaron Swartz's projects complement each other in the data setting. Here I begin to address these questions as we consider how the technological aspects of modern data problems unfold in their organizational contexts.

While we know that bureaucracies organize data technology, it is less clear how the specific data nerd fits into their formalized tasks. The organizational context leads to some straightforward expectations as to which nerd roles should be prominent here. The organization of technology in Gates's view offers a clear and vastly familiar role of a specialized software engineer. How could data nerds fit such specification, or how would a firm script the magic, which data scientists are expected to produce, in a task description?[51] On the other hand, while we could expect in the previous technology setting erratic challenges to the corporate hegemony as responses to the concerns often associated with technological progress, it is not so clear how they unfold inside formal organizations. Finally, roles of the kind we find in the integrated nerd community Linus Torvalds represents, or the anonymous role model Mills mourned, are relatively unlikely in a context that is already institutionally defined and viewed. Early computer programmers, for instance, lost responsibility to the compartmentalization of their work through managers (Kraft 1977, Ensmenger 2010). Lawyers, on the other hand, were able to benefit, at least to some extent, from growing bureaucracies (Abbott 1988). We have also already considered how informally coordinated expertise undermined the established medical community on important questions (Epstein 1996). Thus, while it is clear that we will find bureaucratic definitions of work when we consider contemporary organizations, other forms of coordinating may prevail nonetheless.

The focus on the organizational setting also reveals another contour of the data nerd community. We have seen so far that experts of data problems utilize comparisons, metaphors and analogies in order to articulate the significance of the problem they solve with data. They assume no general understanding of whose problems data can address, nor how. The technical setups are thus unable to solve custom

[51] As fantastic as this may sound, organizations have scripted miracles in the past. In its response to rising Protestantism in Europe the Catholic Church rationalized its classification process of miracles. Candidate activities still occurred outside of its bureaucratic structures, however (Parigi 2012).

problems directly. In this chapter data scientists enter the scene with their “miracles in the middle.” How do they position themselves in the non-technical, organizational context relative to the solutions in place already?

I address this question across accounts on four kinds of interactions of data nerds with their organizational environment. After considering routinized interactions as the central feature of organizational life first, I move on three kinds of moments that are sporadic and thus challenge data nerds to articulate their relevance. Here I focus on hiring, consulting and moments of friction.

5.1 Management

The most visible data problems of an organization are by definition the ones users and consumers, the public, interfaces with, such as search or purchasing recommendations. They offer an intuitive context to consider data science in. I get to them in a later section.

Data science expertise, however, applies much more broadly, and no less directly or routinely, in the operation of organizations. The following description of an internal application by John, a data scientist at a public-facing organization, illustrates one such instance:

> [As the company was growing quickly,] it became very difficult to schedule [the phone customer support operators], right. So, we had two people on the support team, in excel, trying to schedule everyone … they had to slide people around and slide their lunch break around to meet demand, and it was just awful. But I can actually sit down and write an optimization model that defines the entire decision space as a polytope, and then we define an objective and we actually go and search that space, and find ‘okay this is the right corner we can actually solve this problem in’ and provide a schedule back. We actually wrote up in a language [showing a screen full with computer code], this is .lp format, yeah stands for linear program, this is horrifying, this is how the model looks like before you run it, and out pops the schedule, shove back to excel.

Although an initial approach could rely on conventional spreadsheet software and manual arrangements, John provides a custom solution. He shares with the audience not only the specific problem his organization experienced at the time, but also illustrates his description of the solution with evidence in the form of pictures and screenshots of the tools he used and built to solve it. John thus shows us pieces of his process of understanding an organization and translating its problems into a formal and quantitative framework. The radical shift from excel to .lp indexes the lack of definition for how to solve such a mundane problem as scheduling support staff, and is at odds with a bureaucratic definition of tasks. Instead of following formal rules, John began to define them himself with consequences for others on the basis of his knowledge of polytopes. He encoded external ideas into the internal processes.

John goes on, embedding this technical solution in a more general understanding of his role in the organization:

> So now I can provide this point, just means 'you are on point,' meaning you are on chat, actually picks lunch breaks for people, picks when you should do email as opposed to chat, and so the cool thing here was I was able to do what I know how to do, which is data science, and the support folks actually got to go back into support and do chat, which is what they know how to do. You don't want me talking to customers, and you don't want them figuring out the schedule. So this is just a better use of everyone's resources. This was an internal engagement, just do this as a one off, and, do the schedule for folks.

A clear division of labor for data problems and others follows, after all, and thus acknowledgement of bureaucratic principles. Previously managerially defined tasks, such as specific support channels, here result from an algorithm. This implementation therefore reveals one of data nerds' more subtle consequences, the implications of which I consider below. Meanwhile, despite imposing new rules, John also describes how he initially worked on fitting into his organization, dominated by designers. By drawing a line between their expertise and his own, John argues that the positioning should not be understood as the implicit hierarchy it entails practically, with the data scientists through their algorithms controlling the schedules of colleagues in other functions. Instead John leads us to think of it as a clear positioning of data science relative to other roles and responsibilities. Although nerds often opt out of organizational rules, trying to fit into an organization is not surprising either. As Mills lamented, their ubiquitous presence undermines all sense for alternatives. With the specificity John provides, we see the technical correlate of such collegial politeness, or ideological discipline.

Others take a different perspective, as we learn from Riley, who is leading a data science team at another organization:

> So one of the things I have been really working on … is trying to create a culture of, you know, curation of high quality data, you know, outside of our team. It is important to get logging right, and it is important for people outside our team to feel responsible for that, knowing that the better data they create, the better insights they are gonna get and the more quickly they are gonna have them.

As John before, Riley also describes the benefit of others in the organization as a central concern. Instead of defining a specialized problem area to be addressed by data science, as we know from bureaucratic tasks, it fuses with the organization more generally. Others have inspired this view:

> Ahm, you know, there's teams that really have that nailed, and they are typically teams that, you know, are companies that are completely fueled by data. I think square [a digital payment company] is a great example of that.

We can find two important points in this reference to square. First, it signals features of what DiMaggio and Powell (1983) refer to as institutional isomorphism, that is, increasing similarity of organizations within the same field. John's application of a modeling framework that was developed for

purposes other than scheduling support staff has a similar effect, but operates on a technical and more abstract level than this direct reference to another organization signals. The isomorphism is thus not necessarily one that re-institutes bureaucratic rules from a previous era.

A technical underpinning emerges here as well. Upon audience request, Riley explains further that the data science team trained other roles in the SQL query language so that they have an easier time working with the data themselves. In other words, Riley's informal reference to a peer organization leaves formal traces and invokes and indexes more widely shared expertise. Internally, this processes raised status questions more directly, as "the designers hate it, but, you know, we teach them, and ultimately they like it, because it levels up their game a little bit." Contrary to some propositions that made coding unnecessary, which we saw in the previous section, here the coding skills enter other tasks in the organization. Riley's description of a cultural shift therefore slightly differs from John's tool that presented a table to "shove back to Excel." In this interpretation, cultural scripts (Swidler 1986) map directly onto computer scripts. Algorithms are still part of the learning processes, however, as Riley's team utilizes a tool that proposes queries based on what data colleagues have pulled.[52] We once again see the limited relevance of bureaucratic specifications, although the legitimacy of the overall organization remains obvious in these accounts. Contrary we see two more ways of community identification. One process operates through shared technology and knowledge, such as "linear program," and the other through peer role models.

The shift in the organization is not limited to the obvious cultural differences between design and data nerds. Riley also describes technological consequences of his internal project:

> … we totally changed the way we log data, and we made data scientists responsible for instrumenting everything on the site, which is pretty cool, because then they are the experts on the taxonomy of the systems and how things are implemented and all that, where it is stored. And then we also made data scientists responsible for the warehouse, the way it is designed.

Whereas Riley interprets the SQL data query language as a communication device initially, here he emphasizes that data science also defined the infrastructure, within which this communication takes place. And once again, the enhanced role for data science is for the benefit of the organizations:

> We are just this week finishing a monstrous overhaul of our whole data repository, and we have done a lot that just simplifies everything and makes it more intuitive, which will not only help data scientists, but it also

[52] This tool works to a similar effect as the open office layout Turco (2016) describes as a component of modern organizational learning in a physical setting, where employees pass by different desks in order to learn from their peers. In this case people see their peers' ideas through the data query software.

> democratizes data throughout the organization. … otherwise data scientists are like the gatekeepers of information, which is incredibly inefficient and, you know, they just get buried in ad hoc requests for, you know, stupid little things that people should be able to answer on their own.

The motivation is to undermine the bureaucratic position data science could have claimed, because data science sees itself responsible for different tasks. In their overall message these two accounts of data science in its organizational context differ significantly. John explicitly tries to adjust to the prevailing culture, shaped by a design focus. Riley aims at imposing a data culture even beyond the data science team all the way across organizational functions to designers. They also joined at different points relative to when their respective organization formed. Riley explicitly reflects on his surprise with the fact that the emerging organizational culture maps onto the relationships that were there early on. This might explain part of the variable interactions with their organizations.

Aside from this attitudinal difference, their observations also reveal many similarities. In technical terms, the two positions both invoke significant improvisation through contact with others in the organization and beyond. In addition, Riley describes a continuous coordination system, based on shared technology, that runs parallel to formal reporting structures. Finally, their accounts also reflect the process of articulating and defining these roles within the organizations. The public nature of these events where I have observed them turns the experiences they share into possible scripts for others, not unlike those that drove Riley's narrative as well.

Both are instances from relatively young companies. And Riley invoked another technology company as a role model. It could seem much easier for data science to undermine bureaucratic processes where they exist just in rudimentary terms. I therefore consider experiences from a traditional corporate setting as well. Here it turns out that similar struggles unfold as data scientists penetrate formal bureaucratic arrangements with established functional roles. They challenge data scientists more to articulate the specific contribution they make compared to existing solutions, as Rachel describes here:

> Another thing to keep in mind is that most our, you know, most systems create, produce logs data [showing slide with a log text file]. And so, ahm, just one of the first things we do when we talk to anyone in the company, and we talk to the people in the company who are in charge of the business systems about what data that they have and what data we can have access to, they often speak in terms of cubes and enterprise data warehouses, and so we wanna get to the raw logs, and often get initial pushback because they want to know 'why do you want to deal with the messy logs if we have these nice, clean data warehouses.'

And indeed, whereas Riley took over warehouse design, Rachel has to negotiate access to begin with. To be sure, they don't make these requests for the sake of displacing existing competencies:

> And so part of this is that traditional databases are just snap shots in time, and a lot of the interesting patterns in the data are getting lost when you roll it up into these traditional reporting systems, or reporting cubes.

In other words, Rachel emphasizes quality and utility of the data, not the status of her group in the firm. And it is not just the technicality of some information that was not considered in the past, the whole approach is different now:

> And so when we're encountering these internal warehousing teams, we're striving to, it is subtly different for some people but it is substantially different for us, that when we're encountering these traditional data warehousing teams, the traditional data warehouse project involves sort of lots of work to enforce order on a rather chaotic system, and it is just a never ending task to create this final data cube. And instead our approach has been more tactical where we are more comfortable in dealing in the chaos and in dealing in the raw logs at the log level, so we process what we need in order to process, to solve the problem we need to solve in the moment for either for a specific business question or to build a specific data pipeline. And so we find this both pragmatic and it also avoids the problem that you are not prepared to go back to the original source or you potentially lose key signals if you don't have the original source around. And this, I think, is sort of a philosophical difference, or shift, in some ways.

In my observations I have encountered much fewer traditional corporations than startups, though this was by far not the only one. Nonetheless, we see that data nerds can be relevant for them as well. They nonetheless give data scientists a harder time, as we expected from the situation of earlier generations of computer nerds. The hesitance stems not only from establishing a new function. Nerds bring in a different style of work, one that is more "tactical," or less bureaucratically planned. In the next chapter I focus on the question what this might mean.[53]

Thanks to Rachel's perspective we can also understand data science's own coordination. The relationship between data nerds and designers in John's case is ill-defined and therefore unsurprisingly requires more interaction between the two sides than the one we see here between two technical groups that address related problems. Rachel nevertheless describes differences quite clearly, in that one side addresses continuous tasks, whereas her group pursues more "tactical" goals, which resonate with John's "one-off" solution for scheduling. This similarity suggests a consistent pattern of work across very different organizations with respect to age and structure. We also learn that this has to do with the technical details of recording and storing data. In other words, technical setups and styles of organizing tasks are directly related. Rachel's argument also has more general implications as continuous tasks would lend themselves much more to bureaucratic specialization than tactical tasks that change regularly. In spite of all these differences, the account we see here indicates no ambition to place one over the

[53] Later on we also learn that some of Rachel's initiatives in this role reflect her experience from working at Google before. While she does not make that connection here, it seems likely to play a role as well, and supports the isomorphism argument from above by pointing out the movement of nerds between organizations as an additional channel to the imitation Riley described earlier.

other with respect to rigor or utility. Rachel's argument, complementing those of John and Riley, therefore primarily situates data science within a traditional bureaucratic organization, defining their tasks anew.

Although differences in these perspectives on how data science fits into organizations remain, they begin to address the ambiguity left before in accounts of the technological background that facilitates data solutions. The problems shift from designing elegant solutions toward identifying and addressing specific problems organizations experience. This involves, in all three perspectives, an interpretation of the data that was unrelated to the specific problem the organization experienced, thus indicating distinct expertise of the data scientists.

In each case, however, the accounts also pertain to instances where the data science role is already part of the respective organization. This signals commitment and thereby confounds the degree to which their perspectives reflect distinct data science expertise, and was not in fact facilitated through extended relationships that are part of their formal affiliation. The next two sections address this question by considering data science around and beyond organizational boundaries.

5.2 Hiring

As data scientists get hired, they lack trust and extended relationships that could have confounded the significance of their expertise in the previous section relative to bureaucratic task definitions. Hence, by considering perspectives on the hiring processes, we can more clearly understand the role of a distinct data science expertise net of those variables.

This technical context lends itself to hiring approaches that ask candidates to solve formal exercises. One speaker from a prominent technology company for example pointed out that he asks job applicants to be able to articulate their approaches to problems in formal models. A similar comment, although made as more of an aside, suggested a popular textbook on statistics and machine learning as correct response to the interview question which single book a candidate would want to bring to an island. More technologically concerned discussions have been around as specific skill definitions as whether candidates should know how to implement MapReduce or not, a question we encounter later on. All these suggestions assume clearly defined formal knowledge, as formula notations, literature and coding skills. These interpretations of relevant knowledge resonate with formal stocks of knowledge lawyers, medical

doctors and other learned occupations signal through certificates. At the same time, it is at odds with the accounts of many of who were on the job already.

It is telling for data science that elsewhere the formal approach has led to unsatisfactory results, as we learn from Riley:

> Ahm, in the early days, we would do kind of what most people do, you know, we would bring people in and give them a bunch of logic problems, and, you know, whiteboard stuff, and, you know, try to assess whether they are smart. And we didn't hire anyone wrong, like we didn't bring on people who were necessarily bad, but I think we just got terrible reason on people.

Their changes indicate what the formal testing was missing, in their experience:

> And so ultimately what we decided is that it is a lot easier just to see how somebody will do just the work that we have to do. And so now, it's really straightforward, like people come in, they sit with the team for a day, we given them a computer with access to live data, we ask them a question and give them eight hours to work on it. And over the course of the day we work with them, we hang out with them over lunch. They get a sense for what it is like to work on the team, we get a sense for what it is like to work with them, and then at the end of the day they present their findings to us.

As was more at the center in the previous accounts focusing on formal knowledge, this proposition also connects to technical concerns. The idea here is that the final presentation would allow to inquire about the specific technical decisions the candidate made because the time constraint would also reveal skill limitations. More important than noting that they would not completely fall out of the assessment, the novel idea is to reveal some critical qualifications that seem to be embedded in informal interactions and forms of practice that only emerge through them and not the formal constraints of an interview.

The original intuition to test arguably relevant knowledge, fits the idea of relying on bureaucratic task definitions. The shift we see therefore recovers the challenge of data science work on the level of articulating its basis; because there are neither formal indicators nor task descriptions, they have to work with data nerds just to determine skills and fit. Whereas the previous section showed challenges with implementing data science expertise, here we see concerns with defining a context that allows to recognize them quickly and comprehensively, and that this context differs from existing approaches. Because applicants aim to persuade the person hiring of their fit for the job, the redefinition of the hiring process indicates that reliable signals require persuasive strategies that differ from those necessary for other bureaucratic tasks.

The concerns motivating such an approach therefore have some important implications for data science. The necessary skills are complex enough that they require a day of collaboration to surface. At the same time, they also seem universal enough so that training itself is not central here. In other words,

the focus is not on multi-year collaborations to teach specialized knowledge, as has been shown necessary especially in scientific laboratories (Latour and Woolgar 1986). The one-day interactions we see proposed here thus resemble more contact improvisation than inverted heterarchies, or more of Aaron Swartz's style of work than of Linus Torvalds's. The speaker sees a challenge in finding nerds with appropriate skills. This is different from worries over the general availability of these skills, with difficulties instead stemming from formally specifying and recognizing them. This account makes the implicit assumption of sufficiently common knowledge such that applicants are able to join their team just for a day.

This assumption is frequently unmet. Some common strategies to screen for adequate skills in advance beyond traditional CVs are requests of (coding) project portfolios, for example in the form of GitHub and stack overflow. Michael, who we hear next from, leads a project that aims to solve this systematically. It aims to prepare nerds for data science careers. The application process involves, among other components, quantitative challenges problems. The way how applicants perform shows us, for example, that …

> … only about an eighth, let's say, ahm, so half of the people that got it right [a challenge problem involving coding that was part of an application process], really got it right in a way that made you confident about the way they coded. And for the people in the room who are programmers, basically the question required a little bit of recursion, ahm, it turns out you can write recursion, it never even occurred to me, but you can write recursion by saying `function1` calls `function2`, and `function2` calls `function3`, `function3` calls `function4`, and you knew because of the way they structure the problem, it would stop after ten. So, you could do that, and write the body of the function the same way and just use copy and paste, and, ah, be creative, but I think many people here would probably get yelled at in their coding reviews if they did that, I know I certainly would have if I did that ….

Here a combination of formal testing and informal review reveals important nuances in data science expertise. Relative to Riley's approach above, we can note that at least once data science hiring scales up beyond selecting one specific candidate, there is an eye on technical detail, not extended interactions. Still, the formal knowledge expected in this anonymous process, is part of a system defined by experiences from performance reviews, not textbooks.

At least some of the technical skills data science draws on therefore do fit formal interviews we know from bureaucratic jobs, even though the early account above replaced them. At the same time, the performance review basis still indicates that relevant skills require prolonged collaboration. To complicate things further, Michael is able to relatively formally index these skills, in code functions, although informal interactions inform them. This pattern does not reflect the formal knowledge of learned professions. No

less relevant, however, is the way in which Michael describes the result of his formal tests, which quantitatively showed the relationships between the type of language and test performance. The level of detail on which Michael identified skill deficits among some of his applicants signals his identification with the audience he addresses. Although bureaucratic organizations, as well as groups operating through informal expertise, tend to induce highly specialized tasks, which the note on coding reviews is consistent with, Michael assumes a common understanding of the particularity of the situation. In other words, the structure of this substantively specialized knowledge is shared more broadly. We can see this also in Michael's way of leveraging the formality of functions. The idea of functions widely describes short scripts of computer code (the rule of thumb is they should fit on a screen in order for the coder to keep an overview and avoid bugs in the code). On this basis, the speaker is able to share his observations with all those familiar with coding, and regardless of their formal programming language. The idea of a function therefore captures the moment of encoding informal substance into an abstract representation.

There are also overt attempts to impose formalism on the kind of information about a candidate's expertise otherwise revealed through day-long collaboration, as the following appeal to hiring managers indicates. The formalization sorts data science roles into one of three buckets of internal quantitative analyses, machine learning and product-orientation. Jake, who proposes this strategy, has derived these types of tasks from his experience observing the field as part of an initiative not unlike Michael's, except that this one specializes on training PhD degree holders for data science positions. From that experience and interactions with his advisees he recommends …

> … that [hiring managers or team leaders in the audience] get clear on which one of these buckets you predominantly fit in, or which one the role you are hiring for is going to be for. Cause the number one thing we hear from the PhDs in our program when they're considering companies, or who they're going to talk to, is, you know, they say, 'I get what the company is doing at a high level, I get that the data is interesting, but what am I going to be doing, what am I specifically going to be doing if I take this job?' And by getting really clear on which one of these three categories your role fits in, and then articulating that throughout the process, whether it is in your job description, whether it is through the interview process, or when you're trying to sell that candidate to join the team, that's gonna be really help to clarify for them whether it's a fit or not. It will help to attract the right people to you. And also, frankly, weed out the ones that might not be a fit.

In technical terms, they are calling for bureaucratic definitions. This uneasiness with too little control is consistent with that in other modern technology organizations as well, against attempts to get by without bureaucratic rules (Turco 2016). In this context, however, it is orthogonal to accounts of practicing data scientists who often emphasize the diversity of these tasks. This classification into different types of

data science was inferred and not proposed to him by practicing data scientists. Someone in the audience directly challenged it during the Q&A later on:

> Q. I just wanted to push you a little bit on your idea that another mistake companies make is not defining exactly where the data scientist is going to be adding value. It seems to me, especially cause there is a lot of talk right now how the whole field of data science is a little bit vague right now, no one knows exactly what it means, part of the value your graduates could add would be looking at the big picture and saying look, here is where I could add value, here's what I could be doing.

Qualifying his point, Jakes responds:

> A. Agreed, yeah, and so, it's definitely a balance, so my experience; I'm not saying you need to get super detailed, precise, and say for the next five years this is the only thing you're gonna be doing, ahm, but what I've basically seen is a little bit too much into the other extreme, where, which one are you gonna be doing, 'well, this and that, and the other' and they don't have a concrete idea, so I think the point is not so much to pigeonhole, because most people with these backgrounds are really diverse, you know, can really do a diverse number of things, but almost more just give a picture of what you're gonna be doing in the first six months, just to give people a really concrete idea of the specific projects they'll work on. And the best way is actually not even to say for sure it's in this category forever, but actually just to come up with a very specific example, of stuff they can work on and that really helps to get them excited, to get them comfortable about, you know, why it makes sense to join that company. Because that's one of the biggest fears when I talk to fellows, 'I know the company is good, I know the data is interesting, but what am I gonna be doing, and, and, and what does that look like, and so it just gives them more comfort.'

This interaction reveals several aspects of data science in formal organizations. One pertains to the central question of the degree to which organizations script data science work. Accounts have so far associated informal considerations and formal explanations as alternatives to bureaucratic descriptions, on their own. This could be seen across instances ranging from trouble with other departments for expanding storage resources or obtaining access to data to the formal technicalities of collecting log data in the first place. Contrary, here Jake repeatedly returns to impressions of his PhD fellows for an argument that seeks to draw formal distinctions. This suggests that the drive to identify clear boundaries is better explained by his interpretation of the very plausible fears of those who have experienced the struggle of defining a research problem in academic sciences during the PhD training. From research we know that in the scientific setting problems most often follow from addressing a series of specialized questions within a larger field (Stinchcombe 2001). Organizational problems are not so clearly defined with respect to quantitative solutions, which we could see for instance in John's scheduling solution at the beginning of this chapter.

We see some direct traces of academic sciences. We can recall the research on academic disciplinary processes that they facilitate specialization without following bureaucratic rules. Considering Jake's background in academic research that he presented to these data nerds as well, his views likely reflect his earlier role in science, rather than practical data science. It is therefore plausible that these

experiences are real for those former academics he now trains, even if they contradict both the audience's image of data science and accounts of experienced data scientists. Moreover, these data scientists likely experience it as well, but show a willingness to embrace such messiness, as part of the data science role, instead of undermining it through formal task specification. Others have mentors, advisors or managers to turn to. As a community without clearly defined membership, let alone leadership, data sciences lacks this opportunity.[54]

Toward gaining a better understanding of the data science role, the struggle with uncertainty nerds experience, and the wider expectations that they deal with, raises the question of how data nerds navigate this ambiguity such that it justifies a formal employment commitment. Settings that require access to potentially confidential information on a short-term basis amplify this problem. I consider accounts of their relations with short-term clients next in order to better specify the source of ambiguity in the data science role.

5.3 Consulting

Doctors advise patients, lawyers clients, and priests parishioners. Such consulting is a central feature of expert autonomy (Freidson 1988). Whereas diseases, contracts and sins are often formally recognized, in data science, many consulting engagements address a specific client problem. Others are common as well. Ad placement and product recommendation, for instance, constitute such obvious tasks that data scientists joke how "so far [we] have devoted their skills to make people buy shit." Because this application of data science is so prominent, I begin with an account of ad placement that warns against hidden problems. The remaining part of this section illustrates some general problems in otherwise less frequent but more interesting substantive cases.

We first learn that the problem is not as obvious as we might have thought:

> By the way, you really, really, really don't want me to optimize clicks. If you ever considered, please don't. Because what happens is, okay, here is the secret, if you ever want to have a great click-through campaign, all you need to do is to show the ad on the flashlight app. I don't care what you're advertising. It's a whole bunch of people fumbling in the dark. They will click on it eventually. So, you don't want to let me and predictive modeling loose on a bad proxy for an outcome, because all that's going to find, is probably not what you intended.

[54] That we could see this ambiguity through audience participation indexes the common data science understanding. Jake frames his argument as a form of advice for those who make hiring decisions. Yet he encounters opposition. Participants attend these events to learn about new data technologies, their utilities and applications. We have seen before that some of the details may appear magical. In other instances, the audience catches the magic. The accounts we consider therefore reflect less authoritative preaching than thoughtful presentation as the basis of interaction in the technology community.

The question this account claims to solve, with significant cynicism to be sure, pertains to the placement of an ad, and the caution expresses the familiar phrase that "correlation is not causation," in industry jargon. The background to this critique is that advertisers pay for each placement. If they pay for an ad placement in which users are likely to click out of chance, but not because of the content of an advertisement, the click is very unlikely to turn into revenue. Beyond these technical debates, we learn that data science expertise is not obvious, even when the problem seems to be.

Claudia, who just issued that warning, also articulates more targeted criticism that the ideal case of observing a user with and without ad is impossible. She still has ideas how to deal with it:

> So ultimately what you need is something I call alternative histories, or counter factual, you would love to know for me what happens in both cases, showing me the ad and not showing me the ad. And you can't really get this. Not a matter of big data or more data, it's a matter of this just doesn't exist. So, we can't do this. What we can do, however, pretty obvious, is to decompose it into two models. Okay, let me build one model that predicts what happens to people with ads, and let me build a second model to people without ads. I'm not building one model here, I'm basically saying what I need is having two separate models. And then now you put both models to work, to predict the counterfactual. So I never really take into account what happened, I just train two models that can do both, and I can target people based on the difference of the two predictions. Again, you need to do things random and so on, I'm not saying it's easy, but I'm saying it's possible.

The audience gets a fair amount of analytical detail. In substantive terms, Claudia tries to address the practical problem that it is impossible to compare whether a person, let's say Jamie, visits a website, let's say for athletic clothing, upon seeing an ad to an alternative scenario in which Jamie did not see the ad, without reversing history. Instead Claudia compares Jamie to similar others. Similarity, which is often defined as gender, age, income, and so on, here follows from browsing histories, that is prior website visits such as The New York Times, ESPN, Gmail, and all the hundreds of thousands more, because that is easily available and rich in information. Instead of asking whether Jamie and, say, Taylor are both male or female, it considers whether they have visited similar websites. In this setup, one group of those who are similar to Jamie were presented with an ad, the other was not. If both groups are roughly equally likely to visit the website for athletic clothing, although just one saw the ad, there seems to be little need for presenting Jamie with an ad. If the difference is large, it seems useful.

This is not what Claudia said, of course. Instead of talking about Jamie and Taylor and interest in athletic fashion and The New York Times, Claudia talks about two models. Two implications follow. First, Claudia assumes that how these models work is commonly known in the group she speaks to, assuming that she expects others understand her solution. Second, we directly see how much time the formal description of models saves compared to explaining and substance. Moreover, whereas here we have

thought about Jamie and ads, thinking about models accommodates all other people and many more problems as well.

As clever and efficient as this may sound, however, Claudia also issues another warning:

> Right now, nobody has ever approached me and asked me, in my position as running campaigns for brands, to do that. I could, I probably get killed and kicked off the plan in no time because my conversion rate doing this is probably not nearly as high as it would be if I just optimized towards conversions, but my sense is this is what you should be asking me to do, and I can do it, right now nobody really lets me.

Although Claudia has not had the opportunity to implement her preferred solution, she did avoid the flashlight app problem with a design that makes her more comfortable and addresses the client's outcome of interest, as imperfect as it might be from Claudia's perspective. In short, this account indicates a conflict about translating substantive interest in analytical logic. It relies on the rhetoric of statistical formalism. In introductory remarks, moreover, Claudia mentions the technological challenge of running these models in the time it takes for a user to call up a website and for the website to load and display an ad. This technical challenge is not relevant for the client interaction, however.[55]

What about our main focus here? Whereas organizations and bureaucratic definitions had little bearing on data science tasks so far, client interactions shape it here. They are also more sporadic than bureaucratic reporting relationships. As a result, this type of coordination resonates more with the contact improvisation in Aaron Swartz's work than Bill Gates's hierarchies. They are also familiar though from the learned professions where neither lawyers nor doctors work independently of their clients and patients. I consider dynamics of this relationship with respect to expert status later on. Meanwhile we can note how data nerds directly reconcile their arcane knowledge with the respective practical problem and its context, not unlike a specific deal and a legal contract securing it. In addition to technical expertise, here we see the careful analytical framing this task entails. In addition to the substantive specificities, moreover, the data nerd has to conceive of the underlying problem in a way that takes into account the relationship with the client. Whereas the instance of designing a scheduling mechanism required the choice of an appropriate method, here we see adjustment of a persuasive implementation out of several possible applications of an appropriate method.

[55] This also resonates with other accounts that point out the preference of clients for interpretable models over those that might perform better but that remain incomprehensible to them.

The struggle with accommodating client requests into quantitative frameworks goes beyond the modeling strategy. This is easy to see once we move into messier problems, such as the following recollection of experiences in educational problems, which Aaron tries to address:

> The first thing that we learned is that looking at the education system, it's not one-size-fits-all, it's everybody thinks that they're unique. We learned this lesson, we started giving schools, you know, standard, valid, reliable, best-practice feedback surveys. And they'd say to us, 'alright, fantastic, now our committee is gonna get together and edit the survey to make it for our district, 'cause there is no way that our district and the guys next door are going to wanna do the same thing.' And it's, you know, a valid perspective.

Implementing data strategies with clients is not limited to getting the technological infrastructure or the modeling in order. In times of abundant quantification, it encompasses specifying the goal in a respective context. We can go back further in time in order to recognize the magnitude of such an effort. Earlier initiatives of data collection required broad standardization because technology and organizational resources did not allow for much local variability (Conk 1980). And the clients, teachers in this case, relentlessly point this problem out. The ad-click model benefited from the automatically recorded digital traces, though also remained limited to them. Education still leaves relatively few digital traces. But data collection is easier. Once these problems are dealt with, now in education …

> … the question is what works, and the interesting thing is, we don't know. There's no answer for what works, there's no answer for what makes the best teacher. It's about teaching style, it's about classroom, and its challenges. If you believe in a universe where there are very few if any data scientists, I mean very few talented data analysts in education, and all these different contexts and everything is different, how do we actually figure out what works? If you ask about best practices for what a great teacher does, there is no answer. And that's one of the scariest things in this work is realizing that we just don't know.

The tone reminds of those PhD fellows who interviewed organizations that did not specify their tasks in one of the three buckets Jake described above. They are not the only ones who experience fear. Once again, these steps offer none of the direct guidance of academic requirements defining valid data, nor the confidence grounded in prior experiences that would speak to the relevance of this data. All these uncertainties need to be resolved before data nerds think about models at all. Yet both tasks are seen to be relevant in data science. Encountering then such a context without prior experiences sounds scary indeed. It is telling, at the same time, that this focus on data provides sufficient flexibility as to lead this data science project into so uncertain territory.

Both consulting experiences signal some frustration and compromise. In addition to revealing variable applications of standard methods and ways of counting observations, Claudia and Aaron therefore also share their willingness to respond to those they work for. At least Claudia is not shy to admit that this willingness results from market mechanisms. Because the methods they use come from

academic contexts protected from these mechanisms, the accounts once again show the implications of such abstract processes on the level of technical expertise, complementing informal coding reviews and formal task performance, and impressions from peer organizations' internal technology infrastructure and skills.

These accounts also begin to suggest that the prolonged relationships that are part of formal employment and the prospect of which may motivate hiring decisions are not critical to engaging with client problems from a quantitative perspective. Yet, both involved a significant degree of specialization and familiarity with a given problem most clearly. The previous accounts have shown instances where knowledge necessary for data science was specific but not specialized or proprietary in a way that they would require prolonged relationships. The following account shows such relationships are not critical, even in some of those problems where specialized expertise seems central to the task. It is taken from a summary of an online competition in data analysis. The proposed solutions indicate the flexibility of approaches in such conditions.

NASA defined the question of this quantitative problem as mapping dark matter from satellite images. NASA provided a competition website with images that had dark matter encoded in them. In this competition anybody could sign up and try to build a statistical model that automates the recognition of dark matter in a way that improves on NASA's own solution. We see the result of this setup here. This is the reaction to a non-specialist solution:

> Whitehouse.gov published an article saying massive new breakthrough in astrophysics, Cambridge glaciologist, from the home of people such as Einstein and Newton, cracked the field. Literally Martin [the glaciologist] then gets up on Twitter and brags about how he got compared to the likes of Einstein and Newton, except that then, three days later, a professor from Qatar passes him.
> The professor from Qatar had an approach he used for Arabic signature verification, which is to compare two pieces of hand written Arabic text to see whether the edges are aligned appropriately and the kind of the slight fuzziness was the same and so forth to try and guess whether it is by the same hand. And it turned out that his algorithms for that, which he combined with other algorithms he had produced in a past life, which was trying to restore audio to old movies, and he had ideas from both and he combined them together, and he found an even better way to do the galaxy image signal detection. And so he [the professor from Qatar] passed him [the glaciologist from Cambridge].

Just to summarize, we learn about a NASA initiative that sought help in understanding satellite images. The most promising proposals came from nerds with extremely clever ideas but no substantive expertise whatsoever. Unlike in the previous two instances, the question was predefined here. Some might dismiss this setup as distorting most typical conditions for data science. For instance, it ignores the complication of working in a context that fails to define the problem and data and all the uncertainty Aaron

encountered above, because no two school districts were alike. Ignoring this case here for our analytical purposes would overlook the relevance of the applications of solutions that were originally not at all designed to address such a question. To be sure, the competition organizers helped connect the problem to a wide variety of potential solutions, but those solutions nevertheless indicate significant flexibility, given their diverse origins in glaciology and Arabic handwriting. Therefore we see that the knowledge itself is not bound in the same way, even if the application of these approaches requires informal interactions in teams, formal relationships to organizations, or at least specialization and hence reputation in a specific field. Indeed, specifically because of the curated nature of the problem, one could expect that others had already reaped the low hanging fruits. It is therefore all the more striking that two so different and substantively unrelated solutions improved NASA's own attempts enough to receive recognition from the White House.

Aside from the different approaches to encoding substantively messy and ambiguous problems into analytical frameworks, each task also brings with it different consequences. If Jamie, to recall the previous setting, does not click the ad that an advertiser paid for to present it to Jamie on the basis of a statistical model, little harm is done.[56] Foregone educational opportunities can have consequences down the road and NASA receives much federal funding for often risky projects. These are all areas in which data nerds promise improvements. Considering the arcane background of their expertise, however, their expertise is in data and models, which are most often separate considerations from all the details that can go wrong. Yet another complication emerges even after all these technicalities we have seen here are sorted out, as we see next.

Utilizing the flexibility of statistical models outside of the mediated and controlled context of curated competitions is also possible and raises a new set of concerns, as Jake describes here:

> [W]e face this problem [feeling responsible for consequences] building a project with Amnesty International, predicting human rights violations. Built a great logistic model that'd help us understand, based on the severity of a report of a human rights violation, which ones Amnesty International go after first. And it was great, and our classification accuracy was high, and we handed it over, and we started to sweat, 'cause we had that moment, it was like what if this doesn't work, how will they know, what did we not think of, we're not in Jordan, being held prisoner, what things aren't we thinking about, and how will other people be able to use these tools when they're not exactly sure what success looks like. Now this is an extreme case, but I think that's something that we as algorithm designers are creating across the world and really begs the question at least thinking about what kind of bias are we creating and how do we even expose them.

[56] Claudia, the speaker who introduced the online advertisement case, cites this lack of severe consequences in advertising as reason for her move away from analyzing medical data before.

Jake thus integrates concerns with methods and technology not only with those of clients, but also their external effects. This is the most critical and explicit of the accounts concerned with client experiences, which, however, have all noted the complications amid the widely popularized opportunities in data science. This moral thinking resonates with several calls in the data science community for defining ethical standards, which are well-known from many occupations and part of persuasively defining a distinct community (Abbott 1983).

As a few times already, we once again encounter fear. Only this time it is different compared to that of the PhD fellows in the previous section or the educational project described above, where it rooted in uncertainty over what to do next. Here, instead, we see the experience of realizing that deploying data and models in ways outside of one's specialty increases the chance of unanticipated consequences, or at least uncertainty to this effect. The practice of taking models and applying them to problems unlike those they were originally designed for is as old as the statistics discipline itself, rooting in eugenicist political projects (MacKenzie 1978, 1981). Thus, what differs today is not the mere practice of encountering a new empirical context, but much more the consequences that follow immediately from the modern type of data implementation that generate directly or facilitate indirectly almost instantaneous reactions.

Considering these consulting instances has shown how client requests define data science applications. At the same time, the data scientists were able to design the specific solutions on their own. Most significant with respect to understanding the principles of data science expertise are those observations that show how data scientists translate their solutions into different kinds of contexts, thereby reconciling mundane problems with their arcane knowledge.

5.4 Friction

Data science enters organizations in different ways; it comes as new departments, as new roles in existing departments, or in consulting capacities. In some instances, these organizations are young, and data science partly shapes them. In other instances, traditional organizations absorb data science expertise. Data nerds have here repeatedly argued how these organizations can benefit, but others also see room for conflict.

And of course, at least some of the organizations considering data science already collect data, and hence tasks assigned to existing functions and offices, as we see here:

> And so, what do people do? They do this, they do the stopgap, and this is what we see virtually everywhere we look [slide]: Hadoop, little analytical database in the middle, and then the analyst [inaudible] against the analytical database. So, what is this, what is the effect of this? Well, the analyst has this [inaudible] view of the siloed data, there's a whole IT team moving the data back and forth between here [pointing at Hadoop and analytical database], and if you ask the wrong question, you want to drill down into something that wasn't anticipated, guess what, go back to IT and wait another six months until you get that data. We don't think that's a great solution, nobody we have talked to seemed to have liked this model, but it is the status quo.

Six months is surely a stretch, but the kind of experience of depending on others constitutes a plausible concern. This project proposes a solution. While this specific approach is not so relevant here, we now see why Riley happily reported that his data science team had taken control over the data warehouse. If they had not, his team would have been at another's goodwill. This note also echoes a problem that follows from other accounts, especially working with log data, that emphasized the expectation of data scientists to retrieve their data. This specific solution suggests a clear bureaucratic specification of tasks that may have seemed unclear because of the new awareness of data. Other observations complicate this picture.

Here someone shares his observations in the subsequent Q&A:

> I think the word that you used is the stress that is going on within companies now, because of all this data, and I'm finding, kind of on the front lines a little bit, that the, it's not just stress, it's, it is almost like a territorial thing going on right now, between IT and other parts of the company that are growing, and I guess my question is, you know, not where, who owns the data, the company owns the data, but what do you do, what's the collective wisdom when like IT doesn't want to let go of the data because they think that might mean they're not as valuable and that their jobs are on the line, right. I mean, what do you do with that kind of, it's not even stress, it's more like a real conflict going on within companies. And the top executive is saying get this done, you know, what do you, any suggestions, collective wisdom?

The tone here is very different. Territorial "things," or, forgetting about political correctness, "wars," rarely respond to new contractual arrangements, not immediately. The "collective wisdom," at least as far as these presenters reflect it, again emphasize technical rather than organizational changes as remedy to such problems. Speakers were wise enough not to suggest alternative distributions of authority and instead proposed technologies for facilitating collaborations. We have seen this acknowledgement of existing bureaucratic structures before. Returning to the initial metaphor, the collective wisdom suggests the exploration of new territory, instead of redistributing existing territory. Although IT emerges here as the key function under threat from data science, there is no question concerning its responsibility for the technological infrastructure, which also data science utilizes, peacefully.

Perhaps the careful responses from speakers came from their own entanglement in organizational politics. In the following response to a question regarding the role of consultants in the future, such

politics would be less relevant. The different relationships to technology, and thus implicitly skill sets, nevertheless emerge again, and with more specificity:

> … one of the trends here is to try to minimize the number of consultants I think necessary. And I think there's two types of consultants you could talk about, one is the IT consultant to make all this technology work, and then there is more the data scientist type of consultant. I think on the IT side, I think there is a lot of opportunity for consulting right now, because all these technologies are so new and so complicated and so hard to use, but I … think over time these will become more productized and that IT consulting will decline. And I think that's natural as the industry matures.

And prospects for data science look a bit different:

> But I think also from the data science perspective, I think that's another interesting area where there's incredibly high demand and not enough supply. I mean, the true [data scientists] and, you know, we have one at our company, … I mean they are extremely hard to find. So, I think that's an opportunity as long as, you know, these insights are not obvious to folks. But I think that's also what some of these guys are working on as well, to make that easier to see, easier to visualize, even for the non-data scientist.

Once again focusing on the role of technology, this view adds a possible end for data science to its possible trajectories. It lacks specificity in order for us to see the basis for this prognosis.[57] More important here is the distinction, in which data science, at least for the moment, depends less on technological progress, compared to IT, and for its success and hence status. This reinforces the understanding that data science is not stealing an entire jurisdiction, but diverts attention indirectly. To be sure, all the comments on data science's distinct technological positions implicitly also reinforce the question of how it might fit into the existing bureaucratic arrangement.

No one here discusses overt conflicts. Nevertheless, all three accounts consistently reflect the potential for friction between IT and data nerds, with some confirmation from front line reports. The audience comment and the reaction from a speaker to another one also suggest however that the basis for arguments are not clearly defined. And how could it, if we have consistently seen ambiguity over data science tasks.

So let us consider what a data nerd has to say to this:

> I'm a data scientist … [and] … I'm hearing two things, so I want to talk about how you, how both could be true at the same time. The first thing I am hearing is that is almost like a war between the IT/data people and business. And then the other thing I'm hearing is that data scientists are hard to find. Let me make a suggestion why both of these could be true. … I didn't come to be a data scientist to not be a business person. I consider myself a business person. So, to create that dichotomy between the data people and the business people, I think it is false, and it makes there be more friction than there needs to be. One of the problems with that mindset, is that data people are expected to just implement things that are to be decided by the business. … [W]e don't have to think of this as a war. We should instead think of incorporating data people into the business decisions, and in particular spending a lot of time cleaning the data carefully, and asking the questions right, because as a data person I see the biggest mistakes that business people do is they ask the wrong questions.

[57] The following chapter considers several such predictions.

This elegant point of how data science tries to embrace business rhetorically removes grounds for conflict. At the same time, it structurally intensifies reason for the friction. Indeed, at least in front of audiences, data scientists rarely expressed overt conflict with other organizational functions. Understanding of data science as an emerging role would partly explain why nerds don't share in public how established functions may slow down their work. They still depend on them often on a daily basis and lack the strength a formally organized group would have.

In other words, the intellectual project for data nerds to improve and specify the definition of their work, tasks and relevant knowledge encounters the practical challenge of implementing its results in the context of competing occupations. We have seen throughout this chapter as well that data scientists do not try to provoke these conflicts. They seek allies in the organizational context as well. These dynamics enhance the need for rhetorical narratives around the technical and arcane data application. They also show how improvising on the basis of just shallow and fleeting contacts, of the kind we associate with Aaron Swartz's work, quickly encounters resistance from those that operate within bureaucratic task definitions. Instead of carrying out these conflicts, data nerds try to negotiate. They thereby preserve the utility of otherwise agitating task division. I consider the implications of their negotiations at the end of the chapter, and what it is they preserve in the next one. Meanwhile, that others have started to recognize conflicts in spite of data science's silence speaks to the significance of the overall change following from a shifting role and definition of technology. That shift involves a transition from an information infrastructure toward an integrated system of data storage, processing and actions implies.

Chapter overview

How is the data nerd community holding up against formal organizations? Moving on from the technological context of the previous chapter, how can we understand data science in the organizational setting? Data nerds once again describe their work and expertise with references to technological instead of substantive problems. This does not mean that technology remains equally relevant as in the previous chapter. Accounts also appear to be shifting very noticeably toward analytical concerns, which had emerged as the common core of the community in the previous chapter already.

Across the different instances considered here, data science nerds describe how they had to work with technology in order to access and use data, but not the specificities of how the technology is set up.

With the details considered before, we can better understand that just this additional question is complex enough precisely because the more basic problem of how to store the data has different solutions to it. This variability here meets the different demands in organizations. From this perspective, data science reconciles those two sides, the technological with the organizational. In short, moving on from the question that was more about the technology needed to store and transport the data, here it is about data structures themselves and how they fit into analytical strategies that address organizational problems.

Such arrangements can enhance the chances for unwelcome effects. If data science expertise adds significantly to the feasibility of data applications, and nerds apply it as they work with the data of an organization, control becomes ambiguous. This additional complication for public consequences creates individual opportunities in this area of work. These opportunities are not limited to new technical problems around data. Without sufficient competencies in the organizations that generate the data, this work also requires nerds to identify useful applications. This finding itself raises more questions than it answers with respect to what it takes to realize such opportunities, except that these competencies are not tied to formal boundaries.

Turning to individual opportunities and definitions of work, this section has reinforced the divide the previous one left us with between bureaucratic and sporadic task definitions that overlap but create inconclusive boundaries. The job setting reflects this tension. It revealed accounts expecting specialized understandings of what candidates would need to know as well as of the tasks that are already established in organizations, and under attack through data science. Yet, we also saw significant initiative from the side of the data nerds in defining their own roles. In other words, on the one hand there was a clear sense to shut down hobbyists, as Bill Gates put it, for instance with respect to clients who demand solutions consistent with their interpretations, but not an expert view of their effectiveness. More reflecting informal efforts of Torvalds, but somewhat closer to Swartz, we also encountered specialization that was not bureaucratically but substantively induced and in those contexts relied on direct contact, such as from teacher-clients who emphasized their school's unique needs. Indeed, we found clear momentum with respect to approaches that challenge the bureaucratic division of tasks, for instance by installing new cultures. Some of these efforts also signaled autonomous status of the kind we associate with Mills's view, such as in the White House commentaries on successful data solutions. Similar to the earlier

references to the overall significance of today's data applications, this sporadic note offers no basis for drawing systematic conclusions as to how such autonomous and anonymous definitions of work enter the organizational context.

The most significant observation in the organizational context is therefore again the opposition between Gates's and Swartz's definitions of technology work. Their tension is also more striking here than in the previous chapter. There we noted that sporadic and improvised rebellion against the corporate hegemony is easy relative to Torvalds's more coordinated alternatives, because it does not aim for continuous solutions. While that pertains to the use of technology, it is not so clear in the organizational arrangements deploying it. Instead of deploying technological capabilities for one's own purpose, here we have seen data nerds defining applications for organizations. We are therefore left with the question of how Aaron Swartz could operate in a more coordinated context of concrete organizational boundaries. I turn to this question in the next chapters.

Finally, the rhetoric has shifted as well. Because these questions are directly salient to clients, relative to the technical specificities of a stack, there is not much data science has to argue for in the creative ways we have seen among technology experts more generally in chapter four, the first one analyzing the New York setting. As a result, no ambitious analogies are necessary here. Instead, the question has become how data science expresses the clear claim it makes over problems others may consider theirs. Data science could make a straightforward argument based on its more advanced knowledge, and reflecting the arrogance often associated with professional work (Larson 1980, 1977, Abbott 1981). This could directly follow from the practice of creating immediate value through analyzing data (Davenport and Patil 2012). We have instead seen evidence of an alternative, more diplomatic strategy of taking a rhetorical detour of directly considering others and articulating contributions in terms of practical complementarities instead of technical superiority. This distinction offers analytical leverage for the purpose of identifying some further contours of an emergent thought community of data nerds.

Contours: Persuasion

A dominant idea in contemporary arguments of profession formation is the friction that arises as emerging groups interfere with the work of existing occupations (Abbott 1988). We could see traces of friction in the context of data science, although accounts that emphasize compromise were more

prevalent. Most of the accounts we have seen here in the organizational context show evidence of one out of two strategies. Nerds following the first relied on rational arguments to convince others of its utility. Instances ranged from citing technically superior mechanisms of data analysis, efficiency-gains, statistical observations of competencies, and qualitative observations of fit. Nerds following the second strategy relied on relational explanations. These ranged from turning to peers for inspiration of how to manage data in the organization, agreeing to compromise with clients, invoking institutional status of stakeholders and acknowledging tradition. Why do we see these subtle strategies, instead of references to weaknesses other groups may have? To be sure, the ambiguities over the specific division of labor we see as soon as the audience speaks up indicates that those novel activities could have been provided by those functions that have recognized a threat from data science, even if data science emphasizes its distinct contributions and that they complement existing functions. Data nerds therefore have to claim them.

There are three immediate explanations for these observations. First, the public settings might undermine overt attacks on other functions. Second, the organizational setting could support the data science role against others. Third, perhaps data nerds overlook potential infringement because they are new to the organizational setting. After considering each one, I propose a fourth explanation, that data nerds exploit uncertain role relations and thereby claim a superior position.

One argument for explaining why accounts in this chapter have not shown more overt criticism of competing roles has to consider the public setting of these events, because public exposure induces powerful disciplinary mechanisms (Foucault 1995). And most events target not just data nerds but the broader technology community. Hence speakers could never be sure who was in the audience. At the same time, data scientists were forthright when it came to their clients, describing how quickly data science outperformed their solutions and how shortsighted some of their requests are. Interests also differ, to be sure, with respect to clients compared to other organizational functions. It is plausible that speakers seek to persuade potential clients, in a market setting, of their utility even where this undermines their clients' pervious practices, whereas they have no such ambitions for other technical groups in the bureaucratic setting. In those accounts data scientists insist explicitly on their claim over data, especially in its raw format, but often without directly describing the friction that might complicate

precisely those claims. Others described their experiences or ambitions imposing a data culture on their organization more broadly, instead of trying to adopt it as is often seen profitable in the organizational context (Goldberg et al. 2016). It is therefore not all public discipline that prevent data nerds from calling out others. That some freely admit friction here, while many do not, raises the question of whether organizations, or nerds themselves, manage the overt conflict.

DiMaggio and Powell's (1983) idea of institutional isomorphism helps to address this question with respect to the first alternative explanation. An explanation of why data science speakers regularly ignore the IT role that follows from this framework could be that data science has spread as organizations see others define the data science role and implement it as well, and transfer it mimetically. While we saw some evidence of this in Riley and Rachel's accounts, that the nerds promoted this transmission rejects the primary role of organizational support. In such a process we would not expect to see significant friction because the organizational environment defines tasks in ways that avoid conflict with existing functions, or at least mediates them. And this probably is the case at least to some degree. For instance, we can recall one data nerd describing how he turns to leading companies in the data field for ideas of how to organize data in them. On the same note, we also see consistent accounts in the hiring problems, emphasizing the difficulty of designing appropriate selection processes beyond the basic question of formal skills as well as complaints about the inability of organizations to specify data science tasks. This indicates that although the organizations consistently express interest in those services, they remain unable to define the role clearly. As data scientists themselves define their tasks, organizations are less able to separate potentially overlapping functions and prevent conflict. So if not the organizational field, is it the individuals?

Moving on to consider individual explanations, the kinds of claims that reserve tasks on the basis of their fit with the data science process and without acknowledging that this interpretation inflicts upon other functions could be seen to reflect simply naiveté of a young generation on their first job.[58] Perhaps data scientists are indeed unaware of the direct implications for IT of their activities. The entire NASA, after all, took away a problem from those originally responsible for it and made it available for anyone to address. While these are extreme cases, data nerds might be so focused on their own activities that they overlook

[58] See Turco (2016) for an account of similar problems in a more formal setting.

established solutions, from an organizational perspective. At the same time, data scientists were ready to react to confrontation, with respect to IT's role, with ideas of how to reconcile the two sides through technological tasks areas and solutions to interact. General ignorance or perhaps subtleness of this problem among data nerds are thus unlikely as well. If organizations fail to establish the data science role vis-à-vis established roles, and if data nerds are not naive and passive regarding their organizational consequences, how can we explain their consistently covert position amid different ways of articulating their utility?

A context of novel resources provides a basis for comparison to tactics for positional claims in other instances. The one general transformation associated with the emergence of data science is the growth of technological capabilities. Consistent with the conclusions we have drawn so far, this literature points out how external changes reshaped the nature of competition and thus, the field of competitors with subsequently uncertain role relations (Elias 1985, Ginzburg 1992). As tasks shift away from primarily storing data, which requires its efficient structuring, toward arranging it in ways that are most useful for gaining leverage from it, the responsibilities more easily change as well and are not yet bureaucratically defined. In such a setting undirected statements with respect to issues that are clearly relevant for others have the effect of elevating one's own status and a provocation to those others (Leifer 1988, Leifer and Rajah 2000). This pattern closely describes many of the observations from this chapter. Data nerds have repeatedly specified tasks they should take over on a rational basis, but without acknowledging that this takes them away from others, although we could also see that they were aware of them. The strategic role structure literature unmasks that what seems rhetorically persuasive has a strategic aftertaste as well.

Although all accounts revealed instance of how data nerds persuaded others of their utility, they did so in various different ways. Here we see that these different strategies, such as rationalizing data science's effect, or relying on authoritative support, put data nerds in a superior position precisely by leaving those whom they affect unspecified. How can we see this contour so consistently, if it emerges without organizational or direct personal processes clearly involved? While the shared technology offers a plausible direction, we have already seen in the previous chapter that the technology itself would remain

disconnected across different areas of application. It follows that in the next chapters we must ask on what basis this group of nerds forms a coherent role.

6 Work

Technology, software packages, and formal and informal relations in organizations all shape the role of the data nerd. The previous chapter revealed familiar technology nerds solving today's data problems in contradictory ways of accepting hierarchies, but not bureaucratic boundaries. Our observations just partially revealed how these logics fold together, as data nerds consistently invoked informal sources of ideas that have formal correlates as well. The basis on which data nerds encoded the substantive problems into analytical solutions we were not yet able to identify with sufficient clarity as to be able to discern the expertise data science claims as its own. With technology and formal organizations insufficiently explaining these arrangements, we need to consider the specific context of data science work, their skills and the projects in which they apply them.

One reason for such indefinite accounts so far might be that the unit of the formal organization often entails significant ambiguity itself (Carruthers and Espeland 1991). In the presence of such informal activity, observing variability and diversity in data science tasks does not allow us to conclude that this follows from data science expertise itself. In order to understand the definition of their work, we need to consider the specific projects as more granular level of variation. Projects still capture exposure to organizational and institutional effects in as far as they shape data science applications. They also move us closer toward the origin of public concerns. In the case from the introduction on coupons for baby products, for instance, the worry was not with the shopping chain’s main product. The father expressed his discontent over the specific practice of targeting customers on the basis of quantitative analyses of their shopping behavior. Understanding in how far projects provide a scope around data problems is therefore essential for devising ways that address such concerns.

Once we have a better understanding of the project as the specific setting of data science work, we can consider how it unfolds. I organize this analysis around basic skill sets by which data nerds connect technical capabilities, empirical problems and others’ interests. For instance, logistic regression constitutes the most frequently discussed such link—they say it “rules the world.” This evocative characterization clearly over-simplifies the task of applying it. We have seen by now that few data science problems fit standardized examples. It nevertheless reflects the sense from the previous chapters that to data nerds no single external factor is significantly shaping and scripting their expertise, not even

academic research, which defines this method. The tools they know, even relatively simple ones, shape the world they apply them to. This idea, regardless whether true or false, then leads to the question of what they might define these rules to be, what it takes to implement them, and whether we can begin to learn from their implementation, and background, how data science has become so publicly salient compared to other instances of experts relying on similarly arcane knowledge. The practices that are part of the quest to rule the world then reflect a basis for coordinating as an integrated thought community.

The logistic regression interpretation of the data scientists is thus consistent with public concerns resulting from their view of data science's interference with everyday life. Because a model itself does nothing, accounts of skills and how they are relevant index the leverage data nerds propose they can add to the data as such, as well as shape the way it shapes others. This focus also addresses the individual opportunities associated with data science. In this respect, accounts of relevant skills must not be interpreted as an exhaustive list of the things nerds need to know. The accounts instead reveal the principles by which nerds apply skills, as the logistic regression comment for instance indicates by signaling a more casual and consequential interpretation of the fit between problem and tool than others might have entertained. We can turn to our familiar tech nerd characters to consider different principles by which these skills unfold.

While projects can be variably staffed with experts, skills are directly tied to the actors. This should not have any bearing on the Aaron Swartz type of projects. It can affect proprietary and bureaucratic as well as on open projects. Both Gates's and Torvalds's designs of tech work entail specializations, which the former imposes through a hierarchical structure and bureaucratic definitions, and the latter coordinates from the bottom up, including a heterarchy. As Gates unambiguously emphasized the proprietary nature of this kind of work, these skills should be less likely to become widely shared. We can expect accounts of proprietary knowledge to conceal important specifications with a public label. This is irrelevant for the analysis as we are interested here in the variation across proprietary and open projects, not so much the technical details of either one. Against this backdrop, skills in the projects of Torvalds spread more easily because individuals can acquire them without joining another type of technology group entirely. This is of course conditional on their relevance for data problems. At the same time, Gates-style bureaucracies could still dominate these kinds of problems by developing skills specifically

associated with them. Take SPSS as an instance, relative to R. Although it encodes general statistical models, it requires special skills to work with and cannot be adjusted easily to specific purposes or problems. None of these skills speaks directly to the kind of formal knowledge we associate with the learned occupations of Mills's account. In this respect we need to focus on the degree to which they allow to be viewed as distinctive expertise, independent of their respective application and technical background.

I divide this analysis of data science work in two main sections. First, I consider accounts that reflect several key moments throughout data science projects. Second, I analyze different kinds of skills that substantiate the work in the scope of these projects.

6.1 Projects

The data events feature speakers on both technical and substantive projects. While focusing on technology and organizational questions, the previous chapters have already touched upon examples from data cleaning, data accessibility and storage projects, as well as advertising, news, hospitality, education and some others. After focusing there on technological and organizational processes, I now consider the relationship of the scope of project tasks and data science expertise. I draw most heavily on a case that introduces romantic partners to one another online, but also consider other instances for a more comprehensive perspective and for clarifying some underlying processes. We have seen already that in many applications data nerds neither find clear guidance as to the relevant outcome of a data application, nor the ways for reaching it. I use these two problems to organize this analysis of the project setting and scope.

Projects pertain to all four types of technology nerds. We can think of large software projects, like an operating system, and smaller ones within them, such as a text processing application. They can follow a corporate organization, in which hierarchies divide bureaucratically defined tasks, and which would remind of Bill Gates's design of technology work. They can also follow informal coordination, in which developers contribute such specialized components on their own to combine in a larger system in a heterarchical form, such as in the case of Linus Torvalds and his fellow Linux contributors. We can think just as easily of slightly different projects, such as Aaron Swartz's development of the RSS language. While they also have a relatively specific goal, they are not part of a larger coordinated effort. Finally, the

lawyers and medical doctors Mills described solve cases and treat illnesses. Although we typically do not think of them as projects, they share the relevant features as well. The references we have seen data nerds make here to knowledge for “one-off” specialized projects, suggests that this form is common here as well. It follows that although considering projects themselves offers no basis for expecting any one tech-nerd role in the data context, the way in which the projects unfold offers evidence into the set of practices data nerds follow.

6.1.1 Defining ends

A project needs a goal. Goals are often relatively clearly defined for organizations, at least on the higher level of what achievements count. In most basic terms they range from making profit from producing and selling goods or from providing services in the market context, providing public services in the government sector, knowledge and training in education, salvation in religion, and so on. Here we take into account that organizations often entail significant variability amid clearly defined bureaucratic rules that facilitate the pursuit of these goals. As a result, the aims of projects are often more difficult to define. Who is responsible for the definitions?

This can be seen elsewhere in the modern digital context where projects constantly evolve (Stark 2009). And the sporadic observations so far have revealed similar ambiguity. Whereas the instance we have seen early on, of optimizing support staff availability, had a clear goal, the no less obvious placement of online ads gave much room for interpretation. There it was not immediately clear how to address the issue of whether a person who sees an ad would not have purchased the respective product without it. I consider here mainly a case from making romantic introductions with marriage promises because it offers a universally salient context that is as clear, seen in its institutional status and definition, as it is complicated, seen in divorce rates. Vaclav explains the data science view of this problem.

The goal seems clear. Users seek a suitable partner to spend their life with. The data nerd, living the stereotype, turns the romantic problem into a technical one, because:

> You know, for match distribution, we need to decide who to introduce to who, when, and we don’t want to overrun people, right, ah, we don’t want to dump, hey, here’s 50 thousand people compatible to you, you know, there is no way you’re going to get to know each other, one of them and decide who is the right person.

The jargon of “match distribution” describes the question of whom users could be introduced to. Here it signals the perspective of an expert, and one removed from the experience of finding a partner.

Vaclav is concerned with the complication from asking that question for many at the same time. This first consideration toward reaching the goal of matching partners illustrates the work social selection mechanisms do. Once we step out of the context of those with whom we went to school or college, work together, or share other activities, the number of potential partners grows very large very quickly. The goal that seems unambiguous from an individual perspective changes its form when considered in collective terms.

Replacing all these social mechanisms that limit our choices to a manageable number makes the turn to technical abstraction much less surprising. And with this, none of the social complexity needs to concern the data nerd:

> So, let's say for the case of heterosexual matching you have women on one side, men on the other side, you're matching between these two groups. So basically it is a bipartite graph, where some of these edges [between predicted male-female matches] are missing, you know, like some of these edges have been killed by our compatibility system, which says like no, don't introduce these people, because there is not a high enough chance that they will be very happy in a marriage.

Vaclav formalizes the initial problem of introducing two potential partners as a bipartite graph. A bipartite graph is a network analytical concept that assumes two categories of "nodes" (here men and women) and permits relationships only between, but not within these categories (Knoke and Yang 2008). This formalization rests on the cultural idea of heterosexual orientations, as Vaclav clarifies initially. He also notes that some edges will be missing from this graph for a lack of fit, the basis of which we turn to next. For now we see that the social and emotional process of meeting someone and feeling mutual attraction becomes a sheer insurmountable problem as soon as we consider the number of pairs that could technically find themselves in this situation in the digital world of the Internet. The data nerd articulates this selection process through the analytical formalism of graphs. With this setup, another problem remains.

Bipartite graphs only prohibit relationships within the two categories, here of men and women. Besides that, they allow for as many connections, or edges, as there are nodes of the opposite type, multiplied by nodes on the one side. Most marriage institutions, and norms, do not. Nor could users evaluate in reasonable time the potential matches the Internet would yield on the basis of simple characteristics. The data nerd therefore takes one more step:

> … the way you can solve this is that you basically imagine this as a series of pipes, and, ah, where these edges have a capacity of two, or whatever is that match limit over there [pointing to a (man/woman) node on the slide], and here the edges between the users have a capacity of one. And then you can just pump water

> through that, and what you will see is that the matches where there is some water flowing are the matches that you can deliver given these constraints, right. So, the solution, it is like a flow problem [water pump and buckets appear on a slide], and this delivers you the maximum number of matches that, you know, you can deliver, given these constraints.

So this is a data view of the social world. Vaclav formalizes beautifully rich dating experiences as a technical system of pumps, pipes and water buckets. He thereby constrains the Internet to the effect social settings do otherwise, although in different ways and likely with different outcomes. Even improving on the familiar social world, the data solution addresses the otherwise widely occurring preferential attachment problem (Barabási and Albert 1999). Here hopeless pursuits of one's crush delay more promising matches. The data solution, for better or for worse, facilitates a more efficient allocation process that avoids highly skewed dating attempts where our cultural conventions prohibit skewed marriage matching anyways.[59] In other words, following these few steps we could watch Vaclav articulating a substantive problem in formal terms. His translation took us from dating to graphs, and finally to water pumps.

This example has begun to outline the scope of data science projects. Defining goals in quantitative terms in the attempt to address substantive problems creates new problems. Although the graph formalism reveals many possible matches, it introduces the problem of probable relations. Comparing this processes to more familiar technology settings reveals distinctive features. Software projects may aim to provide an interface for rendering text on a screen, or organizing news headlines, as we have seen among the technology nerds. Whereas software projects encounter primarily technical constraints for realizing a respective goal, we see in this data science project a tension between those technological capabilities and the constraints imposed by the data and its cultural meaning. Defining the ends of a data science project therefore requires the data nerds to articulate the problem such that they can apply their quantitative methods. So it came that we heard about water pumps when we learned about romantic dates.

In this instance the cultural understanding of marriage has offered significant guidance for navigating analytical abstraction. In other instances, that is not the case, as we can see here in a project on job recommendations:

[59] This solution also offers another instance of how algorithms interfere with everyday life. While I consider this question in detail below, here we already see that new opportunities and constraints are tightly coupled. I argue later that the nature of this coupling is most relevant for the public evaluation of data science activities.

> So, what does a career path look like? So I said, you know what, let me think about how does mine look like. And actually that's mine [slide showing a line with three point markers in varying distances on it]. And, each of those circles is kind of a major event, and in this case what I did was whenever I changed companies I put it as a major event there, in my career path. But this career path evolves in a different way for different people.

This data nerd encounters the social problem of seeking jobs. Why does it matter how career paths look alike? What is again hard enough practically, to find a suitable job, or marriage partner, turns once more into a new problem. Rephrasing the original question in abstract terms this time avoids pumps and buckets. Unlike the previous problem, Shankar here lacks the obvious aim to find the perfect partner to spend one's life with. Many of us want to move on professionally; it is socially unsanctioned, and even necessary today. Yet, we are not interested in switching jobs the day after we started a new one (although many probably wish they could). The challenge Shankar encounters for translating careers into quantitative formalism is not just moves, but patterns of moves.

This leads to the question of how soon are we interested in looking again:

> And in my experience I have seen three distinct behaviors. The first one, and to me that is the most interesting one, is, these are the people who have a very clear idea of a long-term goal. Long-term in like ten years out, they know what they want to do. And from then on, they kind of look at okay what are the options available today, and then which one of those will take me out to the ten-year goal. Rare breed, not a lot of these people.
> …
> The next one is, that focuses on the next step. I know where I am and I know where I want to be, and time is kind of the variable here. It may take me three years for getting there, it may take me two years, but that's the variability. And the next one is that, you know, time doesn't have that much flexibility. So, market defines what my next one is going to be. So a lot of the things we do focuses more on the second and third kind.

Short of the normative idea to be with one person for the rest of one's life, Shankar turns another basis for anticipating the moment when an employee seeks change. He considers his own observations. With the three categories Shankar arrives at, his team is able to send out job openings of positions and at times that are most likely to meet the interest of a receiver and least likely to annoy her on the basis of observations and comparisons. While the types of careers he bases this strategy on seem plausible, Shankar had nothing like the institutional background of contemporary marriage conventions to rely on for devising them. I have encountered many instances such as this one where data nerds followed their intuition for formalizing the substantive problem they tried to resolve with a data application.

In other instances, more so than in careers, which unfold over a lifetime, the modern data context offered sufficiently quick feedback that allows the data scientist to adjust simple heuristics. As a result of this, data nerds were able to adjust their narratives on the basis of initial results and thereby iteratively modify their formalism to the empirical problem they tried to express through it. Defining a goal requires

from data nerds to integrate a problem and technical capabilities through data. The concern with data has emerged independent of both technology and client problems.[60]

These instances have revealed the complexity required to overcome as data nerds define project goals in quantitative terms. Next they need ways to facilitate those goals.

6.1.2 Means

Let me continue with the marriage problem. Vaclav had introduced us to the system of pipes and water pumps. How to decide which pipes to let the water run through? Similar to how our culture defines for us, and the data scientist, the problem of how many matches to seek, namely one, together with our internal preferences it helps us to identify suitable matches.

Some are obvious:

> Food preferences matter … and you see people who eat healthy, who eat varied, junk food, gourmet or vegetarians. And here you can see on the diagonal [of a matrix shown on a slide] that, you know, there are three big numbers, which are green, that means that, you know, the healthy people generally get along quite well with healthy people, gourmet people as well, and for vegetarians this is the biggest positive lift versus like a normal, average match that we make, you know, there's not many vegetarians in the system, you know, when they find each other, they are pretty much unstoppable here. [Laughter.] On the other hand people who report that they like junk food, they don't, you know, have much success with anyone else [laughter], and especially bad it is, you know, if the two people who are junk food eaters are matched together, right [laughter].

How did we get from water pumps to food? Vaclav does not tell us specifically, but the cheerful audience reaction indicates that the combination of dating and eating is obvious to everyone. Relative to the previous setup, Vaclav moves from the analytical design to substantive terms. Proverbially, after all, the way to a man's heart goes through his stomach.

Although it surprises little that food preferences matter, not all combinations are equally obvious either. The data nerd begins once again very pragmatically with the data and quickly identifies informal boundaries of dating on the basis of quantitative analysis of food preferences. I was unable to tell from my observations whether the entertaining effect was intended or a byproduct of the project presentation. Either way, amusement in the audience followed the stylistically exaggerated note on vegetarians as much as the unedited reversed result for junk food eaters. This finding is genuinely surprising, both amid the other results presented here, and ideas of homophily often confirmed on other dimensions

[60] I should also note that to these social applications come purely technical applications of data science. We have seen a glimpse of early on where people came "out of the woodwork" as soon as data collection capabilities were available, and we will consider another one later on. Because we have much less culturally guided intuition about machines than marriage or careers, the iterative work becomes more central, but also more intuitive.

(McPherson, Smith-Lovin, and Cook 2001). This combination of expectation and surprise is consistent with many reports of these events that describe the combination of testing obvious ideas, and often confirming them, but also remaining alert to less expected discoveries. In other words, whereas we have previously seen ways of defining substantive problems in quantitative terms, here we see data scientists struggle with articulating quantitative results of "positive lift" in substantive terms.

These steps could seem smooth enough. At the same time, data nerds can rarely implement intuition so directly. Looks, another obvious basis for attraction, indicate some friction:

> Obviously the user profile pictures are a very rich source of information. You know, like anyone of us if you look at the picture of someone we can make, you know, some deductions of what kind of a person they are, you know. You can see what their, I don't know, even, you could estimate education, ah, income, obviously attractiveness, and other things.

The initial approach for photos is similar as with food. We can just think about ourselves looking at pictures, and the many impressions we get from them. The data context imposes significant limitations, however:

> But, you know, a lot of things that matter are a little bit fuzzy. It is pretty hard to write a program, a program in particular that will arrive at the same judgement as a human. People when we were interviewing them were saying whether that person is sort of trendy, hip, or cleanly dressed, or, you know, very hard to write a classifier for that. But what you can do is you can run just a face detector, and you can basically calculate the ratio of the size of the face to the whole picture, and that gives you sort of a proxy for what kind of shot it is. …

We can see any ideas, but no magic here. With the note that it is "very hard to write a classifier," Vaclav tells us that he could not encode the ideas people have from pictures into a statistical model. What this means becomes clearer when we recall Claudia's attempt from the previous chapter to build a model, or classifier, that tells whether a person surfing the web will be affected by an ad.

Beginning her description, Claudia reminded her audience that a click on a flashlight app is not equally meaningful as those in less disorienting settings. Other than that, however, a click on an ad is almost as meaningful to us as to a statistical model, it is one activity, or piece of information. A photo translates into thousands if not millions of pixels, each with a color code, of separate observations for a computer, that takes us a blink to make sense of.

Vaclav was nevertheless able to recover utility in this data by indexing qualitative information for quantitative analyses by effectively considering how much of a person appears on a picture. The result falls significantly short of capturing even remotely the kind of information Vaclav sees himself when looking at the pictures. The ease of "just" running a face detector partly follows from their available as algorithms that their developers make available online. It thus indicates the reliance of this project on

knowledge from outside the firm. The social stock of knowledge tells us that faces matter, formally indexed by the easy availability of the face detector. In these two ideas of food and faces we have seen two different strategies of formalizing common knowledge, a survey and available code and photos. Neither step nor their combinations requires specialization or bureaucratic task descriptions.

That Vaclav shares this process is interesting with respect to this organizational arrangement. He outspokenly admits that they were unable to capture the amount and also the kind of information they would have liked to see. How much we see on a picture says nothing of all the qualitative associations we make with them. Yet, even the simplistic approach, …

> … when you look then, you actually find that the picture that have more success in terms of communication initiations, which they receive, is like when you actually see the face and the whole person, …

Although it falls far short of what we might have expected of data science magic, this makes sense. And we do not yet need to abandon these expectations entirely, as accounts at the end of the chapter of some recent developments, or classifiers, will show. Even this limited implementation captures part of how users consider their dating options. Although it seems obvious now, Vaclav discovered this effect through a compromise between the qualitative intuition what we all look at in assessing potential partners and the analytical capabilities he had available.

For the question of how the project setting shapes data science work, this limitation makes an important point very clear. There is often much more in the data than data nerds are able to get out of it. One reason for these limitations is the lack of useful measurements in many contexts. We have seen the creative interpretation of romantic attraction as pipes and water flows, as well as the somewhat more conventional definition of career models. Whereas those cases defined the goals of the projects, here we see the many possible ways to reach them. They cannot receive equal attention, at least not initially when it is unclear how much they contribute toward reaching the previously defined goal. Vaclav compromises with available tools and still finds some leverage. Once again, there are other strategies.

This solution here relied on intuitive and practical simplification. Alternatives exploit computational power as we see in the following example from a project in fashion rentals. Instead of limiting the possible features to three types of photos, this data nerd involved other teams to assign qualitative categories to a large number of dresses in order to leverage these attributes for predicting popularity of certain dresses, but still encountered problems:

> … we try and run regression on it, we tried to see all combinations of attributes, we're trying to find, is there pockets of attributes that actually do really well and do really poorly, consistently over time, and we just tried to help [the purchasing department] every season, just do better, just do better, and I will conceive by the way, this is a very hard problem, I don't, there is a lot of noise here in the data, I think we are improving every year, data has played an instrumental role, but by no means is this problem, sort of, sort of, a cracked science problem here. It is very hard.

This account of a problem without "scientific" guidance leads the data scientist to resort to a data-driven solution.[61] Although this leads to some progress, the speaker excuses his approach as a necessarily provisional suffering from a lack of scientific answers. In other words, his "progress" toward improving the situation of the purchasing department does not count toward closing the scientific gap. At the same time, fashion is of course a scientifically studied problem, at least in sociology and anthropology. Leaving aside the question of whether these fields could have improved the solution, it is not surprising but telling that Vijay does not consider their understandings of fashion here. As the label would suggest, science is the place to turn to for help. As this was seemingly unavailable, Vijay turns to internal collaboration with other teams that coded features of dresses. If not some "fashion-science," external guidance still aided this solution through the quantitative classifiers that facilitate the data-driven compromise. In other words, commonly shared expertise is simultaneously more salient than the organizational knowledge, as scientific answers would have been the first choice, and less salient, as those tools from outside the organization that helped in the end were not recognized as such.

The two cases, of images of potential dates and of fashion tastes, capture the ambiguity data scientists resolve toward their goals of making matches and anticipating interest in dresses. Although just one project emphasizes the specialized knowledge of its organization, and that only as an aside, both pay attention to their community. Their respective tone, one confident and the other apologetic, indexes their recognition of these channels. The problems they encounter not only pertain to specific measurements but the basis for deriving them, and there is no clear definition for what that basis should be. This second account indicates that this is not the way science, which is specialized in resolving uncertainty, would have it. In the scope of a project, however, there is a larger concern with addressing the practical goal, which prevents data nerds from pursuing strategies that other regimes, such as science, might value. Here the aim is to get better, every season, or every marriage. Because of these aims as a thought

[61] The dating project runs field surveys independent of their online matching in order to assess its effectiveness. These results do not provide a benchmark whether a more scientific approach, whatever form that would take, might be more successful, but it does provide preliminary guidance for changes the project implements.

community it helps them to discuss these problems, but they do not need their mutual approval, as scientific peer review does. In other words, these projects are as scientific as their practical goals allow them to be, but not necessarily as scientific as an academic interpretation would require. Some care about this difference, others do not.

These attempts of seeking external and internal guidance, and the varying degrees of acknowledging it, reveal specific moments of community identification. While ways of coordinating solutions address our main concern with data science's contours as a thought community, they raise the question of their effectiveness. Quantitative analysis of course allows comparing different strategies, and there is direct evidence that specifically ignoring external ideas generates the better results. One data scientist recounts how analyses that were part of a competition revealed that case IDs predicted cancer, the outcome of interest. Observation arrangements in a spreadsheet have nothing to do with explaining cancer. There is nothing theoretically comprehensive or otherwise scientific with respect to understanding cancer in this finding. Rather the opposite holds. Only utter ignorance of substantive or conceptual meaning can lead one to consider the ID variable at all, and thereby find out about its predictive power. In other words, although projects often benefit from external or organizational knowledge, they are also self-contained.

This reveals the counter-intuitive challenge of deliberately ignoring what one knows. The following account, of a problem Riley, who we heard from before, encountered in travel destination matches of travelers and hospitality offers, describes such a process:

> […] there was one moment in [our] history, you know, I was getting really excited about this framework for a two-sided market place, and you know, how sophisticated we could get with it. And it kind of blew up in my face because it was too complicated and we had hundreds of people all over the world trying to, you know, interact with this super complex model that was, you know, try to score all of our listings across all these features, and trying to like optimize our listings, you know, and then they were being measured against this very convoluted metric that basically only I understood. And it was just too complicated.

The description style of features, optimization and metrics mirrors the main technical point; the idea was too complex. This project tried to anticipate and interpret the behavior and preferences of more users than one can plausibly imagine in sufficient detail. It is easy to see this difficulty with just two users seeking destinations in the same city for different purposes, and different accommodations with their unique features. What, for instance, if a vacant place is located in the business district, but available for several days when business travels may just come for a meeting and leisure travels, who want to stay

longer, are seeking the historic sights. How do they come together in the most optimal way? Although the "two-sided market place" idea appeared to provide a guiding framework, a much simpler strategy replaced this sophisticated solution. It did not provide optimal results initially, but it also did not collapse either. Moreover, iterative changes, and tests of them, thus largely relying on empirical observations of travelers, lead to improvements and a sufficiently robust application that this data team implemented into the product and shared with the community in a blog post.[62]

Riley thus endorses the idea of iterative work, as unscientific as that might be from Vijay's perspective. These different accounts have shown the deep yet also variable approach data nerds take toward addressing their goals, or problems. They sometimes seek theoretical guidance, of the kind familiar from science, and try to conceptualize their problems thoroughly. Where this is not helpful, they try different strategies. Maybe they come back, however. After all, at least the projects we have considered here do not seem to be solved for good anytime soon. We will continue seeking suitable romantic partners for a while, as well as jobs, fashionable dresses and travel destinations. As data scientists here share their problems and challenges, they indicate much openness toward approaches that might prove more promising.

Projects overview

Here we have mainly focused on a project on online dating introductions, with additional observations from recommending career steps, travel destinations, fashion choices and health diagnostics. Interested in determining in how far data science expertise defines the work in these projects, we have considered two important moments of defining project aims and devising ways for reaching them. In those accounts we have recovered indicators of where data nerds seek guidance from. This analysis has revealed a complex array of tactics with some consistent features but substantial variation as well with respect to their substantive problem.

In general terms, we have seen data nerds translating substantive problems, such as dating preferences and career moves, into formal terms suitable for quantitative analyses. Second, we have also observed how data nerds interpret the results of these quantitative analyses in the terms of the

[62] This reference to publishing the experience online reiterates the observation that all these public accounts have shared their experience of devising project strategies and failing in this process. It suggests that indirect exchange of experiences is not an artifact of the nature of the events I observe them in, but part of the process of identifying with a community.

substantive problems. For both steps they relied on guidance from various external sources as well as their own ingenuity. The external sources range from the generally shared social stock of knowledge, widely shared technical tools, science, to internal knowledge from their organization. Among them, we only saw one critical discussion of the lack of scientific guidance, whereas none of the other factors attracted critical attention. Indeed, we also saw how data nerds primarily focused on the substantive problem they aimed to solve, arraying the various sources for addressing them around that.

These observations raise the question of how data nerds are able to reconcile so much substantive richness with so many sources of guidance and ultimately articulate them in highly abstract and formalized terms. I turn to this question next, with respect to their specific skills.

6.2 Skills

The focus on projects showed better than that on technology or formal organization how data nerds balance the tension between substantive problems and technical capabilities as formal requirements of data solution enhance them. It also revealed their orientation toward external guidance together with direct feedback from data applications. Accounts did not yet reveal the specific basis on which data nerds integrate these different concerns. I therefore turn to their concrete skill sets here.

It is plausible to conceive of skills in terms of specific techniques, such as statistical methods or coding languages in this context. Here I argue that it is more useful for understanding data science to understand skills in terms of how data nerds implement and apply these specific methods and techniques. We could see from their own interpretation of hiring problems support for this conceptualization. Moreover, we have already addressed at least a few of the specific skills in the section on counting and estimating of the fourth chapter.

I focus on three skill sets—“hacking,” “learning” and “black boxing”—in order to consider the basis of data science work and expertise. Broadly, hacking refers to a practice of eclectically and pragmatically combing pieces of computer code and other formal and informal ideas for a given problem in a way that seems to follow no script at all. Learning refers to a family of analytical strategies and is most significantly defined by scientific discipline. Black boxing refers to the automation of data science, which implies that data science follows clearly discernible steps. Reactions to these plans are also considered as they reveal how data science skills might deviate from such scripts.

6.2.1 Hacking

Hacking is important to data nerds. At one event this could be seen as a speaker asked the audience to tweet interesting passages of a talk with hashtag "hacker," and boring comments with the hashtag "thoughtleader." Some would refer to Vaclev's water buckets as a "clever hack," just like to Claudia's two models for ad prediction from the previous chapter. Both take existing ideas and techniques and repurpose them for their respective project. This understanding is broadly consistent with the hacker community from the introduction, but pertains to more immediate problems. As we have followed Vaclev's implementation more closely, we could also see some more of the difficulty such repurposing entails. In order to understand these clever hacks systematically, we need to ask what the basis may be of such idiosyncratic solutions.

Composition

Institutionalized professions rely on educational degrees as strong signals of membership in order to overcome the ambiguity which outsiders experience the arcane professional expertise with. The expertise they stand for would provide a starting point to investigate the basis of clever hacks. Until recently, data science had no such degrees to rely on at all,[63] and previous accounts described how formal knowledge insufficiently indexes a candidate's suitability for a data job. Indeed, with the hacker Aaron Swartz we have started to consider a familiar yet obscure technology nerd role that seems entirely implausible in continuous work arrangements.

Data nerds do not just think about their expertise in terms of formal degrees, however, as we see in this response to a question regarding appropriate educational background:

> I'm biased because I'm working with a bunch of PhDs, but in speaking with a lot of hiring managers, the answer I've heard is yes [PhD degrees make a difference]. Ah, and the reason is the underlying skill set—in data science, you often don't know what questions to ask, and being creative around asking questions and knowing whether you've hit a right answer is something that is hard to learn in courses, which is at the Master's level most of, the only thing you've been doing is courses and sort of homework sets. And the real way you learn that is through experience. And there is nothing that teaches you that well than not graduating unless you find that insight in the data of your, you know, physics experiment.

This endorsement for advanced degrees speaks explicitly to the problem of ambiguity in much of data science work. Together with sorting out the appropriate technology, data scientists need to find the right question. In this context we can recall the fears of those very PhD level candidates, introduced in

[63] Now there are several programs, but there were no graduates from them at the time of the fieldwork. Chapter ten in part two analyzes this aspect in detail.

chapter five, in instances where potential employers failed to specify tasks. This dual meaning of applied value and distress reveals emotional depth associated with data science expertise, specifically because of a lack of clear or definite guidance we could also see above. More relevant for the question of how to rule the world, this description uses the analogy to graduate-level research as a shorthand to index data science skills. They require creativity, and graduate school provides an environment for practicing that.

This interpretation sheds new light on the previous comment on how fashion is not a scientifically resolved problem. In this view here, the benefit of scientific training is to sort out problems where it is not clear what constitutes a useful question and convincing answer. At the same time, we also need to recall that the highly institutionalized context in which graduate students develop such creativity still provides guidance and feedback, even if it does not feel that way, relative to a quantitative fashion startup.

In other words, whereas this interpretation of science has complicated the meaning of formal education degrees, the emphasis on creativity remains vague. Explaining what constitutes creativity that resolves ambiguity requires more specificity than just considering all those activities that are part of advanced graduate training. The following perspective, again on the hiring problem, provides some:

> But I wanted to give you some, maybe some statistics that might be useful for you when you are hiring. So, one interesting factor that we found was, look at language choice, look at what language people code their answers in, and let's see how well they do on that question. So, … you can see that people who did Python … do something about 12 percent better than average, and people are doing MATLAB—sorry for all the MATLAB folk out there—you are about seventeen percent worse. Ahm, and R is sort of somewhere in the middle.

How can formal indicators as languages say anything about the informal activity of hacking? Just like formal educational degrees for other groups, these abstract language names have substantive meaning to data nerds:

> And I think this kind of, this fact kind of illustrates some of the things that we have always thought about data science, which is a lot of it is not only driven around, ah, kind of having the best algorithm, ah, you know, arguably R and MATLAB have better libraries of algorithms, but a lot of data science is really just like data munging, and Python is actually a better language for that. You can see, sort of, see that in the data, and also if you have programmed in all three of these languages, I think you would agree that Python might sort of, because it is integrated in this general ecosystem of technology, is a little better for that.

How reflexive, a data nerd uses data to investigate the role of programming languages in a way that he at least assumes resonates with the audience. Michael apologizes to all the "MATLAB folk out there" because his results appear to have confirmed deficiencies in this language for the kind of analyses

data science entails.[64] Those less familiar with its features learn that this is because MATLAB capabilities complicate the "munging" aspect of data science. It is also not integrated into the ecosystem in the way Python supposedly is. Formal languages matter, but formal algorithms are not so important. Michael partially defeats the basis of his own apology, however, by describing his experience in all three languages. Thought communities define themselves not on the basis of a single feature, and the projects directly showed the relevance of diverse skills across different stages of implementing it.

Previous accounts did not specify the basis of community construction within the scope of a project. With a focus on skills we see here moments of defining its contours that rhetorically appear to exclude some, and practically surely do, on the basis of a lack of integration into the larger "ecosystem." Hence, for our conceptual understanding of community identification mechanisms, these activities reveal a contour of creatively reaching out to other frameworks in the right moments. Identifying what constitutes right moments requires experience. Ironically, this experience comes from academic training, the heartland of specialization. Python's promise lies in the rejection of specialization, which is, besides science, typical of both bureaucratic hierarchies and individual work in heterarchies.

This account has added some specificity to the previous remark on qualities of advanced graduate training. Michael's account resonates with the emphasis on creativity, but also shows connections between this creativity and formal programming languages.

That the speaker articulates features of arcane language presumably salient to his audience raises many questions with respect to a common understanding. For instance, what does an "ecosystem" of programming languages look like, how is it helpful with data problems, and how is it deployed? We have already been told, in chapter four, about the command line and that it facilitates such integration, and accounts advocating this utility indeed also consider a larger infrastructure:

> You might already be familiar with one or two programming languages, or environments, to process data with, you might be familiar with R, or Python with pandas, now I'm not saying that you should abandon that and do everything with the command line instead. What I'm trying to bring across here is that the command line can be viewed as one overarching environment in which you have your R, your Python and a whole bunch of other command line tools that can work together on the command line. …They can only work together if they can communicate, right, if they can exchange the data.[65]

[64] As we have seen in the projects before, the presenter interprets the formal results on the basis of his own substantive experience. As the problem he is interested in focuses on data science skills, which this event broadly addresses as well, Michael runs his own interpretations against those of the audience.

[65] In another talk entirely designed around making these arcane and informal practices explicit, some audience reaction was that this was the same as practices from several decades ago, which the speaker gratefully acknowledged. Although these practices are so common, they are rarely discussed and not formally defined.

Different languages have their own qualities, as we have just learned before. Each language comes with slightly different commands, conventions and capabilities. MATLAB provides great algorithms, but data science tasks also entail problems Python's flexibility is better for. Combining them allows leveraging their respective advantages in order to develop appropriate solutions for substantive problems. And here we learn how to do so.

We have retraced several layers of the initial comment praising the creativity that graduate training teaches and found this qualitative notion in formal terms. Whereas we first considered that for data nerds the creativity requires competence in engaging with an ecosystem of techniques, with this note we learn that even this engagement unfolds in abstract and formalized steps. Through them, data nerds join a thought community without explicitly acknowledging their membership or even recognizing the effect of commonly shared tools, as was partly evident in the project accounts above. The rejection of algorithms might initially undermine the fleeting interactions we expected from a hacker community, compared to the deeper interaction of other expert and technology groups we have considered (Coleman 2013, Kelty 2008). Technically, algorithms are the most formalized components as data science work, as they are even independent of computer code. Here, on the other hand, Michael rejects their relevance following a formalized test of data science knowledge. This suggests that data science skills have a relatively universal component that allows for the fleeting interaction of a test, without relying on the formalized knowledge of classical algorithms.

At the same time, such layering also enhances a project's complexity. It is easy to lose track of the steps that are part of implementing these systems. As these applications bridge languages, there is also no rules for the most appropriate design. In the following instance, Jeroen, who just described his command line approach, comments on some code examples in his demonstration:

> Even here "`cat`" [a command] is kind of useless. It is a bit of a, you know, an ongoing joke in the Linux community that there is this thing of the "useless use of '`cat`' award," and I, well, I am a good candidate for it.

This joke is clearly not meant for large crowds and the speaker does not seem to assume that it is well known even to this technical audience. This clarification therefore uncovers a divide within the community of those literate in coding languages. Whereas the speaker assumes familiarity with different programming languages as far as they pertain to data, such as R and Python, he does not assume familiarity with this joke from "the Linux community." Consistent with the slight contradiction in Michael's

apology to “MATLAB folk,” which followed from his own familiarity in several languages, here again data provides the more significant scope for relevant skills than the formally defined languages. Because the command line promises to integrate these different languages, the joke highlights room for ambiguity on a very granular level. Finally, the combination of the presumed ignorance of this community toward Linux jokes, and the function it serves here for the speaker, we also see the attempt of reflexive thinking among data science nerds amid a lack of formal as well as communal guidance.

Resilience

How could such a complex array of languages for different tasks and purposes and without clear boundaries hold up against the better organized challenges and established frameworks of academic disciplines or proprietary tools defining data nerd work as well? The different moments of community identification in the hacking context, ranging from creativity around asking question, rhetoric around ecosystems above, and cutting across communities with their own jokes, are all consistent with one another.

They are radically at odds with the idea of proprietary technology knowledge organized in bureaucracies of the kind promoted in Bill Gates’s definition of this work. The following experience directly illustrates a challenge to the bureaucratic approach as it invokes the skills we have learned about here:

> … for a long time we’ve been doing information retrieval on massive amounts of data, you know, petabytes and petabytes of data, and we built our own proprietary technology for doing that. And so about four years ago we wanted to improve our information retrieval algorithm with machine learning, so we went out and hired some data scientists. And, ahm, and we thought, well, they’ll just use our proprietary technology. Well, of course they wanted to use the tools of the trade and so, you know, Hadoop and MapReduce and R were introduced into our toolset at the time. And, and so we used that for, ahm, we use that to improve our record linking and other information, you know, learned to rank, and those sorts of things. I won’t talk a lot about those because I think they are well understood, just Hadoop and machine learning.

Data scientists indeed abandon, or overcome the bureaucratic framework. They see the opportunity inherent in integrating available tools and their ecosystems for the specific purpose of this project and ignore the ambiguity that comes along with the question of how to best combine these tools. Besides challenging forms of organizing around proprietary knowledge, this way of integrating tools also speaks against a definition of work that resembles the inverted heterarchy Torvalds constructed. Data nerds take what is available and useful for a problem they encounter, although some of the tools they

implement rely on Torvalds's division of labor themselves. It is not that bureaucratic and proprietary specifications would not work for technological reasons. They did, until data scientists introduced others.[66]

The speaker partly reveals where the data nerds took the momentum for changing the existing setup. This can be seen as the speaker assumes that the tools themselves are familiar among the audience, indicating the significance of the divide between these tools and the proprietary ones. After all, the speaker initially worked with proprietary software, and now, assumes that it is unnecessary to even describe the tools that are new for himself. This makes sense, he experienced that the community was there before when he made those hires. And it shows that although even vague hacking skills can be defined in specific technical terms, the collective effort of doing so as a community contributes to their utility as well.

These accounts begin to clarify, in some detail, skills for designing analytical pipelines and hence, integrating different tasks that come together in a project. Data nerds introduce their own tools in the otherwise proprietary context. Cynics may say that convincing business is easy enough because they lack a sense of arcane knowledge. How does hacking pertain to mathematical aspects of analyzing data? Statistics follows from mathematical reasoning that should be agnostic to the computational implementation. Even more so than organizations are academic disciplines, such as statistics, preserving continuity in knowledge. As far as data scientists rely on this knowledge, it should define their work.

This relationship is not so clear, in their eyes:

> … this thing [slide with formula of sampling distribution], is the really deep idea at the heart of that argument [about a quantitative test]. If the skeptic refuses to believe the assertion that this is the correct formula, your entire argument falls down in aches. In the general recipe for statistical inference, this thing is called the sampling distribution of the test statistic under the null hypothesis, and the reason that statistics 101 was so incredibly painful is that the idea of a sampling distribution is really hard to understand, even under the best conditions. And when it is presented as pure mathematical formalism, as a mathematical object like this, all slavered in degrees of freedom mumbo jumbo, ahm, there, it's just hopeless, right, there might be five people in this room, the people who raised their hands, who could sit down and just derive this from first principles. I am certainly not among those people, and I am a working data scientist.

It is not just coding and infrastructure, analytical principles concern data nerds as well. The tone of this account signals a critical view as well. When taking statistics seriously, with the mathematical

[66] Whereas the previous account rejected the importance of good algorithm libraries for choosing effective programming languages, here the algorithm is central again. The meaning is slightly different as well. The one above referred to a more basic set of data processing algorithms, whereas this one refers to a concrete application. If we understand algorithms in the diverse meaning data nerds assign to them, the two are still consistent as they both have concrete problems in mind, instead of general solutions. Associating both perspectives with the idea of algorithms thus recovers their utility as providing a sufficiently abstract yet consistent reference as to apply to diverse problems.

guidelines it offers, we observe here how these very boundaries can be seen to rather complicate than benefit finding useful solutions. Data scientists find ways around:

> Okay, that was statistics 101, what about this computational method that I promised? Well remember, the thing that we are trying to do is to figure out whether this 4.4 is a large or small difference, so we'll just mark that 4.4 on a plot. … We start with the original data, randomly shuffle, rearrange, tidy them up, compute the means, subtract the means, and now I get another one, +0.1, we'll put that on our plot. And now we can keep repeating that process over and over and over again. So here is three repetitions, here is four, here's five, here's six, and look no hands [holding up hands in the air], right, so what's happening here is you're watching a statistical process unfold, … [h]ere's 20 repetitions, here's thirty, here is fifty, here is fifty-thousand repetitions … a difference of 4.4 is fantastically rare, it happened just 14 times in fifty-thousand trials.

John shows us the transformation of a mathematical problem away from the formal legacy into its analytical components. He deliberately circumvents the mathematical steps formal knowledge would suggest for solving a given quantitative problem. Instead he walks the audience through the logical steps toward addressing the question of whether a quantitative result is surprising or not and implements them in some simple computer code on stage.

We therefore see again a break with an arguably obviously related thought community. This time it is not the informal Linux community or proprietary knowledge in formal organizations. It is the core of the academic institution. The challenge is once again constructive and practical, not on the basis of principles. Indeed, the speaker goes on to remind the audience that this perspective is not new and that earlier statistical thinking that had proposed it originally was primarily lacking the computational power.

This account therefore challenges presumably conventional ideas without aiming to call out a revolution and overthrow the statistics discipline's dominance over quantitative analysis. It contributes to an effort to challenge some ways of applying it. At other occasions, for example, younger members of the audience were troubled by presenters ignoring p-values and statistical significance in regression models. These audience remarks are not to suggest that academic statisticians defend these measures. They often do not. But the comments do reiterate a break that data science also participates in with some deep conventions in ways that puzzles those members of the audience with more traditional training.

We have thus far seen that data science engages in a series of these breaks with established knowledge. Statistics and other academic disciplines debate and resolve these disputes through academic journals, conferences and so on (Abbott 2001). As data nerds have no such institutional forms, nor clearly overlap with a single one, they resolve problems on the stages of these events and the blog posts we saw mentioned as well. We have seen repeatedly that they consider their applications not as

definite solutions, and not even as ideal solutions. Speakers have provided the detail of the contexts in which they defined them and the processes refining them, instead of alternative rules they may have relied on.

As intimidating as it may seem to stand up against the proprietary tools of an organization, or the academic orthodoxy, at least they offer something to stand up against. We have seen this kind of resistance to be typical of hackers such as Aaron Swartz. The kind of work C. Wright Mills had seen being undermined by bureaucracies, on the other hand, entails more continuity. While data science's salience signals potential for such continuity, we would expect that its constituting skills and expertise unfold independent of such direct challenges.

And indeed, we can find this tactic of reinterpreting orthodox approaches also for more mundane problems, for instance when returning to Claudia's online purchasing:

> Okay, is there anything else [than purchases, of which there are few] we can predict, instead. Maybe we can find some proxy for what we really want to predict. Forget buying, just see whether they are interested in the thing. Do they go to that product, just to the homepage of the brand? And it actually turns out that in this case if you build these models you get a much better positive rate. You get about an order to two orders of magnitude increase in the positive rate if you optimize site visits instead of buys. And for the most part, it's very nicely correlated with buys. So you can actually build models that predict purchase behavior very well, by learning a model on site visits. That's interesting, you're learning basically the wrong thing, to get a better model. That stuff's called transfer learning, and basically the explanation is bias-variance-tradeoff.

Here we once again consider Claudia who introduces us to an analytical problem in a way that resonates with the more vivid question of how we can leverage photos when it is too difficult to quantitatively encode appealing looks. This more mundane problem confronts the analyst with a situation in which she observes relatively few people who buy a product online upon viewing an ad for it. This complicates making confident decisions over who else might buy a product if shown an ad. The solution we learn about is that more of those who visit a website also buy a product than of those who do not visit in the first place. Thus, considering all those who visit a website provides a more robust basis for estimating who else might be interested in a given website, and subsequently purchase a product.

For this step, data nerds have to forget about the goal of encouraging consumption, and build a model for the "wrong" behavior of simply visiting a website. The problem of whether someone is interested in a product, or website, or not, translates into the logistic regression framework that rules the world, as we heard in the introduction to this chapter. Although it is powerful indeed, here we see some of the creativity it takes of the data scientist to make it rule the world. This ingenuity is a hack, similar to

connecting programming languages through the command line in order to address problems no one language is designed to address. Open-source software tools in proprietary organizations offer another possible example, just like computational methods for problems otherwise defined by academic statistics.

Especially these last two instances have shown that data science not only entails finding solutions to data problems by integrating different computer languages and packages in ways that are just not available otherwise. It also pertains to problems for which modern contexts and opportunities render existing solutions inadequate. The challenge to articulate appropriate questions thus extends from the empirical problem to both the technical and analytical implementation.

Here we have seen that hacking refers to an iterative process of finding clever solutions to problems that have not been defined before. In the applications to data, hacks rely on a shared stock of knowledge, of arcane statistical models as much as of mundane command line tools that orchestrate their otherwise separate challenges to existing solutions and approaches. In other words, although they break with communities that are close enough to share jokes as well as with the academic institutions and escape formal organizational oversight, they nonetheless integrate a stock of knowledge of formally encoded components.

Finding this relatively loose, contact-based community identification may have been facilitated by the focus on hacking. Hacking facilitates a second important idea, “learning,” which I consider next.

6.2.2 Learning

Problems in modern data analysis are not limited to the shape data comes in and necessary transformations for analyzing it. Their scale, complexity and timeliness challenges the analytical strategies and provides new opportunities for their application.

Just in the last example of online ad recommendations, for instance, much complex information is available on a user on the basis of prior website visited. Statistical methods are often designed to find associations between these kinds of information and an outcome of interest, such as online purchases in this case, or between food preferences and romantic matches in the earlier one. Whereas we understand food preferences, we cannot comprehend all the websites a user might have visited, hence we could not interpret their effect on ad relevance and are not interested in what it is. In times of computers and digital data that is also unnecessary.

Algorithms can take results from statistical analysis as instructions for further actions, such as showing a commercial ad, or introducing a romantic partner. Before this was possible, statistical analyses were useful to the degree to which one could interpret the associations they revealed, and act upon them. This is important still, and in some financial applications required by law (Poon 2015). With the new algorithmic capabilities, however, the data nerd alternatively just needs to provide an initial set of information on those who found an ad relevant and those who did not, and their respective characteristics. This information is then organized in two separate datasets, a training and a test set. Still potentially consistent with traditional methods, the statistical estimation then identifies possible associations between those characteristics and subsequent behavior.

Unlike in earlier times, data nerds do not necessarily compare the results against theoretical expectations. Instead, they apply this model to the second part of the dataset, the test set. Here they also know the outcome, but hide that information from the algorithm and only allow the model to make predictions on the basis of the first part of the dataset. They then compare the predicted and observed outcomes. With that model in place it suffices to just show characteristics, or features, as the algorithm has "learned" through the statistical procedure that outcomes likely follow from those characteristics.

This brief description falls short of even remotely capturing the academic debate around these problems.[67] For the purpose of considering learning as part of the basis for data science as a thought community it suffices to bear in mind that learning relies on a much more systematic procedure than hacking did. The details are not important themselves. As with our analytical interpretation of jargon more generally, here I focus on how data scientists articulate the relevance of these arcane problems at public events.

We can see that this reinterpretation of traditional methods is more complicated from a practical perspective. Even if data is organized in a well-designed stack, analytical strategies mostly require observations combined in a matrix for modeling their distribution. This mostly excludes both large and continuously streaming data as they are common today. Meanwhile there are solutions as well, however, and novel approaches spread to the community. These solutions easily look like hacks, but some of those

[67] This begins with the opposition between supervised (considered here) and unsupervised learning. James et al. (2013) provide a comprehensive introduction.

they learn as part of a different systematic framework, as we can see when we return to John's stochastic gradient descent approach:

> Q. Is there a good recommended way to calculate standard errors?
> A. [pause, relative to previous responses] Ah, for the estimates of these things?
> Q. Yeah, we're statisticians, right, we want standard errors …
> A. Well, the SGD [Stochastic Gradient Descent] is generally used in cases in which what we really want is predictions, not true estimates of parameters. Ahm, there probably is literature on this, but I am not familiar with it, ahm, in large part because basically every paper … I have ever read on this topic has been written by a machine learning person, and machine learning people just don't care about standard errors [chuckle in audience]. … If someone knows one, I'd love to hear about it, but I am not even sure they exist, let alone that they are [so far] usable. [pause] Sorry! Yeah … SGD very much comes from machine learning culture, very much come from people who are just trying to make predictions at sort of massive scales and sometimes sort of throwing some topics under the bus.

This dialogue between John and someone in the audience, reflects differences between classical statistics and its computational implementations or machine learning, which both specialize in quantitative analysis. Statistical methods produce "estimates" of relationships between different kinds of information and "standard errors" for specifying confidence over the estimated relationship. Estimates indicate the direction and strength of the relationship between explanatory information in a dataset and an outcome. Because this is an event nominally about statistical methods, so the person in the audience reasons, it should be obvious that the audience wants to know this kind of information. Yet the speaker is caught by surprise. He understands what the audience member means, to be sure, but would not have thought of it in the context of the material he presents.

This dialogue unfolds in the context of a novel technique, appropriate for leveraging large-scale Internet data. We can therefore directly see the role of how existing stocks of knowledge define data science skills; one leads to expect standard errors, whereas the other does not.[68] This moment of confusion, not over the ideas but their connotations and applications, shows the lack of a formal basis for data science knowledge, or a common knowledge basis. And this is just in the narrow aspect of data analysis.

The implications of this different kind of culture are larger than simply letting a topic fall under the bus may suggest. The following continuation of John's account responds to a question of how to process streaming data, as for instance online activity continuously generates it. Traditional statistical methods assumed discrete phases of data collection and data analysis. We could imagine only taking some part of

[68] We can also recall here an incident mentioned above, in which an audience member was surprised about the lack of p-values, and that the speaker emphasized standard errors instead. What that speaker considered more relevant, this speaker finds unimportant in this context. (The other speaker would probably agree in this context. The more important point is that it was the role of the speaker in both cases, and that this is not a question of universal truth, but context-dependent.)

the continuously generated online data and pretending it was from a discrete moment, but that would deprive it of much of its utility as it ignores change. These new methods, which get by without "estimates," seem promising:

> … you could just be processing [the model] where every iteration you're just processing one and it is not randomly selected, you'd just be pretending it was randomly selected. … You'd just use the newest thing, and that actually, in that context, which is the virtue of setting alpha to be a constant, in that context where you're accepting the newest data, this model, because you're having alpha be a constant, will be like an exponentially weighted average of the values you should have for the parameters, and so when say summer comes and people suddenly have different behavior than they had in winter, the model will shift to be a new type of model. And so that's how it will deal with the fact that the data is not actually random. The constant alpha will actually fix that problem, somewhat, I mean, it creates other problems but it fixes that one.

The change from winter to summer is obvious and much simpler methods could have accounted for it. The idea behind it is useful, however, particularly because this method addresses a concern with "predictions" and not estimates. As a result, the model will pick up changes that are not so obvious and an analyst might not have anticipated initially.[69] John puts much less emphasis on this practical utility, however, than on the technical considerations around implementing it. This speaks both to his interpretation of the audience and their interests, and to his interpretation of the method. With respect to the former, it signals a relatively horizontal relationship in that participants at the event are not interested in taking the method from him, but learning how they can implement it on their own. With respect to the method, it indicates that the framework of such a presentation, together with other available material and knowledge, suffices to implement even such an advanced and powerful method. This kind of interaction is in stark contrast to organizations that would not allow access to their proprietary information in the first place, as well as experimental setups in academic laboratories that often require extended visits to replicate them. The previous conversation over appropriate interpretation indexes this shared familiarity with the method itself.

Despite the noticeable traces of some hacking, we see more systematic aspects in the learning context. The machine learning 'culture' itself is of course not new. But novel technological capabilities have enhanced some of its effects and as a consequence the experiences of this community that applies them. If there is a more systematic framework, this raises the question of how well the data science community is able to articulate a distinct basis of its knowledge in a way that does not just apply what the machine learning culture develops?

[69] The SGD method is more sophisticated compared to the introductory sketch of machine learning ideas. The qualitative shift is similar, however, in that the interpretation of the model is not of interest.

> As I'm sure all of you guys are, all of you guys know, given just a lot of the popular media coverage around this, computer vision is advancing right now at the fastest pace it has ever been. And a lot of that is driven by the latest advances in deep learning, … just a few years ago, most of the cutting edge work was actually done, for in-scene object and concept identification, was simply just by hand tuning a bunch of features and put a classifier on top and seeing how far we could get.

This speaker describes a shift. The approach from before that shift almost exactly resembles the strategy we have seen above in the way in which the romantic matches took into account information of photos or "features." There the data nerd decided to just focus on the proportion of a face on a photo. This is possible, and perhaps even efficient, in the context of profile photos, where we can be relatively sure that it contains a face. In many other visual contexts, like in-scene object and concept identification that is not clear at all. With such a fundamental change, it immediately seems more plausible that modern data nerds define a distinctive stock of knowledge as theirs, even if its roots spread more broadly.

Following the shift, we have a more powerful method, however, and this account outlines the model thus far emerging:

> … in 2011 and 2012 we really saw the change in a really big way. A major, notable milestone was Krizhevsky's work on ILSVRC, which is the ImageNet Large Scale Visual Recognition Challenge, where in one year essentially they blew away the previous state of art with a deep convolutional neural network by about like half of the final error rate on the entire set. And so this was just a really big milestone, that's when most of the public interest in the field kind of happened. And just a few months ago, the main, general interest thing that was shown was when at Stanford Karpathy's group, they did automatic image captioning by simply hooking up a convolutional neural network to an RNN to get something as complicated as 'two young girls are playing with a Lego toy,' out of this particular image [on slide]. And this just all happened in the last few years. So essentially we went from very mediocre high-level classifiers that took features that didn't really generalize very well to some pretty with occasionally human-level feeling results.

This history began just when this research started, and the speaker makes it sound simple enough. It only took a combination of two methods to move from simply relying on how much of a photo a face covers, which Vaclav resorted to for a lack of better solutions, toward recognizing scenes in similar terms as those a person would describe them in. These methods are much more complicated than David lets them appear here. For our question regarding the basis of data science's distinct stock of knowledge it is more important to consider the way he describes this development, instead of focusing on the details David ignores.

With ILSVRC, David introduces the audience to an image recognition competition. Because this competition goes back longer historically, we can be relatively confident that the methods did not exist much earlier, which is also consistent with other observations from the field. Just a few years later, they appear in this presentation on a product that encounters much less well-defined problems than a ranking in a formally defined competition. Complementary to before when we have seen how data nerds

abandoned proprietary tools for open tools, here we see the application of widely accessible knowledge with academic origins to a specialized and commercial purpose. Data nerds preserve the more systematic background of learning problems even when they take them out of their original contexts.

While providing much less specificity, David also acknowledges the formal methods underlying his specific application, which has already widely spread anyway:

> It seemed like every day somebody was blasting through a machine learning record just by throwing deep learning at a particular problem and seeing what happened. It seemed like you could beat or at least match the quality of [inaudible] in your features with relatively little effort. And actually for a while people were just publishing new papers that were applying deep learning to new fields and getting them through and I think that the reviewers now are pretty tired of that and that they move on.

Thus, the cultural shift described above goes beyond merging the two fields of mathematically oriented statistics and computer oriented machine learning. Its consequences are also not limited to winning competitions and building new applications, they matter in academic fields as well. Taking the perspective of the presenter, who is relying on the utility of these methods, it is little surprising that he directs attention away from his own implication as a means of demonstrating the spread and utility of the same methods elsewhere. At the same time, it highlights a moment of community identification similar to the one before with “machine learning culture.” As expected in the learning context, these identification processes overcome the idiosyncratic and anonymous creativity in hacking.

In addition to these explicit community ties, in these settings it seems to be also relevant to specify the diverse applications:

> And so deep learning has been applied to speech, video games, protein folding, stock trading, and all manners of things, and even when you just scope to computer vision, ahm, there is actually still tons of different use cases, whether that’s analyzing medical imagery, multispectral imaging, satellite imagery or finding defects in manufacturing. All of this is just in the realm of applying that to computer vision.

Referring to all these other implementations and recounting the development of these algorithms does not give away or reveal all it takes to replicate his implementation, which David hopes to make a living off. It does reveal, however, that working with these new algorithms neither requires decades of organizational knowledge nor membership in exclusive academic circles even if they developed the methods. At the same time, David also describes their struggle with defining a problem they could focus on clearly because so much seemed possible. Once they settled on video indexing, they still encountered gaps between varying interpretations of scenes and images among different clients. In other words, this systematic spread could have appeared to suggest that its adoption requires less distinct expertise. We learn instead that the ambiguity shifts, and so does the central competence of the data science nerds.

This most recent context directly shows most clearly applications of autonomous work of the kind Mills saw lost almost a century ago. Establishing the credibility of this work, and defining relevant tasks in the first place, heavily relies on observing others also engaging with it and sharing one's own experience.

Data nerds do not only emphasize novel applications. They also evaluate their work in different terms than the sciences that contribute to the development of these methods tend to do. This ranges from analytical concerns with logical clarity and technical elegance from the previous section to economic worth:

> One of the major problems with the current, sort of, state of affairs in data science, I believe, is the disconnect between inference and actually making decisions. Right, we need to connect the inferences that we make with the decisions that we want to take. And ideally, we want to make that completely automated, okay. So deep reinforcement learning is an area of research which combines deep learning which you guys, I'm sure a lot of you have heard about, the promise of deep learning, with reinforcement learning, which is the idea of how to make decisions optimally over time, okay. Again, you should Google it. But to appeal to authority for why it is exciting, there is a little company called DeepMind that does exactly this, which Google bought for 400 million dollars, okay. They don't have any public product, they were bought before, before anything happened, for 400 million dollars. And the core of their technology, at least from all public accounts, is deep reinforcement learning.

This verbose reflection shifts the focus on the technological developments further. Instead of the detail with respect to underlying methods, techniques and applications above, here "authority" is supposed to signal its significance. That authority is Google and the huge price it paid for an arcane quantitative application that did not generate any monetary value. This reference adds to the previous strategies speakers have resorted to in order to describe the utility of arcane skills to their audiences. In this respect, the authority role of Google sharply differs from previous references to other actors. Whereas they invoked collaborative structures, this reference invokes a singular specialization and just the framing of authority invokes a formality that has been rare here otherwise.

Yet, just like other speakers we have considered before, Tristan also provides a specific description of the problem the DeepMind project addresses …

> … it played Atari. Now why is that exciting? So the idea of, the end goal of data science, on way of way to think about this is to build intelligent machines. People don't like to say it because it is too ambitious, and it is, you know, kind of, too crazy. …
> So intelligence, one way to think about intelligence is, intelligence as a measure to achieve goals in a wide range of environments. So Atari is a great example of a wide range of environments. And so what this company DeepMind did is it built a generic algorithm, okay, that, all it did was look on the screen, it didn't know anything but the pixels, it didn't know anything but the pixels and anything but the score. It didn't know where the players were, it didn't know this was a game, anything about this world, okay. And it was one algorithm that they could watch the game, and play the game, now playing the game I mean they just hit random keys, and they just observed what pixels changed on the screen, right, and over time this machine learned and actually outperformed the best human players.

The fascination expressed here with this new algorithm ignores much of the detail better captured by the video recognition implementation described before. We see that part of ruling the world may involve learning to play Atari. This feature, to be sure, is consistent with several observations so far in that data nerds formalized a widely salient piece of the social stock of knowledge. It also differs from many previous instance. With Google, another formal organization indexes the significance of the data solution. Google's reputation differs from even those technology firms defeating the hobbyists still a few decades ago. We can nevertheless begin to see a finer line of how nerds associate broadly salient expertise with more established actors. The processes of community formation provide the basis for autonomous work through improvising with available resources and solving relevant problems with that knowledge, as Swartz demonstrated and Mills described, have to prevail against intervening processes as well.

To be sure, this instance, although salient, is just one of many that apply data science with different prospects. If there are such advanced methods, why is logistic regression proverbially ruling the world? One reason is that these methods are recent whereas logistic regression has been around for a long time. Then it is also that many problems can be conceived of in a Boolean setup with a yes and no answer, including credit decisions, advertisement interest, and spam emails. Data scientists also report how clients appreciate the simplicity of logistic regression often in direct comparison to these more advanced technologies. Most importantly, we have seen consistently the significant effort, partly as a collaborative work either between speakers and their audience, or the field more broadly, of translating substantive problems in the formal context, even against bureaucratic and academic solutions.

6.2.3 Black boxing

If algorithms learn on their own, what work remains for data scientists? Let us consider these kinds of problems specifically. Professions rely on scripted activities, such as those that are part of medical exams of patients or legal procedures at court, at least enough such that they facilitate consistent appearances to clients and the public. Tensions over routinized work have long been known and although professional groups at times conceded some tasks, they could prevail over others (Abbott 1988). The digital context of data science adds reason for public concerns as routinized task can be automated almost entirely.

I have seen a number of projects at these events that promise to automate data science. They mostly focused on relatively simple tasks and not the kind we have seen accounts of when data scientists describe their work and skills. As some of the previous instances anticipated as well, the attempts to automate data analysis go further, as this idea of the "robotic data scientist" suggests:

> So, back when I was in grad school, I was working on this problem, I wanted to help scientists get to discovery faster. So I worked on a lot of algorithms, different techniques to try to leverage data in new and interesting ways. And, it turns out it is a really hard problem to help scientists come up with a new hypothesis on how something works, right. I think we spend most of our time as scientists, analysts, searching around for patterns, but it is really really difficult to find a meaningful, significant pattern. You know, something that is descriptive of what is going on, where we can kind of cut through the visualizations and the analytics and get to the heart, really the physics involved behind the system involved generating, you know, the data we collect.

This is Michael's introduction to a slide called "The Robotic Data Scientist." As we have seen before in quantitative applications in other social settings, trying to replicate scientific explanations Michael admits that "it is a really hard problem." Here it might be less surprising than in fashion. At the same time, it is useful to note that while Vijay described fashion as a difficult problem because science had not cracked it, here we are reminded of the difficulty for science to crack problems. From this perspective, the fashion project might have contributed more to a scientific understanding than Vijay described in his presentation, or was aware of. Conversely, this project that effectively tries to eradicate the role of data scientists describes its strategy in similar terms as we have already seen in other projects. Here the model is science, before it was Atari. Perhaps science is not taken as seriously here as it may sound, just like Google was not interested in good Atari performance.

As this introduction also suggests, the background of the robotic data science idea reaches beyond this emerging role:

> And the phrase that they sort of coined, or latched on to the research that we were doing, was "the robotic scientist." … [W]hat we really mean, what the robotic scientist really is, is this, right, it's a software that's communicating the patterns back to you and behind the sciences are servers crunching the numbers, telling you what's significant, what's important, how things work, right.

In practice this system does not replace the role of the analyst. Michael indeed goes on to describe his idea's success in terms of the community it creates of scientists, researchers and analysts who try out different models of the processes generating their data, and cites their enthusiastic responses.

In other words, this project is replicating the work of scientists on the basis of data science principles. This leads to applications more broadly. It implicitly juxtaposes data science with academic sciences. I have mentioned before that data science does not define itself in the context of other scientific disciplines, although we have seen references to ideas and tools developed in them. This idea here is

different. Michael does not invoke science for the methods it provides, but for the questions it asks. He thus proposes to automate the process of solving scientific problems, which entails the creativity Jake valued about this kind of work. This might be a useful marketing proposal to solicit interest from outside of science, and by his account also from scientists themselves.

It is also unrealistic. We have seen from the vast sociological research in the introduction that the academic institution is more complicated than the scientists who solve problems individually, as smart and complicated they may be. It organizes training, peer evaluation, publication, and funding, all in so many settings. More so than all those projects proposing to automate simple tasks in analyzing business processes, this ambitious project shows the limitation of this promise. At least in this instance of attempted black boxing, the focus of analytical work only shifts slightly, and definitely much less than the label of a robotic, that is, an automated data scientist suggests.

With the accounts from other data nerds we have seen by now we can also anticipate the challenges associated with this promise. We have learned repeatedly that data science involves tasks beginning before and ending much later than these pattern recognition exercises. Another more ambitious project saw little conflict in building a tool that replaces data science somewhat ironically. They prided themselves with employing some of the most successful data scientists, measured in the kind of public competitions already described. From this perspective we should not dismiss the idea of black boxing as overambitious threats aimed at data science. But the case data science makes for its relevance seems to be harmed only little by this kind of encroachment.

The black boxing danger can still be seen to prevail. If the role of the experts shifted from the modeling to the assessment of automated models, black boxing could still take away important data science work. But members of the community continue to emphasize the role of skill and talent independent of systems and technology:

> We both have backgrounds from Google so we've sort of had similar philosophy about bringing the Google tech stack [here]. I'd say that, so, some of the sort of key approach, key aspects of our approach are that we believe in investing in people rather than tools, and so we use open source tools, but also we try to build cross-functional teams of doers and implementers, meaning people who don't sort of just sit around hypothesizing, or telling other people what to do. We want people on our teams who can actually write code themselves, and so the data scientists have to know how to code, the data engineers have to know how to code, the designers code.

This account brings us back to Google, which we encountered above because of its financial endorsement of an algorithm that trains itself. Here, on the contrary, the idea to focus on teams with

technical skills even among less technical positions credits Google with this idea. Unlike in its previous mentioning, where the speaker invoked Google for authority, here it takes the role of a community member that defines the interpretation of work as heterogeneous skills. The last account based its reference on a publicly visible transaction; this one invokes personal experience inside Google. It can be both at the same time, but in the context of our focus on data science work and skills, it is important to note that although some see it focusing on highly specialized applications, others with direct experience working there bring with them a concern of drawing on diverse knowledge and tasks.

This focus on talent and skill anticipates a much more critical perspective:

> But then there is this sort of new angle, where we are actually, the data or the algorithms are actually getting build back into the product themselves, and the logic and the data is the raw material of the product, which then get integrated back into the real world where the users interacting with that product, and then there is this feedback loop that is going on. And so you have to be making sure that you're measuring the effect of the product you are creating and the impact you're having on those users and take that into account in your modeling. And so the remit of any team is to do both these aspects [reporting and product].

First and foremost, this account signals awareness of feedback loops in the algorithms, which the literature has considered in the performativity idea though with relevance for public concerns (Healy 2015). It also makes the point that the effect automation has on the system enhances the role of data science expertise. Regardless of whether this view resolves such issues, it provides another instance of how data science reflects over its own situation. We have seen presentations on new methods for large-scale and continuously streaming data, or on improved recognition of visual data. Here we see references to a new problem. In other words, in addition to continuously improvising solutions to familiar problems, some of which may undermine current data science tasks, new problems requiring different solutions appear in the same context as well.

Modern programming can black box data science work, at least some of it. But it still requires the skills to implement them and adjust to the respective problem. Moreover, data scientists also provide evidence to show their specific utility and contributions. Almost ironically, a project allegedly employing the largest data science team in New York City builds algorithms to outperform human data scientists in competitions at the same time. These competitions are highly scripted, of course, and thus such instances suggest that although capabilities are advancing, they require relatively formally defined context such as, for instance, academic disciplines or business analytics provide as well. Where this is not given,

data scientists seem to be on relatively safe ground to claim an autonomous role for their distinctive skill set.

Chapter overview

How much room is there for community identification when it comes to work? This chapter has directly focused on accounts of how data scientists design analytical practices. A series of instances outlined specific ways in which available technologies were put together such that they could solve a practical problem. The formal organization was not so important anymore for defining these practices. Instead, all strategies focused on the task of data implementation, that is, translating substantive problems into data structures such that existing tools can be arrayed around the problem that data addresses, or adjusted, if need be. In this respect we found clear evidence of community identification. That identification unfolds through a common focus on data, instead of preoccupation with technology, formal organizations or more self-aware discussions of one's identity. At the same time, the larger projects required data nerds to understand and articulate the quantitative results in their substantive context. A key tension has emerged here between following plans and improvising, raising new questions of whether data science can move forward in a shared and coherent direction without pursuing a clear aim. I address this question below.

The accounts have also clarified why data science is often seen to constitute a public problem. Many of their skills directly undermine familiar forms of control without violating legal restrictions. Data science becomes a problem of navigating uncertain problems and relations. This requires to understand data science as a distinct type of work, and thus also provides the basis for seeing it as a source for individual opportunities. Although both consequences induce opposite feelings about data science, they lead to similar subsequent questions. If the skills undermine familiar forms of control, what alternative principles integrate them? Answering this question pertains as much to aims of preventing public harm as to harnessing individual opportunities. Pursuing this direction benefits from understanding the abstract idea of skills and the principles underlying them in more practical terms, which I therefore turn to next.

The abstract community identification principles above become more tangible if we view them through the four types of technology nerds that we have considered as models for defining work throughout. At least for this moment, half way between the concrete technological basis of data nerds,

and the more abstract coordination of their work, which I analyze next, Gates has somewhat disappeared from the scene. Linus Torvalds has joined Aaron Swartz at the center of the accounts. We have encountered the most explicit moment in which Swartz meets Gates, and reinterprets his style of work, without questioning its purpose. Data scientists were brought into an organization in order to solve its problems. They ignored the specialized and proprietary tools and instead used the open-source software instead.

Many of these applications nevertheless serve a for-profit purpose, which we have associated with Gates's emphasis on bureaucratic specialization and proprietary skills. And indeed, several skills we have seen here clearly fit into bureaucratic task definitions. Their demonstrations, however, have also shown that their purpose may remain so unclear as to create a distinct position for the data scientists. As part of this, data nerds often relied on tools that come out of the inverted heterarchical communities resembling Linus Torvalds's Linux movement, although data nerds did not focus on making specialized contributions on their own. Such a position can be thought of as robust because here data nerds appear useful—instead of confrontational—to the organization with its specialized needs, and bring with them a diverse set of skills that are created independently of bureaucratic organizations.

We could see this even on the level of specific analytical designs. Although in ad clicks, for instance, the task is formally unambiguous, the data scientist reinterprets existing methods to implement this goal best. In other words, here we begin to see moments in which data nerds reconcile the two opposing positions of Gates and Swartz, partly by relying on the open but otherwise specialized types of knowledge associated with Linus Torvalds's projects. It is not yet clear how they do that.

If data nerds deviate in their practice from these familiar roles and the organizational principles they represent, what is their basis instead? In order to understand the contours of the thought community data nerds make up, we need to understand how the distinct moments of improvising hang together beyond the specific problem in the face of which data nerds break with established ways.

Contours: Improvising

How do data nerds so confidently sideline established knowledge without invoking or proposing a comprehensive alternative? Unlike the settings in previous chapters, when it comes to work and skills data nerds articulate their approaches explicitly and neither leave it at vague comparisons nor remain

implicit in their relations to others. This can be seen in coding demos that circumvent mathematical solutions with “no hands” computational solutions and Q&A responses admitting that some concerns may fall under the bus, as well as jokes about the particularities of their own approaches. They mock academic dogmatism but offer little more than creativity and experience as alternatives. At the same time, data nerds neither take a position explicitly against ideas contributed by “cultures” they distance themselves from, seen as they use them whenever possible. They only break with the idea of immersing themselves in those cultures or the proprietary and specialized infrastructure of formal organizations. Specifically because these accounts have given no indication of a systematic background of these breaks, we have to consider the concrete moments when they happen in order to discern their basis.

We have already seen that data science puzzles the clients it serves, and alienates organizational functions that claim data problems as theirs to solve. Considering accounts across these public events by speakers from different organizations, cities and over several years, we nevertheless see much similarity in the kinds of practices they describe. Because academic disciplines coordinate across organizational boundaries more than within them (Abbott 2001), and contribute methods data nerds use, they might offer an explanation for those patterns. After all, most data scientists have undergone advanced academic training. Regardless of whether collaboration in small research groups or institutionalized processes, the two camps we considered in the introduction, drive scientific progress and thus legitimacy, we saw data nerds either adopting scientific methods and repurposing them, or outright addressing scientific problems from an outside perspective. In short, we have found a radically different approach, one that is based on improvised problem solving rather than systematic approaches common in disciplines (although they exist as well).

These processes are less surprising once we take into account the practical problems data scientists have described together with the technical decisions. Data scientists face the challenge of connecting the vast number of opportunities stemming from the availability of data and methodological developments that are possibly suitable for the substantive problems they address. Continuous recording of streams of unanticipated information leaves little time for identifying appropriate theoretical backgrounds and the groups defining them. Whereas in the past analysts required guidance, often based

on theories, in order to begin collecting data in the first place, data scientists more importantly need ways for deciding how to look at the data that is around anyway.[70]

Instead of seeking to ally themselves with the established academic fields, data scientists focus on the utility of their ideas in a given context. Somewhat resembling academic practices, although in informal ways, they cite the spread of similar tools as evidence for their status. We can recall the speaker who very emphatically pointed out that modeling taste is really not a "cracked" scientific problem, indicating his appreciation of scientific guidance amid a need to move on without it. Even more telling was Jake's note on academic training, as he emphasized the challenge it entails to find creative solutions, with no mentioning of the scientific method it is commonly known for. Data nerds legitimate some decisions on the basis of existing academic groups but accept that some aspects others care about may "fall under the bus." As indicated above, the criticism does not try to offer a comprehensive alternative approach to the problems the sciences are concerned with. What principles explain their work instead?

This kind of practice is rare, but not unheard of. Even scientists occasionally depart from the direction of central ideas (Foster, Rzhetsky, and Evans 2015, Evans 2010). We are able to see this process unfold more systematically, and reveal its principles more clearly, in different historical and substantive contexts. For instance, we can find evidence of systematic improvising in unconventional combinations of available resources throughout major economic transitions. This begins with the first wave of global trade (Erikson and Bearman 2006) and includes the collapse of socialism (Stark 2009) as well as the breakdown of large integrated corporations in the American industry (Whitford 2005). Respectively, captains violated directives and used the ships entrusted to them for trades on their own account, workers in socialist factories used machines in their own time for work outside of the planned economy, and manufacturing suppliers built relationships beyond contractual obligations in order to provide suitable parts. All three instances reflect forms of systematic yet unguided departure from institutional principles of coordination. Returning to the modern data context, these explanations remain relatively tied to the economic context in the sense that it defines or at least confounds the motivations that lead to these kinds of activities. Notwithstanding overlaps, coordination principles of economic activity by definition differ from coordination principles underlying expert practice. The former consists of

[70] The settings here may shape this presentation as well. They surely leave out details of how they set up their problems, but there is no reason for them to emphasize heuristics here and work with an elaborate theory in fact.

bureaucratic hierarchies with specialized knowledge, even when relationships and exchanges unfold between them, and the latter of arcane knowledge the community shares. It is therefore helpful to consider one more instance, which further back in time and substantively distant, in order to consider how such improvisation unfolds in data science.

Radically changing the setting, we can recall that deviance also defined the transition toward Christianity in the Roman Empire of late antiquity (Brown 1992). Brown recovers from ancient sources how local religious leaders secured support by abandoning the strict reliance on the elite culture with its heavy mark of Greek classics and instead adopted relatively more mundane Christian scriptures that the public could more easily relate to. Amongst themselves, to be sure, the early Christian leaders kept the more arcane ideas alive as well. Against this background, we can consider how the arcane quantitative knowledge together with the practical problems that data nerds put at the center of their work provide a dual backbone for data science, held together through the moments in which they improvise to connect them both by “hooking up” different analytical techniques or letting programming language “communicate” with each other through “exchanging the data.”

Consistent with the specific economic and the early religious cases, data science relies on no institutionalized, centralized or direct form of coordination. Instead, its principles emerge from the technical descriptions of how to most effectively deploy data, seen as data scientists share their skills and present their experiences. The presentations we have seen here implicitly revealed improvising and transposing knowledge as a skill itself. The computer scripts underlying these exercises formally record the improvisation. While such traces facilitate circumventing direct and personal interactions, they lack much meaning from collective experience. Even insider jokes told there remind of possible flaws in these scripts, and anecdotes remind of the considerable effort to design data problems such that the power of the most recent and advanced tools can be productively harnessed. Thus, while direct relationships are not so important for identification with this thought community, the slim contact through group experiences is, because without it the scripts are easily interpreted as authoritative, which they are not meant to be.

This process reveals a contour of coordination arrangements neither mapping onto the existing disciplines, nor emerging from network forms of organizing between data scientists. Data nerds improvise their activities on the basis of technical scripts that are simultaneously anonymous and abstract. While

this account so far suggests a set of practices that indicate a distinct form of data science, for it to be sustainable we need to be able to recognize a community. I address this question next.

7 Community

The preceding chapter has shown that in spite of exposure to organizational and institutional effects, data scientists articulate an arguably distinctive skill set as theirs. For these skills to continuously sustain an expert group or thought community, they require some principles or a form of organization or coordination. Expert groups by definition lack significant hierarchical control. The Bar association, for instance, provides very little hierarchal coordination for the legal profession. It still regulates access to legal work. Many instances of expertise have no such formally institutionalized barriers in place. The same holds for data science, if only because of its nascent status. In order to identify means of coordination and control, we therefore need to consider the experience of group membership and the definition of its terms directly.

How this analytical step addresses the public concerns with data science may be less clear. After all, short of an own formal organization, overt membership experiences most often also correlate with some other organizational context, such as the supermarket chain in the introduction, which are well-known and can be addressed as legal entities with complaints. An additional level of abstraction is nevertheless appropriate and necessary to the degree to which these organizations do not fully define and hence control data science expertise, as we have seen evidence of repeatedly in the chapters so far. Conversely, moreover, this step directly speaks to the concerns that follow from individual opportunities in the form of data science work prospects. Whether bureaucratically defined or otherwise, those seeking to pursue this trajectory not only need to understand the attributes, the skills, but also the context defining their relevance.

Communities are often thought of as local neighborhoods and other personal bonds that organize social life and activities (e.g., Sampson 2011). Professional appearance rarely invokes community ideas.[71] Their relevance in this data science context is therefore not obvious initially. This notion seems particularly far-fetched as we have seen evidence for types of practices that were previously contradictory and often in conflict with one another. Here we can just recall the initial tension between Bill Gates's bureaucratic and proprietary skills, the opposition through Linus Torvalds's Linux movement, where data

[71] The literature has seen discussions of professions as communities in early work on professions (e.g., Goode 1957). It has soon rejected that notion (Larson 1977, Heinz and Laumann 1982).

nerds replaced proprietary with open tools, and the constant challenges from hackers such as Aaron Swartz. In other words, the activities and definitions of work we have seen evidence of so far, if anything, represent different communities.

We were nevertheless able to see that the skills can as well come together from these different groups of technology nerds. We learned in the previous chapter how speakers emphasized anonymous communication around programming languages, as well as building on work of specific groups and the methods they have developed. Here we consider how these nerds, who apply their skills in this way for different problems, define the terms of a shared community beyond casual references to others. We thus move to a more abstract level, where we ask how data nerds acquire experiences associated with different task arrangements and apply them to new purposes.

From the perspective of thought communities, we gathered evidence of what might constitute its contours in the data context all along. Most significantly, we have seen that these contours fall on no formal indicators alone, such as organizational function or programming language. Instead, they emerge from the integration of several informal activities, such as through illustrative persuasion, or technically and explicitly through the command line. Data nerds consistently associated the informal variability with formal markers. So far we inferred these contours from descriptions and examples of data science work. Here we consider definitions of membership in the data science community in those moments in which it becomes salient to the nerds themselves.

How can we observe community identification if there are neither definite terms nor a formal boundary? The accounts in chapter five on conflict with other organizational functions have made the technological side of this salient for data science. Here I focus on self-conscious moments of community identification. We have seen a spark of this, for instance, in the previous chapter as a speaker described coding discipline through jokes of the “Linux community.” Although the idea of Linus Torvalds’s project, and open-source more broadly, rejects formal boundaries to a large degree in the first place, these very principles define a community as well and hence, may shape the accounts of data work. Contrary, if the community overlaps strongly with bureaucratic boundaries in spite of its many informal practices, we should expect Gates’s views in the following accounts. In this case they would at least try to specify tasks for data nerds within formal organizations. We saw external perspectives of this in the proposals for

specialized divisions of technical tasks. On that note, it will be most interesting to consider the degree to which Aaron Swartz's types of arcane and autonomous projects and Mills's emphasis on salient work and abstract knowledge emerge here. Swartz's types of projects unfold erratically and therefore constitute, if anything, an ideological community to begin with, and in Mills's view experts have been unable to define meaningful groups for a long time. From this perspective, neither could have concrete ideas as to what might constitute membership. At the same time, we have seen the very skills of taking knowledge from other contexts and applying it to a given problem, associated with Swartz's sporadic instances, fold together systematically across organizations leveraging data for their purposes.

The previously observed skill applications appeared in combinations that we were not yet able to describe with reference to a mechanism integrating them. The moments of professional identification we move to here are designed to reveal the terms of its membership, even of a subtle and anonymous community. We have seen that in this incipient moment the skills fall under no formal label, such as an educational degree. Where education mattered, it was for the creative experience. We have also seen that many technical skills are required nonetheless. To the degree to which the community shares these observations, accounts of joining data science must invoke the concrete processes of acquiring those skills. Whatever the precise combination of skills this community defines as its own, they must be articulated in moments of acquiring and defending it membership. Pertaining to such moments, the following accounts reveal variable paths into data science work, and thus exposure to different skills and different arguments for defining it vis-à-vis other experts and their stocks of knowledge.

7.1 Passages …

Two important moments mark our working lives. The first is the transition from school into work. The second is moving from one kind of job to another. The educational context may shape perspectives in similar ways across different areas of expertise even if they are associated with different applied problems. I therefore consider them separately and start with experiences of the first, and then turn to the second.

7.1.1 … from school

I first ask what institutional channels leave their mark on paths into data science. Data science itself has not had the time to define and implement training and a career itself and thus common paths into data science activities could indicate sources of data science expertise and practice outside of the community itself.

Here is one alternative experience:

> Well, it's funny how I ended up at [this data science job], actually. I initially intended to be a journalist, and sort of felt like the writing was on the wall, and that there was this insidious thing demolishing the wall, the separation between church and state, between advertising and editorial, and ended up studying mathematics and statistics, and ended up at a company doing analysis of content marketing, so sort of what I was afraid of at the very beginning. But, I think my sort of instincts led me in the right direction, I just sort of mixed up who I was working for.

Instinct, accident and reactions to larger transformations, the key experiences in this transition, remind little of institutionally scripted career paths, or strategic career changes. Such a lack of career paths was mentioned regularly at these events, and is consistent with an understanding of data science as an emerging group. There is a deliberate choice nonetheless. This account also reflects a conflict that begins with resistance to externally defined tasks, here through the market and the bureaucracies operating in it. In the end Zanab describes somewhat of a compromise as she entered proprietary work, albeit in a different capacity than the one she found troubling at first. As Mills pointed out, it is difficult to identify alternative paths without role models into autonomous work.

The time between the initial resistance and later compromise is important here. The initial attempt with educational decisions had a clear aim of turning away from a certain setting. Today it seems "funny" because Zanab ignored that initial aim for moving on. Yet, we have to understand each step in its historical moment. There it is not so surprising because no plan was in place for the next step. This strategy lost momentum as soon as Zanab entered the job market. Her description of the situation thus reveals some clear terms, but no strategy.

How could a lack of planning provide the foundation of a coherent community of experts? If others experience these transitions in similar ways, they would provide strong evidence of simple relabeling of a much older expert group whose established channels a new label disguises. And there is more support for this interpretation in Jake's experience:

> … I worked on the Large Hadron Collider at the University of Toronto, looking for the Higgs Boson, although briefly, before I decided, that, you know what, this stuff I was reading about in wired magazine, and on hacker news and on tech crunch was a little bit more exciting and interesting to me, and the impact I was seeing there

> was really spectacular. So I kind of took the exit ramp, had all those great skills coming out of physics, that I thought I'd leverage. … And then I hit a bit of a brick wall. Ah, I sort of didn't quite know what to do next. All my friends were from academia, I was sort of in this science bubble, all I had ever thought about was being a physicist, and it wasn't really clear what the next steps were.

This time we see an aim toward a data science role. As Zanab, Jake had no plan, and even his friends could not help. The importance of connections is well-known in the problem of finding jobs. We also learn about academic physics as a different setting that in Jake's view provides the appropriate skills but denies ways of applying them productively. In other words, even if data science to some extent relabels existing skills, that relabeling is not a simple process in practice.

We see some of the difficulties if we follow Jake's experience further:

> Ah, eventually packed the bag after a little bit too long, moved to Silicon Valley, got into Y Combinator, which is a great startup accelerator program, and that kind of really opened the doors for me. [It] took two to three years of stumbling around to figure it out. And, I'm sort of not the only one, there's a lot of academics trying to make that transition.

Unlike Zanab's account, here Jake also emphasizes extended periods of struggle in the search for next steps. Again, though, learning about a specific opening did not play a central role in the end. More central to Jake's experience was the abstract problem of a lack of knowledge of what to do next. The steps lined up along experiences, not specific information.

Jake's situation becomes clearer when we consider his new role with respect to the old contacts:

> A lot of my friends start calling me up, asking how they can make this transition from physics, math, other fields. … I think, a lot of people like me are just really attracted to the kind of impact they can make using the skills they have in industry, in high tech fields. And unlike that flat faculty job graph, as you all know the sort of, the demand for data scientists and analytics professionals keeps going up and this trend has continued.

Now it is about information flow, after all. Jake emphasizes again the prospect of impact, which he associated with public coverage of this work before. He had trouble seeing clearly to get into the position of making such impact although he had the appropriate skills. Jake is able to provide the information he was looking for before. What began as somewhat of a discovery expedition quickly turns into relatively concise informal flow.

The steps by which Jake built up his understanding of data tasks collapse into just a single piece of information. In both accounts we have considered here similar information likely circulated as well, such as in the references to popular magazines through which Jake learned about the impact of modern data technology. Both Zanab and Jake however emphasized the experiences they underwent, transitioning from a statistics education into a data science role, from advanced physics training to a Silicon Valley

incubator and then into data science. Contrary to this experience, others who learn about the result are able to seek direct information.

Zanab and Jake share quantitative backgrounds and the experience of navigating unknown territory. Their accounts also differ, for instance with respect to the different kinds of training and different concrete steps into data sciences. These accounts leave us unable to distill a formal path into data science. So far we learn about the consequential nature of the work, as well as resistance to bureaucratically and academically scripted tasks as a possible motivation for exploring the uncertain transition. How Zanab and Jake experienced their respective situations lines up more systematically than the concrete steps they took.

Complicating this synthesis yet again, the same kind of transition overcomes uncertainty in different ways. We see this here in Michael's experience:

> … I have to tell you a little bit about myself and the company, but I promise there is a story. And I am a recovering academic. … and sort of through my rehab I worked as a quant, at D. E. Shaw and here at Bloomberg, and then I became a data scientist, at Andreessen [and Horowitz], and most recently at foursquare.

The perceived recovery process once again signals the difficulty associated with this transition. At the same time, the uncertainty seems much less salient in this description. The stops align more smoothly than accounts of brick walls suggested in Jake's experience. It might be accidental, but this specific path unfolds almost linearly with respect to the kind of data Michael worked with. D. E. Shaw, as an investment management fund, signals narrower data than Bloomberg, which caters a much broader client base. This pattern continues with Andreessen and foursquare, respectively a young Venture Capital firm and a startup, both in Silicon Valley. The steps lead from more to less institutionalized contexts. They define a relatively long chain of connections that signals intellectual development more than direct information flows as guidance into data science. The former has been observed in academic collaborations and the latter in other job markets. They thus pertain two contexts Michael went through along the way. His account describes neither process and instead focuses on the formally recognizable steps. It therefore seems that while the transition is not obvious in any of these accounts, the experience does not have to appear funny or resemble brick walls either.

Even where formal steps align quite smoothly, what is "a recovering academic," which Michael introduces himself as? We know of recovering alcoholics or other addicts, but academics seem rare in

this context. Beginning an academic career, as Michael did, leads into an experience of a more or less total institution without much interaction with the outside community (Goffman 1990). Leaving such a setting leads to a loss of meaning and purpose defined by that institution, which one has to recover from once leaving it. For Michael's trajectory we can infer adjustment from the academic context in which one tries to address arcane questions mostly colleagues find important, toward a context where broadly relevant questions matter, judged by clients, customers and consumers. Or, as the previous speaker put it, the transition from answering questions toward "making an impact."

It is unclear whether that new purpose is visible initially. Rehabilitation comes with the image of a deep and transformative experience, such as hitting a brick wall, as Jake felt he did. It is also a relatively lonely endeavor that while benefiting from the support of others, someone has to undergo alone. Of course only very few who undergo this experience move into data science, although we heard from Jake that many of his friends are interested. Those who end up there share this background and experience. Regardless of its demographic significance, it captures an important aspect in that it may involve emotional adjustment as the skills applicable in one context were defined in another. Presenting these transitions condensed in specific information on the different steps could suggest ease for subsequent generations. At the same time, expertise scholarship shows that picking them up in their original context may be necessary because many of their nuances will be lost otherwise.

Michael's experience therefore refines our understanding of the passage from school into data science. We are still unable to formulate definite steps. We understand better now that our intermediary synthesis of obscure transitions may seem funny or like a brick wall because it is ill-defined but not because the relevant skills and training are in the way. Then we need to ask how they align, if not through consistent formal steps. Here I just begin to address this question by considering steps that resemble those above without invoking as profound experiences as recovery implies:

> Hopefully this is the boringest slide that I show you today, … so, this is me, … my background, probably unsurprisingly, scientific software engineering, math, stats—shocker. My graduate work was at the School of Computer Science at Carnegie Mellon, ahm, doing analysis, inference, simulation, of complex human systems. So this would be social networks, terror networks, ah, industrial systems, things like that. Ahm, and about three years ago I became [the] first data scientist [here] and have been building the team ever since, and it has actually gotten quite large, …

What seems like funny accidents, brick walls, and even rehab to others makes for the most boring moment in this speaker's experience. This interpretation is very surprising in the context of the previous

accounts. In substantive terms, however, viewing a move from an academic background in computational modeling of “messy social problems” toward a data science role in an education project as a “boring” transition, while that from the Large Hadron Collider appears eventful, is more straightforward indeed. It allows us to distill systematic patterns from these individually partly disorganized experiences.

The experience of advanced graduate training therefore offers no direct channel into data science work in spite of its systematic salience in these accounts. Learning skills in their original context is also not necessarily tied to emotional hardship. Speakers are not shy to share either experience here in front of their audiences even if they seem to signal amateurism. Besides the overt skill of having training in defining and answering questions, as data nerds share such intimate experiences in the public settings we consider here, they make the community identification with data science a collective project of navigating uncertain paths.

7.1.2 … from work

Aside from these varying experiences of transitions, and paths, all four have arrived in data science, and all except the first mention advanced academic training. Although a sense of adjustment dominated, everything that comes after school can easily seem different. I therefore also turn to professional transitions. We see one such case here:

> I was formerly CTO and co-founder of [a data startup], and now I am a software engineer at Cloudera, working in the Hadoop ecosystem. I was an early contributor to the pandas open source library in Python. If you’ve done data analysis in Python, you’ve probably used pandas. And ah, in a former life I was a financial quant on Wall Street and also in Greenwich, Connecticut. And my passion is creating better tools to make people working with data more productive.

The struggle here seems marginal, untroubled by the passionate side note on helping the data community. The reference to contributions to a software package describes an entirely informal process in the programming community most prominently seen in Linus Torvalds’s Linux project. This account signals less overhead integration than the applications that become part of the Linux framework and have to gain approval. We have seen references to pandas before in the context of command line data processing. Recovering such a formal trace, this account makes more tangible than others how these nerds participate in the thought community around data problems by finding solutions instead of just drawing on those resources.

Although this experience signals much confidence, uncertainty is not over once school is finished. The trial and error tactic, emotional or not, prevails even in specialized and partly proprietary settings:

> [I]t started at a founder-dating-type thing. … We started out by building an ad server for what's called real-time bidding, and it was about two and a half years ago, right when real-time bidding was starting to really to come into its own and start to take on a real good sizable share of the market, and when we found out that there were a lot of people ahead of us with a lot smarter teams, we decided we should probably start looking elsewhere, …
>
> … and following a series of attempts, they …
>
> … built this great algorithm for content recommendation and for ad recommendation, we based it on a whole bunch of factors like behavior and social media, and what we found was we kept getting asked the same question: how're you guys doing that social thing? We kept saying: You guys don't know? And, nobody knew. We said, well, let's show you. And, ah, that's how we did it. We just started showing people things they didn't have access to. And what we found was that, there is a lot of social media data, that requires not only ingestion and cleaning, but also analysis and display, and that sort of led us into this space where we were looking into small amounts of advertising and content into larger amounts of advertising and content and even much larger amounts of social data behind that.

A founder-dating-type thing constitutes the least specialized context, in terms of professional positions and consolidated expertise, we have encountered so far. It is not much different compared to the trial and error tactics for finding opportunities in data analysis following school. We once again see an aim for working with data without a plan what to do.

In this transition, and the previous account, we learn about the shaping of an application instead of moving into a position. Contrary to the community orientation of the pandas project, this one serves a market. The speaker also acknowledges the community, however, as he recounts how other projects offered more sophisticated implementation of an idea similar to theirs. This helped them identify another problem they could address. In some ways this adjustment just reflects the competition and specialization in a market. In this specific context it also resonates with previous accounts of the lack of clear paths into this area from graduate school. In other words, for the purpose of understanding the data science thought community, this reference to other project testifies as much to market mechanisms as to community identification. After all, the other teams did not offer lower prices. They had smarter ideas. From this perspective it also underscores the point that the experience of advanced graduate work is significant independent of specific knowledge it may teach, by showing that graduate school is not the only place for these experiences to be had.

Particularly studies of patents have shed light onto the innovative potential in the market setting (Powell and Snellman 2004 for an overview). The previous account is still surprising in as far as we have up until now primarily seen science as the origin for relevant expertise and ideas. Science is of course

often the source of innovation in the economy (Moser 2016). We are here more interested in the contexts data nerds identify with, rather than the comprehensiveness of their observations. It therefore raises the question of the kind of settings data nerds are able to define their knowledge in. The process of search in uncertain territory, now most systematically left as a marker of all accounts considered so far, extends from founder-dating all the way into the most rigid corners of corporate America:

> So when it all began for me, twenty plus years ago, I was a nerd in the corner in the cube, nobody really cared about me. Analytics was not a career, it was not something cool, it was hardly even known. My grad department, let's see, from the master's program about half went straight to PhD, from there they'd go to academia, or pharmaceutical, or government, mostly. The rest of the master's program, they didn't go for the PhD went for academia, or government, and about three of us, out of like fifty, actually went into business. And people were like 'what are you doing in business, there's no career in there.' And I said I just don't find these other things interesting, I think it sounds like fun, I'll do it. So the point is, I had no idea this would turn into a great career, and I got in fact very lucky.

Recalling a time long before data science, Bill here shares his experience exploring options different from his peers' career choices, leading him eventually to make surprising career discoveries. The theme of turning against conventional choices mirrors the other accounts of educational experiences. This time the focus is on a career process within a bureaucratic context. Although we have heard before of firms who changed their existing technological infrastructure in response do data science requests, this is the first reflection on a more long-term development in the context thought most hostile to the autonomous definition of tasks we have seen in data science.

And as could be expected from a bureaucratic setting, getting "lucky" involved frustration along the way, with tasks that were:

> [t]otally removed from business and removed from IT. We were these little, you know, ninja teams on the side, ah you know, from shadow IT. But I grew up building the systems to bypass IT, and have all the analytics environment, that's what I did for many years.

Whereas the first moment, twenty years ago, of leaving academia against the common choice is consistent with today's experiences, the existence in the shadows of IT is explicitly not. We can just recall the many accounts of friction as modern data nerds enter organizations and encounter established functions. In other words, Bill describes to us a situation in which he actively worked on undermining bureaucratic rules as a "ninja." This reminds of the hacker project otherwise more familiar from Aaron Swartz's challenges from the outside. The major difference is of course that Bill's purpose was not to change the bureaucratic arrangement on the basis of principles. He was focused on practical aspects of data analysis work. Therefore, instead of moving on to another projects, as hackers would do, Bill remained in his role. The overall situation seems to have changed.

This change is salient to Bill as well, who takes the opportunity to reflect on the new situation. The audience is invited to find comfort in today's situation compared to back then:

> [N]ot only are now analytics people at least sitting with a decision maker, if not being a decision maker, right. But look at these people, Harvard Business Review, Forbes, saying it is sexy, it's the hottest job. Ah, you know, just this past week I got contacted, first time, for, to, ah, they were looking for board members for a public company and they wanted someone with an analytics background. This is amazing. Never, first time I've ever had a call like that. That would never have happened that someone like me would even be considered for such things. This is huge, not for me, it's for the trend, for anyone who's in analytics.
>
> I always say this is the one downside of all of this hot stuff that's going on today, is it's very little upside from being the sexiest job, the most in-demand job, the fastest growing job. So I always say; enjoy it while we can, because eventually the hype will die down, right, we can't maintain this forever. But it's gonna have a very nice, stable career path.

Bill's experience of a surprising new career for him entails the recognition of a larger trend in the job market. This long-term view also provides a slightly different perspective for some other accounts, especially those struck by the struggle of finding their way into the field. Part of this is likely because Bill's perspective observes the situation from within an organizational setting. As his formal affiliation has remained relatively stable, we gain a more robust understanding of how concrete practices are more relevant for the data science community to identify with than the creation of new positions that often come with them. To be sure, in this account we could easily confound larger technological changes with the effect of skills. The following account presents this relationship in more detail. We have also seen in the first chapter that new technologies by themselves do not define data nerd tasks, and in the previous chapter that it takes the integration of skills from different context. This account therefore does clarify the relationship between insurgent improvisation of Aaron Swartz and the scripted bureaucratic rules of Bill Gates. The community of the former penetrates the boundaries of the latter, albeit, or because of, relying on an anonymous basis.

What is then different today from the time when Bill already did similar work back then, such that now it is so salient? Yes, technology has changed. But again, we have seen that specific technology itself is largely meaningless and irrelevant as an identification mechanism. Data nerds still have to integrate that technology into their community. This may be obvious for those starting today, but not for Bill's generation, as we see here in Jeremy's experience:

> When I think back, you know, twenty years ago, when I started in the analytics world, the bank I was working with had spent twenty-million dollars on HNC custom neural network hardware and software platform on top of it, running on top of a new Teradata data warehouse, you know. Today, you know, and there was no community then, you know I had no peers. … Whereas today we have rooms of interesting people, a community, I think it's a very different world now to what we had then. And I find it, I think it's, I love it, actually.

Here for the first time we hear explicitly about the experience of data science as a community of experts. Jeremy's view reflects more long-term personal experience, and thus reflects technological change:

> Data science has the potential to bring together a community of likeminded people who believe that good decisions are made using data, not based, you know, based on a meritocracy, not based on what school you went to or what vendor you are associated with or whatever. And data scientists can use open source tools, available for free, on their laptops, with data that they can download off the Internet. It's a far more democratic process, you know.

Technology matters once again, but here we also see why not by itself. Appropriate data technology was available a long time ago. The community comes with open source tools. This is consistent with accounts on project work from the previous chapter, where new data science hires replaced the proprietary technology with open source tools. But we also know that open source itself is a collective product that induces more specialization than those replacements of one set of tools with another would be consistent with. The question is therefore what creates a distinct and visible community of data nerds that may work with Linux but not know the "cat" jokes of the community creating it.

What is the magic, in the words of the chapter four, which makes the new technology useful? For the first time we see the identification with a set of values. These fare without formal markers of professional status, which data science indeed largely lacks. They have been commonly associated with professional groups (Goode 1957), but their relevance for their status has been rejected (Larson 1977). In other words, these remarks offer strong indication of the silence of a thought community among the data nerds. By itself it does not yet explain the basis of this community, however.

The second remark, on open sources tools, seems more promising after all. For seeing this promise we have to look beyond the communities around Linus Torvalds and others that define this work already. As Jeroen described above in his presentation of how the command line integrates different tools, here we see how others have the sense that the combination of these tools integrates a community. Unlike Bill Gates's proprietary software, the kind the bank worked with twenty years before, the open source tools, which Linus Torvalds's and his movement have much contributed to, allow to spread relatively easily and integrate applications to specialized context. At the same time, they do not claim ownership and prevent others from adopting the tools by themselves and for their own purposes. In other words, the inverted heterarchies that organize their construction do not define their recombination for and application to data problems.

This sense of a community, although not so explicit among those moving from their education into data nerd roles, is shared more widely. As a way of celebrating this new situation, Bill, the speaker who began as a "sort of ninja" in the cubicles invites the audience to participate in the following exercise:

> How many people here are analytics, data scientists, or some variation thereof? And everybody knows them. So here is, here's a little exercise I like to give people to do: When you guys are getting ready for bed tonight, just stop, and like put down your toothbrush, your comb, look in the mirror, smile at yourself and say: 'I do analytics I am sexy.' Okay? And for the first time in any of our careers, a lot of people would not argue with this, right, so enjoy it.

Reading the last two notes together, we see that the transition can be experienced as much as part of a larger, exogenous trend, as it is experienced to be created by those data scientists I have considered before. And even Bill admits that he has anticipate the transition for longer than it has been recognized That very account, however, stems from a public presentation on data science, just like all the others. Independent of the external processes, at these public events speakers construct narratives around personal experiences, sometimes colored with intimate feelings, which a larger community can identify with.

These presentations render processes of community formation that we have seen traces of throughout previous chapters in substantively richer and more comprehensive terms. They reveal academia and the corporate world as equally relevant institutional contexts, connected through open source software. They also show significant variation. Plain statistics education has not been seen to produce the practices associated with data science today, nor do they follow exclusively from advanced research training. In particular, the rhetoric of "recovering" from academia and hitting "brick walls" on the way into data science roles reject the notion that academia as such defines the skills underlying the shared data science recognition. Finance, and corporate experiences more generally, occurred in accounts of former academics and those without PhD level training. It nonetheless appeared less significant. Nerds mentioned it in passing, although it is likely to have contributed to their level of quantitative experience.

Across these variable contexts, accounts consistently show patterns of pragmatic search for, and exploration of opportunities, guided by data analysis. Here it seems quite clear therefore that data science does not directly follow from the fields and disciplines defining components of its core knowledge and skills originally. But to also explain data science's distinct expertise, such an argument would need to be able to specify what it is that channels these varying backgrounds into a common direction. We have

seen here several mechanisms of community identification. They were most significantly dominated by an experience of overcoming uncertain conditions as to how to apply supposedly relevant skills and expertise and by sharing these experiences the sense emerged of a community of likeminded people. Although it is not surprising that uncertainty entails a bonding experience, it has nothing to do with data specifically. While data offers a common theme, without further specification it remains so general as to also include many experts and occupations not part of the data science movement despite their involvement with data. I pursue this direction in the next chapter.

These accounts were focused on entering the field. Next I consider moments of defining the data community and ideas specifying it against opposing views in order to better see on what basis data nerds draw distinctions.

7.2 Defense

Unlike established professions, which rely on specialized training that leads to formal degrees and certificates, accounts of joining the data science community have revealed the sense of a collective exploration without formal coordination or recognition. We found a combination of informal trial-and-error tactics that were anchored in the frameworks of formal coding languages and analytical strategies.[72] These observations show moments in the definition of the emergent data science thought community.

The focus on situations in which data nerds entered the field could inflate the uncertainty that we have found is part of it. Everything seems new and exciting before one gets to know it. I consider further situations here with a focus on accounts in which data nerds define their contributions. For data science to be fruitfully considered a distinct thought community in spite of diverse backgrounds and applications, we would expect that its members define their tasks and overcome conflict and friction in comparable ways. Because ideas are easier to have than to put into effect I consider the two aspects separate from each other.

7.2.1 Ideas

One way to articulate distinction is by resorting to broad and general rhetoric:

[72] These public meetings and the interactions between speakers and their audience capture one moment in which this anchoring takes place. They are so informal that both speakers and audience members have beers or soda and pizza before, and sometimes during the presentation. Some groups go to a bar afterwards. Yet, these interactions clarify the content of data science work. None of this would take place without the developments of coding languages and quantitative techniques outside of these events. These developments enter through presentations and the formalized scripts, packages and applications they show.

> … Thou shalt analyze thine data in its natural form. … [S]o, this is what big data looks like [showing slide with some text]. This is literally what big data looks like, because this is the Wikipedia article about big data. [Laughter in audience.] But, you know, big data is free form text, paragraphs, you may want to do search here, fastening, some simple aggregation. Ahm, my friends in finance—anyone?—should be able to recognize this [different picture], this is the fixed data format, this is, you know, what finance and census look like, it's a bunch of key-value-pairs, right. And there is tons and tons of that gets generated.

This is one of Ten Commandments the speaker envisions for big data, with this one specifically describing implication for analytical applications. Considering their cultural connotation, this rhetoric of the Ten Commandments indicates the sense of a community that seeks, or is believed to at least benefit from, the sort of directions commandments offer for ensuring appropriate conduct. This very specific note has a technical basis as well, which emphasizes data in its "natural form," or raw observations. The presentation of a Wikipedia article indicates that although we think of quantitative analysis in terms of numbers, raw data refers to words as well. In terms of community identification, data nerds seeking membership count the words themselves, instead of leaving that to others who prepare the data for them.

This sends a clear statement in a way that is at odds with the organizational arrangements we have considered so far. If the terms of practice are so clear as to be articulated in the form of commandments, how come data nerds had such a difficult time finding their way into this work? We can also recall the contact improvisation of the kind Bill's ninja team reflected as it undermined formal rules for over a decade. At the same time, the idea of natural data forms is sufficiently broad as to include different activities, pertaining to different natural occurrences. Although it takes the style and connotation of a rule, its content rejects specialization. It is more a protestant than a catholic reading of the scripture, or translated in the secular world of modern technology nerds, it introduces a general principle in order to preserve autonomy of the professional groups Mills saw lose it.

Looking closely, we can also see that these commandments are in fact based on practices in place already, which is consistent with their religious origin (Brown 2000), but still break with earlier ways of analyzing data in their composition:

> This is JSON, probably the trendiest data format of all. It is sort of semi-structured, multi-structured data, where things like JSON, Avro, Parquet have made these possibilities. Mongo, right, mongo has made these huge bet on making sure that data should stay in this format and not just for performance scalability reasons, but because there is an extra bit of expressiveness here that you just cannot do if you put the data into this next format [new slide showing a spread sheet], which are tables, which everyone knows and loves.
>
> You know, this is big data too, there is lots of just straight tabular data that exists in the big data world, the difference is there is hundreds of billions, or trillions of these, they have lots of columns and you still have to do lots of relational joins.

Shant, this speaker, describes some of the diversity of modern data with respect to available technologies addressing it. He also acknowledges table formats as an established and accepted way of organizing data, and by extension work associated with it. That way this account directly positions a novel area of problems relative to that adjacent, established one. The lesson here is not that modern data scientists ignore older data problems. They embrace both. The distinction between data scientists and other functions working with data emerges to the extent that some might choose to focus on a data technology, either new or old, and not data itself.

Shant emphasizes as well that the effect is not additive. The new data formats, in addition to storing different kinds of data, also provide "expressiveness." Identifying this benefit resonates with accounts of "creativity" and pragmatic "tactics." How does the work with old and new techniques differ specifically?

> … one of the things I remember having some … lively discussion to figure out how we're gonna integrating [a newly acquired startup] into all our stuff. And they had this big data analysis process and it had a very logical flow through, well you start with the problem, you do these other things, and I tell him, 'okay, this isn't any different from what I have been doing for twenty years, it's the same process.' 'Oh no, totally different, totally different.' I said 'I agree the problem you're solving is different, but the fundamental process, and the thought process, and the skill set, everything is really the same.' And so I remember how I finally got it through to him, and this was one of the Silicon Valley guys, didn't have any experience with traditional methods, and to them this wasn't it, and that's great, I appreciated it. I went and got CRISP-DM, and if you don't know what this is, look it up, it's a data mining process flow from like 1998, 99, … and I stuck it on a PowerPoint, next to his flow, and then I put a little table underneath where I lined up the steps just in little bullet point. And I said 'what do you think about this, what do you notice here?' 'Bill, these are almost the same thing!' I'm like 'yes! …' …

And because skills remain unchanged in this account, existing guidance continues to apply:

> A lot the disciplines that we have around how we define and execute analysis, again, still very much the same. Specific methodologies and packages can be different, but you still have to come up with the idea, get the data, cleanse, clean, prepare the data, run some analysis, determine if it actually works or not, if it does, deploy it. Those general steps are constant. So, there is my rant.

In this anecdote Bill recounts a conflict over appropriate frameworks for data solutions, and his evaluation of the quality of the dialogue. We can see from this again an argument for a comprehensive and inclusive view of data. But Bill's motivation here seems to be sufficiently different. His framing around the ongoing utility of existing disciplines rejects the need for a novel set of commandments or additional expressiveness of the kind the previous account brought forward. In other words, new and old work differs not so much. Or put differently, whereas the two sides of this argument agree on the mutual benefit of the different kinds of data, one sees this inclusiveness in a redefinition of older practices, whereas the other view integrates novel problems with the older practices.

There need not be consensus. These conflicts amid core agreement strongly underscore an interpretation of the data science thought community that relies on direct coordination and collaboration as much as on more commonly shared knowledge. Perhaps it is also too early for consensus. Even the side endorsing old ways recognizes that something is different here. As Max Planck famously said for science, “advances take one funeral at a time.” We see this combination here in the two sides directly arguing with each other with positions of which we have seen support separately already.

7.2.2 Practice

Things are often easier said than done. Intellectual perspectives define modern data nerds on the basis of inclusive views of data and give little relevance to technological or organizational forms of specialization. How do these ideas work in practice? I now consider accounts of problems at work, where these factors hang together more closely.

We begin to see first complications here:

> Generally engineers are often involved in [the process] as well. The data is then delivered to the analyst. Analysts generally work as a team, right, data science these days is no longer a solo activity. You have people working together, they have different expertise, have different areas of business domain knowledge, and also statistical knowledge. They work together through this data discovery process and the result of what they discover is then made available in the form of reports and dashboards for the rest of the organization. And of course it is never this nice and neat, and it is never a linear process. You might have comments and questions in the collaborative process, and, what’s really painful is when you get the data and you find that it isn’t really ready for analysis and you need to further transform it.
>
> … when a lot of analysts look at their data and say ‘hey, I need to do a transform,’ they then need to cross the boundary and the data needs to go into a different tool, they need to pick up the phone, call the engineering team, and then they need to wait a few hours, if not days, before they can get new data back.

In the soberer context of practical problems, we do after all find a bureaucratic solution that tries to mediate any potential for friction by assigning clear responsibilities. As much as we have heard about integration of the stack, here we see the tasks organizationally separated in ways that correlate with its technological components. This division preserves the established data analysis flows, even as specific methods change, because each group of experts can separately adjust as necessary. This view of a highly specialized and clearly defined role undermines notions of a distinct community, as the definition follows from the organization, not the nerds themselves. As Mills predicted, bureaucracies absorb autonomous work.

This was one view. Another interpretation of bureaucratically specified and separated practices reconfigures the arrangement such that the experience changes noticeably. The following interaction among panelists reveals the nuances of this interpretation:

> I think it is about having a contract between what your data scientist does and what your infrastructure team does. So, often I think the best form of that contract is SQL ["Yes" from a co-panelist]. So, you simply say look, infrastructure team, we want to create a separate analytic sandbox where we push the data to, because the last thing you want is you data scientist running queries against your operational system. [Laughter.] I think we all know what happens when that goes down. And if you have your data scientist given the expectation of they have to know SQL, then if you get Hive installed on your Hadoop cluster, they run their queries and get their data. But it is critical that they have a certain level of data engineering expertise.

Here we see a solution in which it is not bureaucratic divisions between teams preserve older technological limitations, but technology encoding bureaucratic rules. We still have separate teams. Instead of picking up telephones and giving orders to each other, in this solution skills are continuously distributed such that those who primarily focus on data analysis can retrieve relevant data directly through the SQL interface. This interface only gives them access to part of the data, however, such that some jurisdiction for the team focused on IT preserves its jurisdiction. Whereas the infrastructure side needs not to worry about data scientists interfering with their data, or bothering them with requests, they also lose control over the data they keep aside for data scientists to integrating into the analysis process.

Several panelists agree with this idea, indicating another moment of community identification that is consistent with the earlier emphasis on data in its natural form as well as presentations on the command line that moves data from one tool to another. How should we understand the opposition between this and the previous account? If data nerds want to preserve their distinct and autonomous status, what are their arguments?

> One thing that Claudia [Perlich], who [my co-panelist] was mentioning about, at strata, she said 'I have one rule, which is,' and she is, you know, a renowned data scientist, said 'I never let other people pull my data for me.' It is critical that the person who's doing the algorithmic work have the ability to pull the data out of the system, because it's an iterative process, this is why speed is so important. Because if it takes four hours to run a query, often that first query you run is not the right one. So I think it is about having a good contract between your data scientist and your infrastructure team.

Many accounts of projects and skills have emphasized or reflected the iterative aspects of their work. Whether data nerds try to predict appeal of fashion dresses or improve education, their work requires moving back and forth between data collection, coding, formalization, analysis, and all over again. Contrary, the idea of a contract between different teams introduces the definition of bureaucratic division of labor, which implies separate and specialized tasks. The speaker sees no contradiction in this. This is because he invokes the idea of a contract as a description of computer code. This techno-

bureaucratic argument distributes skill requirements across functions that were previously separated between them, even introducing redundancy, at least on one side.

As we consider the defense of a community here with an interest to see if the uncertainty of data science tasks is more profound from an outside perspective, the question arises where we can draw a line with respect to sufficient coding skills. To begin with an extreme case, could there be data nerds without coding skills? After all, data analysis software packages are available that provide graphical user interfaces for interacting with data and methods to create models. We have also heard about proposals that aim to automate data science altogether. Someone on the panel we have heard from here considers this point of doing data science without coding, and quickly dismisses it, meeting agreement from the others. A data scientist needs to know how to code "really really damn well," in "SQL and a whole bunch of other things." After we have heard much of SQL now, to data nerds it does not provide a very clear boundary. The sense among novice data scientists that their transition was "funny" resonates with the sense among practicing data nerds of a "bunch of other things" they need to know.

Turning to another extreme, we can ask whether there is an upper limit of how much coding a data nerd needs to know?

> Right, but what they don't need to know is MapReduce. I don't think,—it is okay for a data scientist like Claudia. The evidence is when they install Hive, which is a SQL layer on top of Hadoop, Cloudera's Jeff Hammerbacher has told me that once you install Hive on top of a Hadoop cluster, the usage of that cluster typically goes up by a factor of ten. Because you lower the friction of getting data out of that system. I don't think that, I mean, I speak for myself, I've done a lot of data hacking, I'm not a java programmer. I can write R, I can do Python, I can do Perl, I can do UNIX, but writing down to the level of, you know, MapReduce scripts, in java, I thinks that's …

This panelist reiterates the list we have heard much of, especially R, Python and UNIX, but also Hadoop. We also hear again about Claudia, and her arguably special skills and status. She is a prominent data scientist in New York City, who had the flashlight app ad placement recommendation and whose role the next chapter considers in more detail. Focusing on the skills for now, we also learn that his colleagues on the panel don't agree with all of it. Careful not to directly contradict their co-panelist, one suggests that "I guess it's a continuum. I mean, you know, I would much rather somebody did know MapReduce."

This disagreement makes up a fleeting moment of the discussion on stage, but allows for more profound insights than this salience would suggest. One side excludes those analytical tasks data science encounters as they become so large as to require a different programming framework. It thereby implicitly

reinforces specialization in a pattern where we have so far more often observed arguments for integration, even from different motivations. Moreover, this account's reference to Claudia Perlich's work seconds the earlier role of techno-bureaucratic contracts between different teams.

Similar to how contracts define bureaucracies, work and activities of prominent individuals, like Linus Torvalds and Aaron Swartz, define the informal technology world. In the data science context, however, we find disagreement whether some skills are so specialized that only outstanding individuals can be expected to be competent in them. More in support of the view that questions this definition, we can recall accounts describing the ease of moving tasks to MapReduce, for textual processing in that instance. Settlement of this issue is of course not required for considering data science as a thought community. Quite the opposite, that these different perspectives emerge in such a focused group indicates its richness, heterogeneity and hence applicability. From this perspective, accounts defending data science work fall short of revealing definite boundaries of relevant tasks, of settling on what is part of it and what is not. The contour emerges much more from the debate over what is relevant for a problem instead of settlement of such questions. So far it seems that much remains negotiable even among practicing data scientists, expect that it must involve data, and coding.

If coding is so central, why are we then not conceiving of this new role as "data coders?" Porous definitions of the responsibility associated with the role of data nerds emerge in non-technical aspects of their work as well. The following response to an audience question regarding the highly visible turmoil at the magazine The New Republic illustrates this point:

> … as you know many startups failed not for any like product-market-fit hu-ha, but just because of people problems. And so there, I'd say what you have is you have a craft, you have the craft of journalism, and like any craft, you have like a guild that has perfected that craft for centuries, … there was some clear real culture clashes,… a lack of respect for the workings of the people in the guild. Right, like those people they have a set of values, that you should respect. … And I think that, I guess that speaks in part to what a data scientist should do, which is be a really good listener. Right, 'cause you really cannot go into a biologist, or a social scientist, or a bunch of journalists and say 'I'm going to deep learning your craft, and you need to stop what you're doing and need to replace it with convolutional,' whatever, I mean. You really need to be a good listener and understand what their values are and then figure out the extent to which your skill may be useful in advantageous those values. And sometimes they are not.

Seeing how this speaker accepts that data science services may not be useful under certain conditions, and definitely not in the form some data nerds may like to anticipate, extends the conclusions we have drawn so far. We have consistently seen views that emphasize how data science integrates new and old forms of data, or tasks from older organizational functions with those it primarily focuses on, as

well as how data nerds combine different analytical techniques. This account adds to this list of practices the integration of different views of underlying problems, also those from non-quantitative perspectives. The kind of negotiation over relevant approaches extends beyond appropriate code to include questions of how to encode substantive problems.

The discussions of data science practices reveal as much of the experiences this group shares as the specific information it provides on a general set of tasks and skills. In combination, the evidence in this section suggests that data science is indeed not a result of clear definition, or deliberated agreement. Community identification unfolds, and makes data science visible through integrating experiences shaped by heterogeneous contexts and instances of applying such expertise. In this view, data science relates to the social stock of knowledge and gains salience not despite, but because of the disagreement over the details of its practice.

But if there is disagreement, why do the different camps continue to talk? Indeed, this last question with respect to defining its boundaries, and why not focus it just on code, has still left us with an important question: Why science?

Chapter overview

In the chapters so far, community identification has emerged through shared recognition of data with respect to substantive problems and technical capabilities. How salient are those common principles to the nerds themselves? The first section suggested that data problems attract a wide array of experts whose respective backgrounds did not reflect diversity in data science practices themselves. Regardless of the specific training or previous occupation, data nerds struggled with applying those backgrounds in part because problems are not narrowly defined. The second section initially lent more support to this interpretation by revealing bureaucratic definitions of the problems that divide skill sets across different functional roles, most often data science and IT. Additional arguments that revealed ambiguity in those definitions on a technical and skill level remained sufficiently contested as to further reinforce doubt of data science's integration as a thought community. In other words, speakers who explicitly reflect on their community experience did not provide much direct evidence of the shared principles we were able to infer from accounts in different contexts. Closer analyses revealed, however, that oppositions emerged from agreement on core questions viewed from different perspectives. What seems like inconsistencies

indexes data science's diversity without undermining the idea of a distinct thought community. Seeing disagreement of appropriate definitions of the work as evidence against the robust role of data science expertise would, moreover, prematurely dismiss less practical comments, such as invoking authorities on the problem, as seen in the previous debate in the references to Claudia Perlich and Jeff Hammerbacher. I turn to them next.

What should we make of data science then? The evidence we have seen here rests on considering modern data problems on this abstract level that facilitates integration in the absence of a singular definition. We have seen how the expertise addressing these definitions to a large degree comes together through the nerds themselves, and to a much lesser extent remains in the organizations, which they apply it in.

This finding is significant for addressing public concerns, which in the past have quite successfully centered on organizational actors. That is still possible and likely prudent in the data context. It also overlooks continuity in the knowledge that underlies the applications causing wider discomfort. This suggest that we also need to clearly understand not only the terms of the community but also the deeper principles that hold them together in order to embed in it terms that meet public acceptance. We have seen some beginnings of this in the last account that emphasized openness to non-quantitative views. This could be extended to public views.

The complication that still remains for this step is related to the nerd careers, and hence also individual prospects. Their paths into the organizations, and their roles within them, follow no familiar patterns. This finding is relevant as it indicates opportunities outside of institutionalized channels. It also challenges my conceptual argument of a thought community as it requires to clarify on what basis these steps align instead. Before we turn to this question, I take one more step where I consider the significance of this way of organizing expert knowledge relative to alternative models, seen through the familiar nerd roles.

There are also conceptual conclusions as to how we understand the data nerd role. Consistent with our expectations, here we are left with a resurgence of Bill Gates's proprietary style of organizing as the basis of facilitating access to the community, and letting members stay. The bureaucratic task definitions providing such access remain sufficiently broad, however, for the data nerds themselves to define their

role in multifaceted ways. Considering accounts as nerds joined the field, we have seen repeatedly how they were quite willing to address some specifically defined problems, but we also found opportunities for broader applications of the skills they used to do so. In other words, once data nerds have jumped the hoops of signaling specialized knowledge, they use the space they have gained access to for implementing solutions independent of their bureaucratically assigned task. We have seen here directly that the different roles layer on top of each other in careers. That layering is also facilitated in the organizational setting itself. Computer languages are seen as contracts, hence introducing a particular kind of dynamism that is directly linked to those nerds literate in these languages. In other words, there is no learned definition of the otherwise contradictory skill combination. Although we recovered the technology nerd models on the basis of their different practices outside of the data context, here we have seen on the level of community members that have taken the different perspectives for different problems, and combined them along the way.

Here we have seen significantly more traces of anonymous role references of the kind that fits Mills's hero, who guides later generations. Careers offer no explanation for how such an anonymous role could outlast the generation of those nerds who experience the layering of their skills and knowledge now. The literature on the learned professions that followed Mills emphasizes abstract knowledge as mechanism for integrating such diverse experiences and perspectives. As far as it has remained unclear here what brings the different careers together, these integration principles will be the focus of the next chapter.

While these questions need clarification through further empirical observations, some others can be answered on the basis of those from this chapter. As data nerds share their stories and views of how they have come to do this work and what they think it should be in these informal settings, they blur the line between personal and professional interpretations of their work. Because the settings in which they share these views and experiences involve other nerds, professionals and experts, we need to consider the implications for the definition of a distinct community amid these otherwise distinct experiences.

Contours: Intimacy

These accounts embed the skills and practices from the previous chapter in their social context. They share a notion of undefined career paths und contested principles of applying expertise, which

speakers rationalize specifically with references to sequences of formative professional moments. These narratives implicitly reveal experiences of initiative, uncertainty, struggle, opposition, and indeed, isolation, as community identification moments. While transitions through these experiences are not formally mandatory for individual members, their effects and open discussion defines the community nonetheless. Remembering the public setting of these events, sharing their anecdotes of initial struggle and uncertainty regarding subsequent steps creates intimate moments between speakers and the audience. Whereas such common memories contribute to the formation of thought communities, they undermine formal images of expert status and thus unsurprisingly appear on neither side of the sociological debate as a central concern. I therefore draw on the literature on intimacy, which leads to a different interpretation that reconciles community formation and status preservation.

This focus on intimacy emerged from accounts of speakers who, in addition to nominally describing steps they have taken and skills they apply, share with the audience their personal struggle that was part of it. The audience is invited to imagine the process of navigating mathematical statistics and machine learning cultures and of pursuing deviant ideas, which we saw in the previous chapter, in arcane terms, through the much more broadly applicable experiences of education and training choices and career changes. These descriptions admit struggle but emphasize encouragement more than warnings. To be sure, some speakers propose attempts to ease those paths, but others interpret the same kind of transition into the field as mechanism for acquiring the appropriate combination of otherwise distinct formal and informal knowledge. In the context of these settings, those individual experiences become collective memories of integrating and defining areas of expertise.

I now consider the relationship between intimate knowledge and community formation beyond the information that data nerds share. I turn to the sociological literature on intimacy, first in the expert context, and then, because of limitations in the treatment there, more broadly. In the sociology of expert work, intimacy appears in analyses of experts and clients, such as doctors and patients (Freidson 1961) or lawyers and clients (Uzzi and Lancaster 2004) and experts and the lay public more generally (Collins and Evans 2007). Two main arguments result from this work. First, providing intimate settings and maintaining intimate relationships impacts how clients and patients evaluate the expert relationship, both substantively and monetarily, independent of the expert service itself. Second, experts also need to

demonstrate familiarity with the practical problems their arcane knowledge applies to in order to gain expert status in that matter. We have seen evidence that speaks to both aspects throughout the previous chapters. With respect to the first, we heard of relevant instances as data nerds described their clients' preference for inferior but more comprehensible solutions. With respect to the second, however, evidence pointed into the opposite direction in that while data nerds admitted their lay status in certain substantive areas, they consulted with experts in that field instead of supplementing their own basis. We lack the evidence for assessing the effect of these positions, however.

The first observation also helps us anticipate dynamics of sharing the intimate moments on stage. Just like patients and clients appreciate a personal relationship with those they consult, so may peers. We need to recall that the audience is not limited to peers, and thus may have the opposite effect as nerds focus on themselves and not their work or clients. Moreover, neither focus captures the personal statements of joining and defending the data science thought community from this chapter, where we have seen the nerds speak of themselves in relation to the group, not work. I therefore take one step back conceptually in order to consider more applicable occurrences of intimacy albeit in less similar contexts.

The most significant sociological work on intimacy focuses on relationships in private life, including family, friendship, romantic and sexual relationships (Zelizer 2005). It thus also primarily addresses the kind of dyadic relationships the literature on expert work addresses as well, but the greater scrutiny it has applied uncovered findings that help us understand how the role of intimacy within the group of data nerds may affect their appearance to outsiders. Zelizer (2005) has offered particularly detailed and comprehensive analyses, as the following instance reflects. Here Zelizer dissects changes in the legal interpretation of an intimate relationship between a mistress and her lover, or client and customer, depending on the kind of evidence, as we learn. Treating this matter as a tax-fraud case, a court had to decide whether to interpret financial exchanges from the man to the woman as payments, such that the woman would need to pay taxes, or gifts, so that the man would have to pay taxes. Considering that the man gave checks to the woman, a first court ruled to interpret them as payments. A second trial, however, led the court to consider letters that described the continuous affection of the man for the woman independent of specific services he may or may not receive. On this basis the second court ruled to

understand the financial exchanges as gifts (Zelizer 2005, 96-98). In short, once the intimacy was revealed, the quality of the relationship changed, although the nature of payment had not, in the view of others. Although substantially far away from quantitative analysts, I argue that the basic dynamics Zelizer uncovers can help us understand the public sharing of personal struggle in becoming a data nerd and in defending the data science role against others.

Translating this romantic case into the data science context requires some abstraction with respect to the different roles involved in it. For our analytical purpose, let me replace the two romantic partners with data nerds, respectively, and the court with the public audience attending the events.[73] With this setup we can consider the effect of the kind of conversation on how others view the group having them. We have seen variation in how speakers presented their experience. In this view, consistent accounts that present the transition from graduate school into data science as "boring" would have a different effect on the way the community appears to others than if there were instead, or at least as well, accounts that describe "recovery" and "luck." What is the interpretation then? If transitions are boring, they imply established specialization. Yet there is no formal recognition of data nerds. Therefore, more established transitions would indicate that nerds vanish in established formal organizational roles. Transitions that are unexpected, on the other hand, lead to images of no existing paths. In addition to this technical interpretation, Zelizer's arguments points out a qualitative shift. Whereas the boring transition signals just another transition from institutionalized training into a job, the intimacy signals a different quality of the process and thus the group experiencing it. Intimacy signals commitment, as the court that revised the original decision shows, even without formalized relationships, such as a marriage. For data science, the intimacy might just play that role of indexing the depth of an experience of which the terms are too arcane to vividly communicate to those who have not undergone it, and where no formal certifications are in place.

Is the comparison to Zelizer's case a stretch? Other reports on similarly intimate moments that are part of joining an expert community exist as well, although not systematically. Since for these other groups we have a way of thinking of them, such as profession or expert group, we can consider whether on the basis of these accounts here we might conceive of the data nerds in a similar way, in other words

[73] This framing is consistent with the Deweyian definition of the public we have adopted in the introduction.

replicating Zelizer's analytical use of the court setting. Reflecting on her experiences in professional training to become a midwife, Cohain (2009) recalls an exercise in which she and her colleagues examined each other, as they would as midwives with patients. In this view the intimate experience is irrelevant for the public appearance, but contributes to the formation of a community. The lack of more systematic evidence of these processes for lawyers, doctors, scientists, or other experts, might partly result from the highly formalized settings in which professions organize their training, and the highly concealed laboratories and library cubicles in which the sciences train graduate students. In both cases a formal certificate defines community membership. The expertise scholarship, to be sure, has also pointed out that informal interactions matter as well, but has focused mostly on those that transfer arcane knowledge, and not on the mistakes only instructors and advisors see their students do.

Contrary, several data nerds repeatedly admitted how much they negotiated each step of their transition, or of their application of analytical practices. The immediate conclusion would therefore suggest that data science differs from those groups. This conclusion would commit the same fallacy as the first court in Zelizer's case of equating the degree of overt intimacy with the substance of the relationship. In other words, only because data nerds publicly admit uncertainty and struggle other experts do not admit, does not necessarily suggest that their competency is inferior and that they constitute a different kind of group. Instead, these processes preserve room for data nerds to continue defining their roles by considering each other's experiences. Intimacy therefore supplements the community identification project without necessarily undermining expert status. I address the question of what holds the community together instead in the next chapter.

Two main implications follow for data science and its contours. First, existing research shows that different ways of describing the same relationship change its status in the view of others. We have seen data scientists describe their transition in different ways as well. While accounts of uneventful transitions and obvious claims over certain tasks cast it in a light of an existing and continuous group of experts, those reflecting unexpected transitions and conflict over tasks cast it in a light of a collective experience. Second, these informal interpretations are tied to formal applications of knowledge. What may appear initially as a somewhat naive self-presentation is directly tied to accounts of formal achievement. Together, this contour of the data science thought community reflects rhetoric that seems to contradict

conventional interpretations of expert roles as much as it offers arguments backing up those claims nevertheless by revealing substance that signals durability and commitment amid a lack of formal guarantees.

8 Discipline

The previous chapter has recovered an at best fragmented basis of the data science nerd community with respect to both career paths and to a slightly lesser degree for its terms of practice. We have reached this understanding after first untangling data science work, tasks and skills from technology and organizations, and then reassembling it to reach the level of community identification processes. What is then the basis of its integration?

Accounts from the last chapter could initially be seen to support the interpretation that data science describes a loose array of skills and tasks held together either by media hype or the organizational belief in their promise. In spite of complementarities across data nerds' different views and experiences, they have so far lacked consistency in the form of a systematic integration of arcane problems. Even the previous chapter on processes of joining the community and definitions of its boundaries offered no way of integrating the different components. It revealed consistency on an inclusive view and general experiences of joining data science. These terms reflected practical tasks and organizational contexts. They would leave us with an incoherent community, in spite of all efforts of nominally identifying with the generic data science label.

An alternative interpretation may acknowledge the heterogeneity and view data science as a set of skills and tasks that apply in such different contexts that they render any singular definition unsustainable. In this second view, we might still consider data science as an expert group and professional thought community with an eye on their shared focus on data analysis as a career project. Such a conclusion requires consensus among data nerds also, for instance on basic principles that this responsibility begins before data fits into spreadsheets and stretches beyond the coefficients, or plots thereof, that define results. This argument so far would need to assume that somehow the existing channels producing data scientists are arrayed in ways as to sustain this supply such that data nerds then continue to recognize each other. Such an explanation ignores the disagreement we see on questions as important as the minimal skill set, even among those who agree on the general utility of the data scientist.

A more powerful argument therefore also provides mechanisms, such as identity defined in practical experiences but also in a sufficiently abstract set of references for defining those mundane hacking skills along with the much more disciplined statistical methods. In addition to integrating these

existing though distinct areas, these references would also need to account for the adjustments nerds make amid the practical problems defining a given task as part of overcoming disagreement that may result from the dissonant experiences that mark heterogeneous practical applications.

Some of the narratives of joining data science and of defining it indicated traces of a common set of references. They are familiar from projects of community formation ranging from religion to the sciences and including all kinds of social movements. Whereas in the contexts of technology, organizations, projects and skills the accounts mostly focus on methods and techniques, we have sporadically seen data scientists referencing the groups or individuals behind methods and practices, particularly as the data nerds defined distinctions to applications that are not part of data science. These references to other nerds and experts encode expertise that we have seen consistently in data science and aspects that just the shared methods and technology would not capture. This expertise entails knowing how to integrate specialized tools in order to solve a problem they were not designed for, or for which kind of problem a solution might be available, even if one is not aware of it specifically, and for which not.

In such a framework, the public would find unconventional and unfamiliar, yet also explicit processes to rely on for addressing issues of how data science shapes everyday lives. This framework requires or assumes no governing body that could take complaints into account. It does, however, indicate that data science itself relies on coordinating mechanisms that bring relevant issues on the agenda for members across the community. In the Facebook case from the introduction, for instance, the institutional affiliation and oversight covered academic contributors and signed user agreements covered Facebook's data scientists. Legally there was no wrongdoing. The subsequent public debate nevertheless attracted commentariat from the community more broadly. In other words, the community recognized shared concerns without being affected directly or given a formal mandate.

The following accounts uncover how such simultaneously indirect and robust processes unfold systematically. They thereby also pertain to the individual opportunities for data nerds. Joining the community does require diverse skills, some technical and some not, as well as reflexivity over the processes governing the skill development and application independent of the specific practical problem they work on in a given moment. Such reflections provide a basis for defining and adjusting formal skill

applications in specific problems and tasks for which there could not have been a script, and without undermining technical requirements in that process.

In terms of the familiar technology nerd roles, this chapter aims to address the gap left previously. There we have seen how contradictory skills associated with Bill Gates and Aaron Swartz fold together in specific applications and throughout careers. The accounts have also shown that in the context of data, the older conflicts between views of specialized skills do not need to remain, as the opposition between Bill Gates and Linus Torvalds, with proprietary and open definitions respectively, would have suggested. Going further, these specialized practices can even be reconciled with the type of projects Aaron Swartz pursued that otherwise challenged them both on practical terms. While this combination would indicate a new type of technology that on its own might constitute a source for public concern, or an opportunity for individual work, it has remained unclear what brings together these careers that stood in opposition to each other in other ways.

We can turn to the literature for some explanations that reconcile these contradictions. The key idea in the research on profession is the significance of abstract knowledge for defining areas of expertise. As part of discerning the contours of the thought community data nerds form, we have focused on the constituting principles of abstract knowledge through the perspective of technical skills. This has revealed layering of skills along careers. If there was direct coordination of these careers, the accounts that define community membership would have shown less of Swartz definition of work, but likely that of Gates and even Torvalds. Instead we found recovery processes and ninjas undermining bureaucratic rules. Where we saw direct references to Torvalds's project and other open source efforts, we also learned about the need to integrate their respective specialized capabilities around data problems. In other words, it seems relatively clear from the observations so far that to the degree to which data nerds form a community, their integration follows processes other than direct coordination. While abstract knowledge offers one alternative that has been observed in the past, those existing accounts fall short of specifying its characteristics outside of their respective cases.

As far as abstract knowledge has remained unspecified, academic disciplines offer the most obvious and pertinent framework for understanding the kind of patterns around knowledge of distinct skills we have seen so far. There is no plausible way to think of Gates in the context of academic disciplines,

nor of Torvalds or Swartz. At the same time, we have also moved beyond those relatively clear and by now documented and established roles by this point. We have repeatedly seen that skills from Torvalds's legacy come to apply in the context of problems previously defined through Gates's perspective and through their unorthodox application that seems to have no direct connection to either of these contexts also invoking Swartz. We have moved beyond considering the observations in the data science community with respect to their similarity of the different familiar technology roles and instead by now defined a distinct and integrated type of data nerd.

The question is therefore how discipline unfolds in the present context of modern data processing technology. So far we have considered all the empirical sites that are relevant to those who address these problems. From this basis we have already seen another type of discipline operate, one that Foucault described, and that is implicit in Mills. As data scientists address bureaucratically defined problems, they continue with the compartmentalization of their work that Mills has seen on the rise since over a century ago and as part of that adopt the docile body Foucault (1995) recovered in modern society. Leaving it at a focus on their acceptance of the overall structures would ignore that data nerds carve out niches of significant variability within them. It follows that the disciplinary processes we have to consider here resemble very strongly those seen in the academic context where oversight comes not from departmental colleagues but peers in the wider community. This requires in turn that data scientists somehow frame a context of relevant expertise in more general terms than the integration of skills that follows from their careers. Although science is about knowledge, in disciplines prominent individuals are important still both for scientific work (Merton 1968a) and defining identity (Stinchcombe 1982).

Here I focus on those reference and consider three different ways of associating practices and knowledge with their creators. I begin with those engaged in the definition of data science directly. I move on to consider role models, where I again distinguish between those still actively developing relevant knowledge and those not.

8.1 Leaders

The "prominent data scientists" whom we learned about in the previous section, of Claudia Perlich, Jeff Hammerbacher, and some others appeared at these public events often with little introduction, and a clear sense of authority. They have also featured in media reports and interviews. And they speak up

themselves, such as DJ Patil in the 2012 Harvard Business Review article on data science, or Jeff Hammerbacher below.[74] While the definitions of data science practices in the previous section were significantly shaped by the respective problems those who delivered them had encountered, this group I consider here works on articulating a more common perspective. This should not imply that their ideas define definite boundaries of the data nerd community. They have no formal consequence. Rather, these strategic accounts provide a script for the audiences they were delivered to for interpreting experiences, personal and reported, more generally.

8.1.1 The beginning

First things first; communities need a beginning. Sociologists have defined their beginning with Auguste Comte (Stinchcombe 1982), Christians with Jesus Christ's birth (Zerubavel 1997), and Americans with Christoph Columbus' arrival on the continent (Zerubavel 1992). When did data science begin, from a data nerd perspective?

> Alright, so, thank you everyone for coming out, ahm, so turns out we just [hit] our five-year anniversary, ah, of me accidentally inventing this term data science, data scientist, or, you know, borrowing it from Bill Cleveland, or however you want to say it. But in any case, given that time has passed I thought to pull up the email, in which I sent to my team telling them `Hey, I am rebranding all you guys data applications scientists,' [showing email on slide] …

This is quite a moment. Facebook and Jeff Hammerbacher, who in this account refers to a time when he was there, are relatively widely attributed with initiating the data science role. Together with commemorating that event, Hammerbacher acknowledges William Cleveland's association with the term. How can both be true? Cleveland, then at Bell Labs, had published a paper in an academic journal with data science in the title. I had heard of the origin at Facebook from conversations with others, also academics, even without prompting for Hammerbacher's story. Among the many views and experiences associated with data science, this seems to recount their inception, somewhat. Just like with sociology, Christianity and Americans, it is clear here that there is room for interpretation.

What is more important, to Jeff, is how that event unfolded …

> … I got one person who ah, who responded back and said `You know, ahm, maybe that's not such a good idea, there is this whole title structure called research scientist, so why don't we, why don't we keep research in the title line?' And I said, 'Nonono, we're gonna drop 'research,' no one really talks about research at Facebook, we're moving very quickly.'

[74] For an extensive list see Shan et al. (2015).

Hammerbacher revisits this early argument while also acknowledging Cleveland's role. From this combination we see that the distinct contribution follows from the application of whatever abstract interpretation the idea of data science may entail. From this perspective, the two supposedly original definitions do not contradict each other.

And Hammerbacher goes on specifying that implementation:

> And so the primary focus of rebranding everyone with a single title was just reconstituting the workload that everyone had. So I thought the premature specialization that could happen where this person over here would be really good at generating a report, and this person over here would be really good at pulling data out of a database and putting it into a data warehouse, and this person over here would be really good at doing statistical models, and I kind of felt like everyone should be doing all of those things.

On the level of work there are no surprises. This justification maps onto the disagreement seen in the previous chapter where within the data analysis process opinions differed on the precise skill set of the data scientists. Those other accounts might have been influence by Jeff's description, which many in the community know of. At the same time, several of the accounts we have considered so far preserved specializations and did not invoke situations in which tasks were redefined, nor did they reflect Hammerbacher's motivation. Moreover, five years had passed at the point of this observation. That Hammerbacher still recalls this time at these events signals a sense that the idea of such a formative moment is more relevant for the community than the causal significance of its content.

Hammerbacher ends the practical explanation with a strategic note:

> So I wanted just one term, and in fact the primary focus was to get those annoying research scientists to do some real work and not on getting the business analysts to become, like somehow, like magically more sophisticated.

This rhetoric signals a widely shared stance of those interested in solving applied problems toward research. It also conflicts directly as far as the label of the frame for addressing those practical problems invokes 'science.' How can this work? Reconciling this opposition requires considering that those building and defining the data science community do not share the connotation science has for others. I have suggested in the introduction that just observing data scientists would not reveal much we could not see among other professions, or academic experts. Hence on that level the science label seems valid and perhaps strategically useful. It still ignores the less visible aspects of peer review, panels, committees and research. To others, science has a normative and a practical meaning. Normatively, science describes advancement of knowledge through systematic experimentation and observation. Practically, science

describes as a community of nerds doing work few others understand.[75] On this high level, data science fits both. In other words, it is possible to think of it as a science and reject research as being part of it at the same time.

What else could it be? We have seen by now that data nerds neither follow bureaucratic rules nor contribute a specific component to an open source project. They outright undermine existing structures and organizational principles that familiar technology nerds follow. The reference to science supports the interpretation that Hammerbacher's specification did not singlehandedly define the field. A more longstanding connection could be seen in those accounts from data nerds who apologized for their inability to present more scientific approaches. In other words, science could affect data science in a way that is inconsistent with scientific orientations and principles of work, and thereby address the puzzle other observations have left us with.

Such a radically deviating interpretation of one of the most durable and widespread institutions in our society today puts the bar up high. Hammerbacher responded to the weight of the implications of his idea. He did not revise his definition. Instead, Hammerbacher specified his idea in terms of a comprehensive course curriculum, which was to consist of a "data preparation phase, data presentation phase, experimentation and observation, which are really two sides of the same coin of causal inference, …, and then data products, which is really about prediction, and putting predictive models into production." And these are the steps we have seen in accounts of data science projects throughout, as Jeff acknowledges as well. Although the ideas may not be new, the community distinctly integrating and applying them can be.

At the same time at which Hammerbacher provides a comprehensive view of data science's background and purpose, thus promising clarity for our question of what integrates the otherwise fragmented field, the contradictions in that account raise new questions. We saw a more prominent role of sciences than most previous accounts indicated. At the same time, sciences' core activity, research, was

[75] This is largely independent of the debate I have outlined in the introduction.

rejected. From this perspective, Hammerbacher identified in science a formal model for data science, not a substantive one.[76]

8.1.2 Other myths

Hammerbacher's story is also somewhat at odds with some of the previous observations. Here we have learned about a specific moment of inception with a clear motivation and definition even in an academic program. Previous observations were consistent with these ideas of data science, but primarily focused on practical problems and applications. I have seen many times introductory definitions of data science at the beginning of those other presentations, but not necessarily this one. We began this chapter with the question of what may integrate data science. We considered scientific discipline as a conceptually plausible mechanism and found it empirically at least salient. Then we discovered that science can be interpreted in a way that ignores its chief mechanisms. What might fill this formal hull Hammerbacher presented science as?

Consistent with those other observations, and with other community experiences with incomplete origin narratives, these events feature different views of data science's roots. The following account by Hillary Mason, another prominent data scientist, illustrates the experience of an emergent process further:

> … there is this extremely powerful algorithm. It's actually a human algorithm, and it's something I've seen work a couple of times. This algorithm is that you find something a lot of people are doing, and you name it. [laughter, pause.] Now Chris Anderson at *wired* is the master of this algorithm. Ahm, this is where phrases like big data, which we will get to, come from. Data science also came from this. And it was amazing to see how there were a bunch of people who sort of landed on to this. And so when I started at Bitly, three and a half years ago, I insisted my title was scientist, because I was coming from academia, and I didn't want something that would ruin my CV [some laughter], … Data scientist didn't exist three years ago, three and a half years ago, as a field of practice. It had just started to emerge at this point. And it did so in a couple of different ways. There were a core group of people at the West Coast, who started promoting it, and a core group of people here in New York, who started promoting it as well. And we used to have brunch and like argue about should it be data science, or should it be something else.

Similar to Jeff, Hillary combines somewhat competing, if not contradictory accounts. One the one side, she describes an obscure "human algorithm." On the other side, we learn about concrete groups that defined data science, one of which she participated in. As in Jeff's account, this contradiction promises analytical leverage. The second and the more obviously conflicting one juxtaposes Hillary's account with Jeff's description. Instead of a singular definition, here we learn about not just one, but two separate collaborative processes. Even if both experiences are correct, which is possible, Hillary's

[76] This process resembles one I consider in more detail below. Ben-David (1971) describes the origin of the American graduate school system in higher education as the result of a misunderstanding of returning American scholars who falsely interpreted observations from their academic visits to Germany, which had a leading university system at that time.

experience proposes directions for some of the questions Jeff's left us with. Here we have a collective process that begins to resemble the activities defining sciences. I return to this direction in a moment.

Meanwhile, Hillary's first idea is interesting as well. She makes unorthodox use of the idea of algorithms. We have encountered algorithms many times and always in the form of computer code. Here Hillary introduces us to a "human algorithm." This reminds of the note in which a speaker showed the Wikipedia article on big data in its raw text form and without markup, thereby simultaneously showing how big data often looks like, and describing its features. On this level these references signal a deep engagement with key components of data science. They also point out a reflexive view. The specific reference to algorithms shows the salience of such an abstract idea that it gets casually applied to a radically different problem. In order to see this better, we need to look more closely.

Hillary describes how a newspaper editor takes what many know already, and gives it a name. This, in Hillary's view, defines a human algorithm. And surely the human activity involved in that it is not a natural evolution that there are suddenly data scientists. At the same time, she also notes that there are many people doing this already. Moreover, as an editor, Anderson addresses a larger audience. In other words, there are many humans involved in this algorithm. From this perspective, we see a process in which we describe a puzzling process in common terms. There are many people doing data science, so we think of it as a social process. It is neither natural nor organic nor automatic to think of that suddenly under one label, so Hillary sees an algorithm. After all, by now we are well aware that algorithms change things in subtle ways.

Finally, and to return to the way Hillary's view relates to Jeff's understanding of the relationship between research and applied work with data, Hillary's recollection of her prior experience implies that she sees data science not so different from more conventional understandings of research. At another occasion she has taken a bit more distance to the term, while still accepting it, at least relative to big data. In this respect the two clearly differ in their view of data science's identity.

What about their views on data science practice? After all, many of the less visionary accounts were surprisingly consistent. Here is one of the ideas that came out of the discussions Hillary mentioned above:

> In 2010, I wrote this [following data science definition, on a slide] with Chris Wiggins, of Columbia University, who is also another co-founder of hack NY, and this is where the OSEMN comes from [Obtain, Scrub,

> Explore, Model, iNterpret], we actually sat down to write down what the process of data science was, because we could not find it written down anywhere. This was in 2010, this is not that long ago. And there's the URL, if you actually wanna pull up our, our little essay, where we go into each bit of this. But this seems completely obvious to those of us in this room today, right. You get some data, you clean it up, play around with it, you build a robust model and, you know, you make a graph, or write about it. This was not obvious in 2010, at least not to me.

When it comes to the practical level, Hillary's experiences resemble Jeff's. Here we learn about a blog post, whereas Jeff responded to the weight of his claim by substantiating it in a course curriculum. Both come up with similar steps that we have in one form or another seen across several other accounts. Jeff acknowledged that his proposal can now be widely seen, and Hillary notes that this "seems completely obvious to those of us in this room today." We also see in more detail here the groups defining data science, which Hillary mentioned already. We learn about collaborations as well as the foundation of a formal group, both of which are important processes governing academic work as well.

They also unfolded historically parallel to each other. It is not surprising then that not only Jeff and Hillary, who have been publicly credited in the community for their contributions to defining the term and as leading practitioners in the field. There are others such as DJ Patil, who co-authored an article in the Harvard Business Review, or Drew Conway, who designed a Venn diagram that is widely used, with credits, as data scientists provide a more general introduction to their specific project they are about to present. The following tweet by Josh Wills, another prominent data scientist, gained similar routine prominence:

> Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Somewhat unlike the previous definitions, Josh invokes established fields for defining data science. This definition is consistent with previous recognitions of the relation between academic science and data science, and practical challenges of implementing it. We can turn to the familiar nerds in order to understand what this composition leaves for data science. The reference to the software engineer maps onto both Linus Torvalds's and Bill Gates's definition of organizing technology work. They address specialized problems either in bureaucratically or heterarchically organized strategies. Statisticians also most often do highly specialized work in a disciplinary organization. The type of contact improvisation we have associated with Swartz is possible in sciences as well, though rare in practice (Foster, Rzhetsky, and Evans 2015). The definition in less than 160 characters is therefore evocative and substantively

useful. Considering these specific connotations for work and practice at the same time enhances the puzzle of how to understand data nerds.

Leaders overview

We have seen here some attempts to provide succinct summaries of data science. These accounts have revealed somewhat inconsistent, but nonetheless quite thoughtful, ideas of its identity as part of data science practices. Despite different narratives, we saw consistent views of central practices. It was also interesting how they positioned data science with respect to academic sciences, with one seeking distinction and the other similarity, just like a glass can be seen to be half empty or half full. The agreement on these practical terms signals a basis for community integration and the variable associations for salience.

Although we have learned about some very specific formal and informal processes, these accounts have still not revealed a mechanism with sufficient abstraction as to integrate the heterogeneous areas and aspects data science defines its own. Once we begin to consider an integrated understanding of the skills that facilitate data science work, we also need to turn to the origins of those skills and the knowledge defining them. It suffices for the community to put more emphasis on the implementation of ideas than their initial conceptions, but such easy dismissal in this analysis would raise questions regarding data science's salience as a result of distinctive knowledge. Therefore the following two sections ask how data scientists define their expertise as being distinct from those origins, given that its constituting ideas are not new. To address this question I distinguish between ideas of those who actively produce that knowledge, if not as data scientists, and classics, through whom data scientists canonize their ideas as well.

8.2 Patrons

These were stories of data science's main protagonists and narrators. They directly shape the field and propose general definitions more so than most previous views. At the same time, they did not forget their practical background either, which we were able to recover from the subtle contradictions in their accounts.

It is for this reason that that their definitions remain too complex as to explain the coordination of data science expertise. As much as they design courses, write blog posts, and speak about all that at these events, it does not account for all the data nerds with whose work their descriptions resonate. Formalizing concrete and substantive activity and information into abstract notations leads to more scalable rhetorics (Carruthers and Espeland 1991), the specific process of which I turn to here.

The community also recognizes role models on a more abstract basis. Some of those role models have passed away long ago and cannot help shape the field today. Others still do. They interpret data science through their area of expertise, which they have often shaped directly. This ranges from academic to economic and technological contributions. They thereby provide the data nerd community with an additional layer of abstract references. By invoking them as references, data scientists undertake an additional step of interpreting others' ideas for their own purposes. While this specific activity itself rather supports a view in which data science relabels old disciplines, the collective process of doing so suggests otherwise.

As before, I distinguish between practice and ideas, or more fitting in this setting, conduct and terms. Unlike the distinction in chapter three, however, where ideas were easier to have than to be practically implemented, in sciences ideas are much better documented than practices. I therefore reverse the order here.

8.2.1 Conduct

Let us first consider the basic tasks data nerds work on, and for which this speaker thinks …

> … everyone has a set of processes they do. I often, I wrote a blog post a couple of years ago, called, ahm, 'the three sexy skills of data geeks,' and in retrospect I should have called it data scientists, you know, if I had known that that was gonna be such a hot meme. And I talked about, ahm, munging, modeling and visualization as the three, ah, the three skills. And that was really playing off of a quote from Hal Varian, who said that statisticians will be the, this next sexy job profession, you know, of the coming decade. And I think that Hal got it mostly right, except that it's not just statistics, right, it's not just modeling, it's really important …, munging data is really hard. We all know when we work with large data sets, eighty percent of the time we spend is in structuring those datasets into a format that we can use and then put onto something like, you know, [a competition].[77]

The basic definition is mostly consistent with that of Jeff and Hillary as well as the descriptions form previous chapters. Recounting the origin of this comment, however, this speaker invokes a prominent economist, who moved from academia to Google, and the speaker suggests an extension of those ideas

[77] Considering competitions as an aim here entails at least some irony, as the speaker says it with a wink to his co-panelist, who organizes them. We should therefore read the comment to suggest that working with the data prior to analyzing it formally entails a significant component of the overall analytical work.

he cites. We have seen throughout the chapters so far that what sounds like a simple extension is easier said than done. Statistics draws on and extends a comprehensive stock of knowledge for quantitative analyses. Here we have heard in so many accounts practical steps for transforming data and arranging it. We have also learned about analytical challenges that, while not requiring mathematical proofs, go beyond implementing textbook cases. The rhetorical challenges of articulating the utility of all this arcane work with respect to a practical problem concerned data nerds as well. Finally, this proposal entails a challenge of then integrating all these aspects in more inclusive themes than statistics offers. In other words, we cannot understand this description as a practicable set of instructions. Instead, it seems to function more as a guide to all the relevant knowledge that substantiates the statistician's status.

The extension not being as simple as the previous accounts make it sound, can also be seen in further observations. Although the initial idea of data science was, as Jeff Hammerbacher acknowledged, conceived of in statistics, we can see other reference to non-academic origins and foundations:

> So, Michael Olson, co-founder of Cloudera, chief strategy officer, when he was up here [a prior iteration of that event], he kind of said something in regard to oracle, I'm going to paraphrase, but he said something oracle gives you min, max, median on tabular data, and it is a hundred-billion-dollar industry. And, you know, we do advanced analytics on a thousand times that data, and I feel like it's got to be worth more than that, right. And then in response to that he said, 'hey, I've invested a lot in this industry, so, everything I say will be self-serving.'

Both Hal Varian, who was cited before, and Michael Olson are prominent in the tech industry and beyond. Whereas Varian's comments speak to the analytical side, Olson's offer an interpretation for the relevance of data technology. Here his view of the value in those tools and methods helps the data science community delineate its purpose and prospects. This valuation refers to the simplest data operations. These are also the kind clients often find more appealing. Seeing an economic reference here reminds of the status associated with analytical sophistication. In his study of more collaborative technology projects, Stark (2009) has found competition of monetary and technical value systems. In the less bounded case of data science, they rather seem to complement each other.

In other accounts, data nerds rely less on institutional status as they seek guidance in broader ideas:

> And the main thing that I spend a lot of my time doing is plumbing. Ah, and this [slide with two versions of a flow diagram] is from a blog post by John D. Cook from a couple of years ago. Ahm, if you don't follow is blog, he's amazing, ahm, a programming philosopher, I guess, ahm. But, he talks about a lot of diagrams of, you know, data systems, or any kind of programming systems, look like the one on the left over there, where you have like these big data bases and you have got these big processes and the cloud and whatever. And you have got these simple little lines connecting them together. And that's nice and easy, all you have to do is, you know, send it from one to the next. But really what you end up spending your time on is this, the arrows and

> the lines are huge. Ahm, the amount of time spent making sure that your data flow is correct and reliable and solid is enormous and then as long as that's there it makes it a lot easier, so the amount of effort required to use it is much lower.

We learn that implementing data processing pipelines is much more complicated in practice than their parsimonious paper drawings suggest. That formal representations rarely capture empirical richness is well known in sociology, seen in the sometimes sharp divide between quantitative and qualitative research. Here we see that data scientists directly struggle with this ambiguity between substance and abstraction, and appreciate that others point it out. The reference to the relatively undefined position of a 'programming philosopher,' which, in spite of its label citing an academic field is telling with respect to finding orientation because it carries no institutionally recognized status. We cannot establish here John's accuracy or originality. He appears to have helped this presenter with his work, however, and his blog post finds exposure in this public data science meeting. More institutionalized settings would have been much less likely to consider these sources. The call for guidance in such a mundane problem and the public discussion of available solutions indicate that the questions that concern this community are not limited to moving beyond the statistics discipline, or reconciling analytical and economic rewards. This orientation on the details of the work are contrary to those we would expect to see in science on the basis of arguments that view it as an institutional system with formal status signals.

The John D. Cook reference therefore corroborates the previous idea that data nerds seek guidance for navigating uncertainty. Status does not seem to be the only reason for turning to others' ideas. At the same time, status seems to be the reason for the uncertainty in the first place. A prominent finding in research on professions shows that higher status experts seek to avoid messy problems (Abbott 1981). Moving data is messy. One conclusion for data science could then be that data nerds do the lower status work.

This also implies that others just focus on developing arcane knowledge for addressing these problems. Here we see accounts of developing some of the arcane knowledge the community is also concerned with. A recent development in this respect has been in deep learning, as indicated before. The following account directly captures Yann LeCun, one of the main drivers of that effort.

> [I]t's been an amazing year, and it has taken me by surprise, really. Ah, I mean the whole success of deep learning actually has taken some of us by surprise. We were all sort of convinced from a long time ago that deep learning was going to take off at some point, but the speed at which it has been picked up by industry, research and startups and everything is nothing short of amazing.

And it was surprising because over ten years before there had been a time "where neural nets fell out of favor a little bit, in the machine learning community," or, expressed differently, they were "seen as lunatics, you know, like ten years ago in the machine learning community, which was sort of enamored with methods that could be completely understood but not necessarily very powerful computationally." In other words, LeCun shares some of the struggle creating the methodology that is widely adopted in data science today.

The audience learns also the background behind the technical specifications. The account of the struggle becomes more specific:

> So, we got together [at NYU], and with funding … we started a research conspiracy. Deep learning is a conspiracy [some chuckles], ahm, really. And the conspiracy was to kind of, you know, pick a bunch of problems or techniques that we think would be kind of interesting enough for the community to really pay attention again to these techniques, and move away from, you know, SVMs and things like that. And I have to say it was an unbelievable success.

To recall, this was a non-academic audience. When sharing that background in an academic talk LeCun articulated it in terms of the specific debates and their main protagonists of different scholarly camps, well-known among academics but barely visibly outside their exclusive circles. Though even at this level the audience here, which seeks to apply those techniques and cares less about getting the debates right, can get a sense of the struggle underlying the methods and technologies it sets out to adopt and deploy. LeCun speaks of a community, but leaves it undefined. Whoever was meant, now this group here becomes part of it.

And even these accounts of that beginning with the deep scholarly struggle over appropriate research and methodological advancement echo the earlier observation in the data science community of considering more comprehensive guidance than that academic authorities provide:

> [T]here was a very interesting phenomenon with deep learning, which is that industry picked up on deep learning faster than academia. So, it was quicker, it was picked up very quickly by, you know, Google, IBM, Facebook, etcetera, ah, Microsoft, ah, in the space of a few months, whereas in academic circles there is still people, in computer vision for example, who kind of, you know, waiting and seeing to kind of figure out how it is going to pay out. You know, between young people, you know, not just, you know, old, very old people. So there is a bit of inertia there, or resistance.

As we have seen before from many data nerds, LeCun, who unlike most of them has an impressive academic record, cites non-academic indicators for acceptance of his contributions. Just a year earlier LeCun had changed his primary affiliation from NYU's data science institute, of which he was the director, to Facebook's NYC artificial intelligence office. Transitions like this one can be seen systematically, as

even the question LeCun responds to here had cited several moves from prominent academics to large corporations. His comments are therefore unlikely to result from personal rejection by academia.

The struggle for relevance does not halt before those highly regarded from both sides. Here someone whom others turn to in search for guidance echoes their experience of working with quantitative methods outside of the academic circles primarily concerned with their development. To be sure, LeCun still "keeps a foot in the door." It is therefore not only that the audience gets exposed to nerds who stumble upon useful combinations of skills and practices for which there was not much of a systematic basis, they see some unlikely transitions of highly defined institutional status as well. LeCun's perspective thereby gives evidence of two types of processes we know from arcane work. First, we learn about a conspiracy, the community at NYU. Then we also learn about the quick spread into other institutional contexts. Both seem to be operating in data work.

Withstanding status and advisory benefit, the community does not turn to those prominent voices as sole role models, even as they at least implicitly endorse skepticism of purely academic guidance. Evidence of their utility is important as well. For instance, deep learning methods are well known in the data science community partly because of a series of prominent victories in data mining and machine learning competitions. We have seen some of this in chapter three on skills and their applications. At the same time, competitions are, as we have also seen, sometimes discredited for their limited representation of the data science task areas. All these accounts have identified sources of uncertainty and ambiguity that data nerds encounter in practice.[78]

All accounts motivated their references to arcane knowledge and its contributors to their practical implementations. This suggests that the ideas might not be sufficiently universal as to guide data science practice unless one engages directly with them. Formalized rules spread more quickly and widely. I turn to them next.

[78] We could also rule out a lack of data nerds' awareness of arcane solutions for these problems, as academics struggled in these areas as well.

8.2.2 Terms

Data nerds seek guidance for practical challenges from experts with more experience or specialization. These references integrate diverse applied skills which we have seen consistently though unrelated before. For other problems, some rules are in place as well.

Here is one instance, as Kirill sees it:

> And, ahm you know, frankly, one of the issues I think with any new data technology, and we're talking about big data, is one of the benefits of relational databases, fortunately or unfortunately, is the ubiquity of a language that everyone knows, and understands, and knows how to use. That's why if you look at some of the database products that Mike Stonebreaker created in his life, they were always extensions of SQL, of this standard dialect, and he would add time series functionality to it, or he did geospatial. Because actually there's a way of thinking or there is an infrastructure for doing data and queries.
>
> So, with NoSQL databases, with Hadoop, Pig and Hive and things like that, ah, there's sort of a new language and a new paradigm of thinking about things.

Michael Stonebreaker is an academic computer scientist, thus adding another type of background to the series of academics considered in accounts so far. Unlike in the previous instance of the programming philosopher, formal status and technical utility overlap this time. This insider tries to show through the case of databases how programming languages matter. This note recalls some of the patterns we have seen emerge systematically throughout chapter three with the consensus that a key skill for data nerds entails working proficiency in several languages, instead of expertise in a single one. Kirill also picks up on the core question chapter six, on work, left us with. There we were unable to discern how the range of specific languages constitute an integrated stock of knowledge. Kirill describes this as a "paradigm" and cites Mike Stonebreaker as a reference for such systematic relationships.

Kirill's account also entails an inconsistency that once again reveals an important point. He draws a direct connection between Stonebreaker's efforts and the collection of tools associated with more recent developments in data storage and analysis. In other words, while the idea of a coherent paradigm may apply to both, this account should not be read to suggest that data science simply relabels an older effort.

Consistent with this sense of diversity, Kirill goes on. The heterogeneity of backgrounds not only emerges when we consider the accounts from previous chapters. We can find it here, within single accounts, as the continuation of the previous database comment indicates:

> But let's be careful and let's not get too excited about data for data sake, and technology for technology sake needs stuff. Ahm, I love Edward Tufte, I loved his graphs, but his graphs are there, are done that way for reasons so you can look at it and make a decision very quickly. Look at a vast amount of information and actually do an analysis, see something. And that's what we need to be driving towards.

Edward Tufte, a well-known academic statistician, has specialized on visualizing statistical data and models. He appears in this account not directly as an inspiration for others in the sense that his recognition for visualizations should universally signal their relevance. Instead, this speaker emphatically reminds the audience of the careful considerations that undergird those visualizes, but can be easily overlooked. Tufte functions as a reference to a rule, not an authority for, say, the importance of visualizations. The rule here remains implicit at best, but recalling asking oneself whether a visual feature adds utility makes for a timely call, when web and digital publishing eradicate constraints of black-and-white prints with respect to color choices and animation. Invoking this idea with reference to a person connects it to a body of work with more specific and systematic ideas.

Moreover, this speaker moves from Stonebreaker to Tufte and thereby covers in the course of a single narrative the whole range, the whole stack, of what data scientists have defined as their tasks. He combines references of similar status by vastly different intellectual contributions all the while also emphasizing technical details of their ideas. Kirill thus presents the community with an array of different views and perspectives that would have been unlikely within institutional boundaries they have separately defined their status in. Although he exploits that status, this bridge across scientific fields offers a new angle for considering the uncertainty data science entails. These prominent voices offer some guidance. They have technical implications for problems data nerds face regularly.

Finding directions in more subtle problems is even more difficult, as we see in Hannah's warning

> So, very few computer scientists or engineers would consider developing models or tools for analyzing astronomy data, without involving astronomers. So why then is so many methods for analyzing social data developed without social scientists? I think in part it is because we have really strong intuitions about the social world. And in fact my colleague Duncan Watts' recent book, 'Everything is Obvious,' addresses this exact point; humans are really good at using intuition and at rationalizing and at narrativizing. Intuition is often wrong, and narratives are not historical fact. For example we all possess attitudes, stereotypes or other cognitive shortcuts that unconsciously influence our understanding and our actions and our decisions.

Watts plays a prominent role in NYC's data science community. He for instance participated in the public debate following Facebook's emotional manipulation experiment, appears frequently at events hosted by various institutions throughout the city, and is involved in teaching efforts as well. He enters this account as a reference through which we learn about the specific idea that when data nerds make observations in social data, their analyses have to take angles beyond the ones that seem plausible with respect to intuition. Hannah reminds of other areas for which we do not have an intuition and therefore, turn to experts immediately. Now, with Watts, we have an expert reference to tell data nerds to turn to

experts such that even qualitative ideas about the terms of data science work get encoded in formal references. At these events, data nerds thus consider references eclectically for all problems they might encounter or be concerned with, including reflexive ones.

Although we begin to see a pattern of connections of different problems through formal references, we need to acknowledge an elephant in the room. We have seen references to technology entrepreneurs, computer, and even social scientists. Where are the statisticians? It is central to data science yet it has been suspiciously absent so far. To be sure, in a side note Jeff Hammerbacher mentioned William Cleveland, Josh Wills acknowledged statistics as half of data science, and we have seen a few others in passing. Specifically because it is so central, data nerds recognize many canonical statisticians. I treat them separately in the following section. Meanwhile, what about the specific guidance it might offer data nerds?

In part this is because it is so obvious anyway. Statistics largely defines quantitative analysis. That needs no reminding. At the same time, we have seen more general remarks of the two different cultures of statistics and machine learning. We learned that they interpret very similar, if not identical, quantitative analyses in different ways.[79] There we saw misunderstandings and confusion among audience members and presenters. John, the speaker at that event, described the current state of research as part of his response. I have seen similar descriptions about Bayesian methods, for instance. In other words, statistics is salient to data nerds in much more specific terms than many other fields.

Because data science and statistics are so close, I take a different perspective here. Let us consider a view from statistics on data science. For this I briefly turn away from the New York City site to consider a perspective that was first presented at an academic conference at Princeton University and then discussed widely in online data science forums.[80] Here we see the prominent Stanford statistician David Donoho directly reacting to data science as it is promoted in a range of recent university programs to point out its close association with machine learning, and concludes:

> It is no exaggeration to say that the combination of a Predictive Modeling culture together with CTF [Common Task Framework, see below] is the 'secret sauce' of machine learning.

[79] Just take for instance type I and type II error, and precision and recall, or classifiers, and models. See James et al. (2013) for a comprehensive description.
[80] http://www.r-bloggers.com/50-years-of-data-science-by-david-donoho/

The CTF is credited to Marc Liberman, yet another academic scholar, and consists of some version of public training data of a prediction problem, competitors engaging in the prediction task and a test set against which a referee compares the predictions. It therefore describes in technical terms the competitions we have seen accounts of and references to repeatedly. We can recall, for instance, the case in which the White House commended efforts of non-specialists to improve the analysis of NASA satellite images. Part of this effort involved the provision of the satellite data as training data which these non-specialists analyzed and built statistical models for. They submitted these models to the website that hosted the analytical competition and which applied these models to a different portion of the same dataset. Depending on the performance of those different submissions, they assigned scores and ranked contributors based on them.[81] Some data nerds rejected the relevance of this framework because it does not require many of the tasks they find important in data science. So what is it that Donoho finds important about it?

Donoho references a technical description of this practice, the common task framework, and provides his own, non-technical description, the "secret sauce." This label goes to signal his main argument with respect to the CTF, that "the single idea from machine learning and data science that is most lacking attention in today's statistical training" and that the CTFs require significant skills in designing appropriate information technology environments that accommodate those competitions. If this is data science's advantage and statistics should adopt it, it would seem that Donoho envisions data science as the future of statistics.

Consistent with many of the community identification moments we have seen among data science nerds, Donoho points out that such adoption would entail novel training programs as well as specific programming languages and packages associated with data science. A central argument, however, remains that:

> Insightful statisticians have for at least 50 years been laying the groundwork for constructing that would-be entity [of data science] as an enlargement of traditional academic statistics. This would-be notion of Data Science is not the same as the Data Science being touted today, although there is significant overlap. The would-be notion responds to a different set of urgent trends—intellectual rather than commercial. Facing the intellectual trends needs many of the same skills as facing the commercial ones and seems just as likely to match future student training demand and future research funding trends.

[81] As I mentioned before, specialists provided the highest improvement over NASA's own initial efforts in the end. The non-specialists who had made significant improvements before as well nevertheless remained involved in a conversation over the underlying problem beyond the competition.

Statistics would of course not become data science. In Donoho's view, there is instead a "would-be" notion of data science. This alternative data science he envisions is consistent with some of the central ideas of the way it is defined today.[82] His focus on intellectual problems over commercial orientations constitutes the most striking difference to many of the accounts we have considered here. To be sure, this view still differs also from those applications we have learned about here that address non-commercial problems, such as public or civic services.

This comment mirrors those we have seen speakers propose in their efforts to defend data science against competing work with data. Those defenses mostly focused on the question of which tasks of the data analysis process should be considered as part of data science. Contrary, this argument pertains to the overall choice of problems, and not just specific practices. This implies a substantively more specialized structure in terms of organizational arrangements of data science work. The sciences are of course diverse as well in the problems they address. Yet, consensus in the literature is remarkably consistent in recognizing that individual work tends to specialize for esoteric reasons of knowledge construction (Stinchcombe 2001) and mundane reasons of securing academic careers (Foster, Rzhetsky, and Evans 2015). This pattern is noticeably interrupted with non-academic collaborations or incentives (Evans 2010, Foster, Rzhetsky, and Evans 2015), precisely the features of today's data science work that would be dropped from a "would-be" notion of data science. Where does such an opposing vision of data science leave us?

Many accounts have proposed their own directions for data science. Donoho differs from those other accounts. His view implies a "re-specialization," instead of the inclusive diversification we have seen so consistently. With these considerations we have hit the starkest contrast with respect to organizing work. Science relies on a much more developed institutional basis than any of the technology nerd roles we have considered otherwise. Those who leave it experience "recovery," and those who are in it defend its terms. Donoho dismisses the relevance of navigating the uncertainty those accounts brought forward:

> A broad collection of technical activities is not a science; it could simply be a trade such as cooking or a technical field such as geotechnical engineering. To be entitled to use the word 'science' we must have a continually evolving, evidence-based approach.

"Science" is about continuous specializations, not improvised diversity. It also means entitlement, in Donoho's view. At the same time, we have just seen above in Jeff Hammerbacher's account that the

[82] Donoho cites university programs and some popular definitions of data science as basis for his understanding of the field.

purpose of his relabeling was to undercut some of the connotations research has, which is of course part of science. Indeed, just one speaker reflected on the lack of a scientific basis of the fashion problem he addressed. Considering all these accounts that construct the data science community, we find no evidence that it neither requires nor desires such entitlement to legitimate its practices, although they worked with scientific techniques and methods whenever these seemed useful.

Entitlement implies status. We can recall again the status finding that those professionals who have much of it avoid impure problems and consider colleagues who treat them of lower status. Conversely, those who embrace impure problems give up regards from their peers but gain public status (Abbott 1981). Abandoning scientific "entitlement" is directly associated with gaining public status. How should we think about science in the data science context then? On one side, there is the clearly defined scientific method. On the other side, there is also the scientific institution. While the evidence for different scientific fields varies, we know that scientists vary their research activities depending on their orientation toward securing their career, or gaining broader recognition (Foster, Rzhetsky, and Evans 2015), or that evidence and results are associated with measurement technology and investment in it (Collins 1998). The interpretation of evidence therefore seems to at least partly result from a *dis*-continuous processes. Similar to how we were able to reconcile Jeff Hammerbacher's dismissal of research and endorsement of science on a cursory level, we can see features of what Donoho calls a "trade" when we consider science's specific makeup. The entitlement coming with it comes from its institutional status. Data science, from this perspective, could gain entitlement from its own institutionalization.[83]

Donoho grounds his criticism in ideas of his 'heroes,' statisticians such as John Tuckey, Leo Breiman and others. Despite his dismissal of data science's current form as a mere trade, its members seek guidance from some of the same heroes. In the next chapter I therefore consider how data science reconciles drawing on some of the same canonical references with interpreting its practices as an integrated stock of knowledge, embedded in a thought community, that may be unscientific.

[83] On a more technical note, the alternative to being a science is not just a trade to begin with. The conceptual framework motivating this study offers guidance. Both sides of the debate, the sociology of professions as well as of expertise, equally rest on the premise that work that is based on arcane knowledge divides into more nuanced practices than the residual category of "trades." Besides sciences, there are professions or expert groups such as lawyers and medical doctors, and specific groups within them. What Donoho thinks of as trades this framework considers as occupations. Dismissing data science as an intellectually motivated movement for its lack of scientific integration would be premature from this perspective as well.

8.3 Canon

We began with the premise that data nerds find guidance in disciplining processes, and indeed found principles which data nerds integrate their ideas and practices in disciplined knowledge, though without directly replicating the academic sciences, which we know disciplines from. The question with respect to their status as a distinct thought community remains in how far data scientists make those ideas their own. The way data nerds have found and introduced authoritative references has mapped closely on the practical problems we have seen reports on throughout the accounts so far. This has resulted in formalized representations of the specific analytical steps and processes that require much more description otherwise. Amongst some other remarks, the last account has also reminded us that some of the central ideas go back further than the modern data context.

Considering the way data nerds engage with those classics exposes a deeper challenge. Unlike the guiding ideas in the previous section, the classics did not write or act with today's data-abundant context in mind. Data scientists have to bridge this gap in order to identify common grounds and thereby reveal more of how they interpret their own activity through those classics.

8.3.1 Reinterpretation

We can continue with the directions Donoho pointed out at the end of the last section. From his view they would lead to a different direction for data science than the one we have seen descriptions of.

Here canonical statisticians are discussed in a data science context:

> Ah, so, I do think it's a good idea to go and study one thing really hard, really long, make a lot of mistakes in it, but, also, … if you look at the history of applied computational statistics, as we now call it machine learning, I think there is a long, under-discussed, in academia, thread, of the importance of working in complementary teams. … if you look at the history of the people who framed the intellectual foundation on which data science now sits, and that includes people like Leo Breiman, and John Tuckey, you know, these were people who were high fluent mathematical statisticians, and then went out and did dirty, dirty consulting, right.

This account interprets the classics Donoho considers his "heroes." Yet, the conclusions seem to differ from those Donoho had in mind. The account in the last section focused more on the technical aspects of data science work and based on them an academic orientation for data science applications. This claim here points toward a different direction, noting that academics consider too little the applied work previous generations did.

This argument goes beyond the specific and institutionally recognized methodological contributions. It considers the classical scholarship not just with respect to the technical work academia remembers them for and thereby recovers practical terms:

> So Tuckey spent all this time working for ETS, Educational Testing Services, as among other people, Breiman had a proper tenured position as a mathematical probabilist, wrote a beautiful book on mathematical probability at UCLA, and then just walked out, and just like walked the earth in Santa Monica, taking these crazy consulting gigs, and then he like gave us CART, and like, which begat, you know, he gave us random forest, I mean like all these beautiful, very applied ideas in data science, came from interacting with real, messy problems, in the type of collaboration …, right.

These are the careers of scholars who modern academics still recognize for their contributions. We recall those contributions today on the basis of formal reference to the work in which they articulated them. In this account we also learn about the practical situations of those scholars. They are lost in the formal references and citations. It turns out that important contributions to science are associated with practical and commercial problems.

If statistics already claims these classical scholars, albeit for slightly different reasons, where does that leave data science?

> And the thing that makes data science different from machine learning, is not just getting epsilon better predictive accuracy on learning cats faces from pictures, it's this thing where you interact with somebody from a different discipline. And this somebody from a different discipline has this hundred years of domain expertise, they don't necessarily speak calculus, they have an understanding of how some system works, and then, brhhh, they're being challenged by abundant data. And then you as a data scientist are working with them to reframe their problem as a machine learning task, interpret your machine learning task in such a way that speaks to their language.

Here we see an argument that distinguishes data science from machine learning. It also once again maps onto one of Donoho's arguments. "Getting epsilon better" refers to the performance of models, for example as part of a competition, or "common task framework" application. In this account, that is not a central aim of data science. Instead, we learn here how data science involves the work of translating practical problems in quantitative analyses. As Hannah pointed out above, data nerds' own intuition would be biased by their personal experience. Data science's main promise, in this view, unfolds where there is no CTF in place.

Yet, this way of invoking classic scholars could still fit into a model of a "trade" that addresses idiosyncratic problems of clients, and thus training that is less defined by a community and more through the specificities of those interactions. Even if these interactions turn out to be significant for data science work, however, the methods nerds apply in these tasks are general and hence, accessible to a community of practitioners connected in spite of idiosyncratic experiences. The data might be proprietary,

to be sure, but that has not much bearing on the methods to analyze them. Formal reference to those classics misses this transferability, which, however, denies the trade-interpretations much basis. On the contrary, the accounts we have considered until now from these events capture it time and again.

This community identification process builds on accommodating and acknowledging varying interpretations of statistical ideas. In data science even such deviance can be legitimized with reference to classical ideas where not context is missing in formal references, but formal references have remained too selective, as this data scientist argues:

> In 1936, Sir R. A. Fisher, one of the titans of modern statistics, in a paper described almost exactly this technique [of computationally simulating statistical distributions]. He notes that 'the statistician does not carry out this very simple and very tedious process,' because he didn't have a computer in 1936, he would have, '… but his conclusions,' he [Fisher] goes on to say, 'have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method.'
>
> So Fisher himself, right, one of the greatest statisticians of all time, has given you his blessing to go off and write random permutation test, which is the name of that simple computational method I just walked you through.

Nothing is sacred. In addition to taking into account the context of work of statistic heroes, this data nerd proposes a purer reading of the original framework as well. This reinterpretation of classical scholarship ends a presentation that critiques the degree of mathematical sophistication necessary to compare two simple distributions to conclude that:

> … to do the statistics, that we just did, you needed three essential things: The ability to follow a straightforward logical argument, random number generation, and iteration. You were born with the first of these three things. And the last two are provided by any programming language with a decent library.

Here the data science community legitimizes the use of computational 'hacks,' simple enough to perform live on stage, as a replacement of mathematical procedures. They still use those procedures often as well, but the specific argument here is nonetheless revealing. Whereas before the emphasis was on translating practical problems into quantitative frameworks, the focus has here shifted to the question of quantitative rigor. Classical scholarship offers modern data nerds comfort to use their intuition, rather than the complicated systems of rules they had to devise in order to compensate for a lack of computational power.

To see how these two ideas come together, we can recall Aaron Swartz from among our familiar tech nerds. From an organizational perspective Swartz represents the process of contact improvisation. The previous account invokes contact with its call to listen carefully to others. This one facilitates improvisation, but circumvents the key components of quantitative data analysis. As part of this nerds rely

on programming libraries, as we have already seen them taking advantage of the work of Linus Torvalds. With these canonical references, we can also recognize a systematic basis to the otherwise sporadic contacts. They define a body of shared knowledge that, while not formally defined, provides a level of abstraction otherwise associated with the learned professions Mills saw disappear.

But what is "shared?"

To be sure, references to classical scholars do not always reach this level of depth:

> The methods that we have today around data science, the tools that we have today around data science are not necessarily the tools that are gonna deliver the miraculous outcome, this total big data revolution, okay. … there's hope, there are miraculous methods. So, one is probabilistic programming, okay. What the heck is that? … rather than tell you, you know explain to you in depth what probabilistic programming is, I'll appeal to authority. So the first authority of course is Thomas Bayes, and the second authority is John von Neumann. And so probabilistic programming is the marrying of computer science and Bayesian statistics. And it is something that is in the research world, you should Google it, if you're interested in doing formalized statistical modeling.

Both Thomas Bayes and John von Neumann were important scholars who are still relevant today. In this account they serve as "authoritative" references for the type of research and scientific contributions their ideas initiated. It is once again a formal interpretation, and one that does not provide additional specificity. None of the previous accounts discussed references with any scientific depth. They did point out the key idea they took away from there. They thereby established a connection between concrete ideas and abstract references. This is missing here.

This might seem unnecessary at this point where the aim is to recognize that:

> … the current statistical methods and tools like R, it's sort of a grab bag of tricks, right. And it doesn't give you a recipe to tackle a new set of problems, to model a new domain. And what probabilistic programming does, is it allows you to say, if you can model a problem, and by model I mean you can simulate a problem, we can do a whole bunch of magic ….

And we are back to magic. The summary here is superficial from the perspective of anyone who has at least heard of the ideas before in a more technical setting. Perhaps anticipating such reactions reflecting this absence, the speaker notes that he had considered providing a more technical account. This audience was broader and with a few hundred participants relatively large. No one asked for more detail. Indeed, other audiences discuss the most recent software packages developed by academic statisticians that implement Bayesian ideas, just like many other methods, with significantly more detail and technicality. Rather than indicating superficial readings of established scholarship, this presentation signals the practice among even applied data nerds to invoke classic ideas where the audience couldn't

care less. Despite all applied activities, academic “myth and ceremony” (Meyer and Rowan 1977) prevails as well.

Across these accounts we see nuances of the community identification mechanism of sharing ancestors. The focus differs though from the more familiar practice in the academic institutions, even for the same ancestors. Data nerds focus on the technical ideas as much as on the context their creators had them in. We have also seen the argument to focus on the analytical ideas rather than the technical interpretations that followed from them. Some, to be sure, rely on classics as formal references. With this redefinition, data science might appear perhaps not as a relabeling of existing disciplines, but not much more than a modern version of it either. What is distinct?

8.3.2 Redefinition

All the scholars considered so far form a relatively homogeneous canon and are also prominent in academic statistics, even if read there differently. These accounts have interpreted their contributions more broadly to include the social context around the development of the more abstract ideas they are being remembered for. Yet, at least in terms of strict intellectual lineage this choice still positions data science relatively directly as a descendant of academic statistics. But data science is more inclusive, albeit not always equally comprehensive:

> … what I’m going to show you today is an actual demonstration that looks at a problem that is more than two thousand years old. And the program required the greatest mind of its generation, a guy named Hero, who was a professor at Alexandria, to solve, without big data tools; in fact, he used just basic algebra.

The audience meets the most classical scholar so far, by age of ideas at least. His role here is not so clear, however. The problem is not two-thousand years old, of course, because Hero solved it then. The solution is two-thousand years old. That does not seem like a hard baseline.

Yet, the presentation goes on:

> Hero, the, our technology represents a shift, because today, for the first time in public, we’re going to show you that the tool can automatically derive a formula that it took Hero several years to build and literally his creative genius, to figure out how to do. And now our software does this automatically from data, in a few seconds.

Thus, not all references to old ideas work to the effect that they formalize practices. We are told in this demonstration about an algorithm that is as good as a single mind two thousand years ago. If we consider the intellectual work that has gone into the technology that allows for this improvement today, it is doubtful that there is any left in the analytical work. In this presentation data science seems to support

or justify a reputation as a grab bag of tools by suggesting a grab bag of classical references. Because there have been many smart ideas over the last few thousand years, just that connection makes for slim and undifferentiated formalities. We seem to have another myth and ceremony instance.

Yet, this still shows the difference between usage of classic before and here. The previous accounts emphasized the practice of problem solving, this is about having a reference that sounds big, and they deliberately take a different approach. Here is no experience or practice, just strategy. Tellingly, this project aims for replacing data science, not establishing it.

How would a more productive redefinition look like? We have seen before that it is easy enough to invoke authorities without describing their ideas in terms of practical relevance. We see it here again for heroes unfamiliar otherwise. A more productive approach integrates representations of activities contributing to data science's distinctiveness.

Haile presents another unorthodox reference, one that is less ambitious and only takes the audience back to the captain of Charles Darwin's explorations. That captain subsequently turned to weather forecasting, and in that context:

> … decided that it was as particularly important despite inaccuracy to present predictions for weather. Ahm, in about, so in the 1850s, Admiral FitzRoy was particularly concerned about how storms essentially decimated ships and the men manning them. And he decided that, ah, and it's fascinating to me that he decided to take this leap, but he decided that the time was ripe to start displaying and, ah, pronouncing predictions for storms. Ah, he actually, … in 1861, ah, presented the first meteorological report in The Times. Ahm, it involved a series of basic meteorological readings in a variety of cities, in the UK.

This account invokes a similarly unconventional reference in quantitative fields compared to the previous one. It differs in all other ways. Instead of presenting FitzRoy's achievements as a bar they improve on, this one emphasizes the identification with the intense practice of putting together data of relevant problems and making them useful. The experience is a central component, as we can see in the following comment to FitzRoy's suicide upon the dismissal of his reports:

> Actually, if you've found yourself making a prediction, that, ah, for which your business, sort of rests very heavily, and if you've made the wrong prediction, ah, you probably know what he felt like, even if you didn't do what he did [commit suicide].

Just as we became comfortable with our situation in nineteenth century England, we find ourselves in modern predictive modeling. How did that happen?

Haile focuses on the emotional experience of data analysis. For that argument it is irrelevant that the methods FitzRoy had available were so different as to undermine any specific connection to modern quantitative analysis. We have seen similar remarks across several accounts when speakers mentioned

"blood, sweat and tears" in the context of preparing and structuring data, or just sweat amid the thought about potential harm resulting from unintended consequences because of the context of an application. The central point here reiterates observations anthropologist have made among modern hackers (Coleman 2013, Weber 2004).[84] In a mostly formalized world of problems, it is not easy to express such emotional aspects.

This historical reference encodes feelings that we have seen throughout the community in a level of abstraction. Emotional experiences thereby loosen their specific relationship to a project. This speaker clearly oversteps the convention of classical references, not only with the specific individual but also the purpose for which he cites him. The activity of formalizing rational ideas and observations, which classics usually do, we see this presenter repurpose for formalizing feelings. This formalization allows a community to share and identify with otherwise deeply personal experiences.

Canon overview

Data nerds draw on canonical ideas in two different ways. The kind of references last seen constitute an incoherent extension of the elegant statistical canon. We saw a different use of classical ideas where they served as benchmark instead of guidance. It is not clear how this would help data science integrate distinctive expertise. The possibility of such integration became clearer in another instance that described early efforts quantifying weather forecasts and the emotional effort associated with it. Although the technicalities have changed, the emotional involvement has not. Yet, it is rarely considered part of the process. This extends the set of canonical references of other quantitative fields.

We also saw nerds apparently misreading original contributions. They soiled the purity of the statistical ideas with references to the circumstances in which they were had, including to economic and other practical problems. This reading could be interpreted to reinforce the view of a 'collection of technical activities' that might be associated with any 'trade,' to recall the suggestion seen before. At the same time, the way the emphasis is placed here reveals a concern that gets easily overshadowed by the more formal role of Tuckey, Breiman and the other statisticians. Equally important as their specific ideas are, for data scientists, the practices of turning empirical problems in formal representations through direct

[84] It is a sad fact that Aaron Swartz, the hacker through whose type of work we have relied on to understand the organizational process of contact improvisation, committed suicide as well.

observation. As we saw as well, formal references can stand for the systems of rules designed around early ideas, as well as the early ideas themselves. They can likewise stand for the substantive context of those ideas.

Together these ways or reinterpreting older ideas and redefining the set of relevant ideas on the basis of formal references creates an integrated stock of data science expertise.

Chapter overview

This chapter has revealed further mechanisms of thought community identification, moving from technical and specific skills, tasks and career paths, toward the collective and systematic interpretation of those experiences. We have seen data nerds circumventing existing formalism, thereby enriching their meaning, and formalize substantive experiences. This way they integrate otherwise distinct activities. Recognizing these processes adds to our understanding of data science as a distinct and robust expert role. The data nerd community builds a basis not despite but because of the variation in those references. The abstractions allow the nerds to integrate practical and arcane ideas in a way that remains systematic across different applications and for each other. Data nerds redefine and reinterpret purely technical ideas such that they become formalized again albeit entailing richer substance. Paradoxically, the pristine definitions, proposed by those most removed from the messiness of applied problems, gain purchase once their purity is ignored. The abstract references thereby rationalize data science practices independent of varying substantive applications.

They still end up with specific heroes they consider theirs. A subtle feature emerges once we consider that others claim these heroes as well. Data science nerds see things in them others do not find relevant. Through these heroes it becomes possible to define relevant knowledge independent of concrete applications. Similarly, because no one can claim ownership of these heroes, data science is free to assemble its own set of characters with their respective specializations that would not fit together from the perspective of other groups. It is still possible to reason over the principles, as we saw in attempts to articulate them on course curricula and in informal groups and the blog posts they produced. Just like references to individual ideas, those to statistics and software engineering point toward concrete rules for conduct without imposing definite steps. This way data science establishes anonymity with respect to its role. Mills's view seems directly addressed.

These coordination principles have implications for our understanding of the problems data science confronts us with. We have found here a model for integrating expert knowledge. Whereas the previous settings have rejected the relevance of organizations with offering just incomplete accounts of alternative coordination principles, here we have direct indication of how data nerds integrate their breaks with traditional ways of dividing and combining work. This supports previous conclusions that the concerns the public has brought to data problems cannot be understood through the organizations through which they have unfolded alone. By now we have seen different contours of an in many ways distinct thought community, and which jointly map out its coordination principles that hold them together. With the abstract integration of different specializations from this chapter, we need to consider interpreting the consequences of this expertise with an eye on how we engage with other expert groups. I outline more practical and specific implications of this argument below and in the conclusion.

Meanwhile, this has implications for individual opportunities as well. Whereas the previous chapter has already suggested that organizational careers do not directly lead to data nerd roles, here we have indication that the steps they take entail complementary characteristics nonetheless. The struggle between arcane training and knowledge and concrete applications, which we have seen before, nerds here rationalize as basis for their distinct expertise. Data science individual careers and public consequences constitute instances of abstract patterns that are not salient on the level of organizational or other institutionalized forms of coordinating, but do find a very explicit set of “heroes” as references and for guidance. Therefore models of leveraging or addressing them cannot rely on those familiar indicators.

What does this make data science? Data nerds take ideas other groups consider theirs, interpret them differently, and combine them with yet other ideas those original groups did not take into account at all. In order to consider this question in detail, I next focus on the science in data science.

Contours: Performance

We have seen here data nerds conceptualizing data science on the basis of their synthesis of work they do and observe and of interpretations of their role models, or heroes. One way to read these definitions would focus on their substantive and historical accuracy. With some of the classics data scientists have invoked being familiar in other disciplines, which may have a more accurate recollection of their specific contributions, invites to pursue this direction. At the same time, it was not evident that data

science itself interprets those classics as definite sources of directions for their practice.[85] This could be seen early on in the chapter as a speaker simultaneously claimed founder status while also acknowledging the legacy of established scientists. Taken together, we need to untangle how data scientists, whose work, as we know well by now, is similar to academic disciplines with respect to the arcane knowledge they use, take scholars who are significant in those disciplines and interpret them such that they suit their own cause. We observe in these accounts data nerds performing science, the metaphor that defines their group, and thereby create, together with the illustration of relevant problems and technology from the first chapter, a frame for the persuasion, improvisation and intimacy, which we have discovered in between.[86]

The significance of performance may seem self-evident. All we have seen even before this chapter were data nerds performing on stage. These kinds of presentations, on practical steps and considerations in their daily work, are common in many professional contexts. Some of those groups are salient, but many are not. Focusing on just this aspect therefore offers little analytical leverage for understanding data science. Throughout the accounts of this chapter specifically, data nerds describe activities of others in order to distill predications for their own work. I suggest that this practice resembles conventions in the sciences but that the way data nerds interpret them lead to different implications for the kind of thought community they form. In order to specify this mechanism as a basis for yet another contour, I first outline, in basic terms, the templates other sciences provide. Next I suggest an alternative structure, which follows from a different interpretation of scientific principles. This alternative structure resembles more closely the patterns we can observe among data nerds.

Sciences continuously build on and develop ideas. The reviews in the introduction have revealed three main positions that offer competing explanations of the development of scientific knowledge. The role of classic ideas remained relevant across otherwise different views. To recall, the formalist view emphasizes that the classics define the ideas for further scholarship and research. As the expertise view emphasizes informal interactions, classic scholarship would fit the framework but play a less central role. Lastly, a functionalist view that in this respect overlaps with work on thought communities points out the

[85] We can also once again recall the evidence I have cited above that points out how those disciplines construct memories.
[86] Performance has several meanings. In one sense, performance describes the successful accomplishment of a task. The main part of the chapter discussed this aspect, and its relevance for data science. The following discussion focuses on a different meaning of performance, that of rendering a role.

role of defining the scope of a scientific community, if a somewhat arbitrary one. Although the three views differ in the effect and significance they attribute to classics in the science, they all consider the role of classics with respect to their ideas for future work as all three focus on continuously evolving groups. More specifically, we can take the formalist view, for instance, which holds that scientists look at classic work in order to identify relevant questions (Stinchcombe 2001), or the informal expertise view, in which we would also expect classic work to emerge from the close collaborations where new group members join whereas others may move on (Collins 1998). On the level on which they agree on scientific knowledge, that it continuously develops some classic ideas, these models are nominally consistent with data nerds.

Moreover, data nerds do not contradict interpretations other sciences offer of the classics they have in common. Indeed, data science takes the kind of arcane ideas and abstract guidance, familiar in academic disciplines, and applies them to the practical problems it encounters. At the same time, citing emails to the effect of explicitly denouncing “research” components of data science work, speaking in public of research conspiracies and invoking blog posts for central definitions departs in many ways from the formal channels of scientific discipline, and in parts from the informal view as none of these channels gains significant qualitative depth that arcane ideas often require. This has the consequence that defining its expertise through abstract references, even if historically inaccurate, provides data science with a way that integrates the practice of improvising into a systematic framework and specific direction.

Just take the common reference to Leo Breiman. We saw a statistician naming him as his hero for contributions to quantitative knowledge. The data nerd acknowledged this as well, but also emphasized the life, outside of academia, Breiman conceived of them in. We learned less about the lives of Tufte and Stonebreaker. Instead of limiting our focus to single ideas they had, however, we again learned about the consistency of their specific focus over many ideas. There was also Captain FitzRoy, whom we might not have considered at all if it was not for the context around his specific ideas. Then, however, we also encounter the ancient Greek “Hero,” who did not seem to connect to the others as he did not provide an idea, just a baseline.

The scientific discipline thus offers just incomplete explanations for how data nerds pick their heroes. How can we understand data science’s interpretation of classical scholarship instead? Research

has found similar behavior in more common contexts of work as well as in different contexts as religious communities. John Levi Martin's (2000) analysis of children's literature and the way animals it describes perform tasks associated with different occupations leads Martin to argue that social interpretations of the class structure and division of labor are reproduced for children through animals. He demonstrates that this unfolds on the basis of their bodily characteristics such that animals with certain bodies do certain jobs, and not others. Whereas fiction writing provides the freedom to ignore that no actual pig works as a ditch digger, no fox as a state official, and no dog as delivery person, as children's literature imagines (Martin 2000), principles of scientific discipline ensure that data nerds preserve substantial relevance between the classical role models data scientists choose, and the work they choose them for. At the same time, neither classics nor animals have a saying in the way narrators describe what tasks they perform. In other words, the Western society thought community arranges abstract animal species rather than specific people. Because animal species are natural, it thereby naturalizes the division of labor. Data science not so much naturalizes, but in this setting rather institutionalizes its practices through abstract references as well.

The fiction author whose presentation Martin analyzed takes animals because he thinks children can relate to them better. How do we get from the mental images Martin describes with which children (and later adults) interpret the social world, to the practices with which data nerds solve problems?

Data science has a clear basis in the sciences it takes ideas from. In order to understand contours of thought communities, we need to connect their interpretation of ideas to collective behavior. Anthropologist James Fernandez (1972) takes this additional step as he analyzes religious communities in Africa. Fernandez argues that descriptions of religious assertions have performative effects, citing instances of understanding a pastor as a "bull who maintains order in the cattle kraal" to lead to the pastor engaging with the community, and the community being disciplined, or that ideas of being "the voices of God" lead to studying the bible and listening to sermons (Fernandez 1972, 55). Translating this finding to the data science context requires that we consider its own main metaphor. As we can see in its title, and from the practices above, this is the sciences. Fernandez's argument suggests for this context that although sciences are commonly understood as developing knowledge, here they offer a metaphor for principles the performance of which informs behavior. Focusing on the performance, more than on the

content (although data science does that too, only in a more specific way), data science creates a new array of references. Somewhat ironically, a similar performance of a misinterpretation of American graduate students returning from stays abroad in then leading German universities shaped graduate school, a fundamental component of the university system in the US (Ben-David 1971, 139). The basis is performance of ritualistic interactions of science, except in a different way. Science offers a metaphor that comes with the limitations they often have, that they are imperfect. It nevertheless leads to behavior that is consistent throughout the thought community.

Importantly, as we have just seen above, circumventing academic channels does not undercut all discipline, seen as members of the community debate their ideas with each other as well as in the context of their origins, and not to forget their applied utility.

The kind of discipline data science undergoes when performing its definition at these events lacks the opportunities sciences offer in disciplining ideas and developing new ones. They neither offer the formality of many academic channels, nor the closeness of scholarly groups. These settings also forego their constraints. By emphasizing the process around defining data science as well as associating its abstract ideas with their widely recognized origins, we observe the definition of a coherent narrative, which in turn provides the grounds for a thought community that is better able to evaluate the quality of unorthodox or at least unfamiliar ideas. This rhetoric facilitates a more comprehensive discussion of analytical strategies. It also facilitates new ideas in the sense of practices and as a result of discussions not only over abstract knowledge but also concrete practices and applications, thus adding dynamism. And so here emerges not despite but because of the performance of a scientific model, a thought community that upsets scientific principles.

The data science community

New York City's data science community is real, vibrant and varied. It is so diverse in fact that a number of familiar explanations of control over tasks and problems provide only limited guidance for understanding how data nerds explain their work to the public. Data science practices and applications cross technological infrastructures and organizational boundaries. Their consistent appearance across specific projects gave little ground to explain its salience as an expert group that defines its work across these contexts. Accounts have shown consistency in their application of expertise and the challenges they encounter but no conclusive basis for it. Data nerds' strategies for addressing these challenges reflect informal activities to overcome them, such as some direct interaction with peers and other involved parties. We also encountered reports of friction between data science and other experts of data problems, which those organizational functions, primarily IT, expressed, however, whereas data science did not without others asking. As we saw also when data nerds explicitly defined the requirements for doing their work, they remain preoccupied interpreting the basis of their expertise. This array of observations reveals data science robustly embedded in the social structure of the technology industry, many other sectors that collect data, and in more arcane scientific roots. They also demonstrate, however, that none of these contexts clearly defines data science tasks, practices and expertise. We see that data nerds do that, but not clearly how.

We have then turned to consider the details of their skill set and the ways in which they define it in order to specify these processes with greater precision. This shift of perspective has led us from the concrete computer scripts data nerds design over revelations of personal trajectories and to the individual interpretations of complex problems. From these accounts we were able to describe further the level at which data scientists define their practices. We saw a community structure emerge from the interplay of formalized skills and definitions and informal practices and interpretations in the process of solving concrete problems of data representing different aspects of social life as well as technical systems. Computer languages and strategies for overcoming complications associated with data of varying sizes and for differing purposes connect applications across a series of areas data scientists all define as theirs. Other tasks also related to data may require more specialized experts and hence induce isolation.

It remained unclear, however, how these data nerds collectively define their integrated expertise on a continuous basis.

We finally focused on the experiences of community membership and discipline as a coordinating mechanism of groups that balance large size and minimal contact or formal control. Amid still varying backgrounds, we found consistent experiences of building up knowledge toward defining data sciences roles along otherwise largely ambiguous paths. Nerds differentiate these roles from other data tasks most consistently on the basis of transcending bureaucratic divisions toward more inclusive and interactive technological frameworks. They anchored this agreement with guidance of well-known technology nerds as well as classical academic ideas. Unlike academic fields, for this purpose they focused on the substantive experiences of these scholars who had them. Data scientists thereby redefined the substance formally encoded in references to them from abstract into practices. They thus integrate heterogeneous applications through this kind of abstraction.

These processes reveal a number of steps of defining arcane data science work and community identification mechanisms. What kind of community do they produce?

While the initial formal contexts gave data science some momentum shaping collaboration and inducing competition, the more subtle processes of solving unfamiliar problems by improvising and adapting existing skills and strategies were critical as well. In order to understand how a distinct group of experts reconciles these two levels we have moved on to considering the contours that emerge as data scientists define their work, challenges and developments. Descriptions of such processes have emerged since the beginning as accounts were illustrating technology in vivid metaphors and analogies and persuasion in the relations with competing functions in ways that put data science in strategically advantageous positions. The effort data science invests into defining a community emerged most clearly, however, through the intimacy in accounts of personal trajectories and in the public definition of intellectual lineages to heroes. These observations have revealed the uncertainty that the community jointly overcomes by interpreting the transitions they make to be associated with the practices they develop for reconciling practical problems with arcane tools and ideas.

8.4 Data science as an instance of role layering

Here I consider the different moments in New York City's data community from the perspective of the more familiar technology nerds. We find a sequence in which their respective ways of defining work layer on top of one another. The view we have associated with Bill Gates's bureaucratic task definition is being taken over by Linus Torvalds and the heterarchies his style of organizing technology work induces. Some more additions follow and Torvalds's integration of open systems makes room for C. Wright Mills's emphasis on salience as a feature of autonomous work on the basis of abstract knowledge. These transitions follow from the systematic presence of Aaron Swartz's style of contact improvisation that becomes increasingly recognizable in the accounts defining data science work as we move from technology and organizations to skills and formalized knowledge. In combination, these different characters have jointly constituted a distinct data nerd.

This layering unfolds across different moments in the construction of a data science thought community. We have considered the settings on the basis of their utility for capturing macro-level processes of contemporary data technologies and micro-level aspects of the skills constituting that work directly. Whereas skills have introduced us to more abstract ways of coordinating, we have considered the data nerd community and discipline as meso-level contexts that organize the distinctive skills, where we found most explanatory leverage. In terms of the organizational arrangements of familiar technology nerds, this design has shown that Gates is prominent at the beginning with respect to problems of technology and organizations. The middle plot, which centers on specific projects and skills, sees the momentary retreat of Gates into the background following project definitions, and increasing prominence of Linus Torvalds with the focus on skills. In the penultimate moment, around the community aspects, Gates regains a voice in the data nerd accounts that is just slightly more prominent than Mills's, at that point only to silence significantly, leaving most to Swartz and Mills in the finale, which focuses on discipline. That Gates is relevant in the context of formal organizations is of course little surprising. But there we found that the community defines its own contours amid organizational arrangements by devising arguments on the basis of the tactics with which they solve problems organizations encounter, largely ignoring existing organizational function. Similarly, as we considered data science discipline, we found less evidence of the sciences that have defined disciplinary ways of organizing knowledge in to

begin with, even though data nerds acknowledged the basis those sciences drew on and developed. The data nerd role thus emerges from layering of practices that pertain to the economic and scientific settings.

To be sure, Aaron Swartz's footloose strategy has a somewhat easier time defining work in an emergent area, and thus more easily shapes the modern data nerd thought community. After all, it takes much less coordinating around an erratic initiative than an ongoing community, which is part of all three other strategies. Yet, unlike Swartz, who became an icon for some and villain for others, data nerds emerge as an anonymous role that others recognize for the underlying practice itself. It instead takes the layering of these experiences on the level of careers, and the folding of careers into a robust role.

Against this background we can remove the cover of the historical figures and consider the more arcane processes they represent directly. To recall, we have viewed Bill Gates as the face for bureaucratically dominated occupations, Linus Torvalds and Aaron Swartz as those of improvising and heterarchical different instances of expertise movements in the technology context, and C. Wright Mills as representative for the anonymous professions whose demise his account focuses on.

This technical level helps to generalize the processes of knowledge construction we observe. Most significantly, against sharp opposition between informal expertise groups and formal professions, here we have seen a layering of both types of processes across different settings, where occupational principles have left their mark as well. On the one hand, they have acknowledged the formal organizations whose problems they addressed, as long as they did not define their work. This significance has also been marked, however, by moments of irrelevance on the level at which nerds do the work. At the same time, we have continuously seen a significant role of some informal definition of work in which the data nerds abandoned formal definitions of tasks and competencies in the interest of the problems they found relevant, or the tools they found useful to apply. Here informal and coordinated expertise groups, centered on open technologies as well as substantive conversations on what a problem might be and how to solve it with data science, have driven these movements away from bureaucratic task definitions. In other instances, the interactions were more fleeting, as those between speakers and audiences at these events have indicated, the casual references to other peers, communication through blog posts, and so on. Although limited to relatively brief contact, these informal ways of defining their thought community left formal traces. On a practical level, informal communication on novel solutions to shared

problems manifested itself in formal trace in code that might have been shared in this process, or even the mathematical ideas underlying it. The abstract knowledge of professions has also gained prominence particularly by means of articulating data science tasks in the language of those who have preceded them. This way data nerds have integrated otherwise distinct problems in a more abstract way.

In short, data science nerds do not come out of nowhere. Their relation to statistics and computer science was clear since the introduction, and never seriously in doubt. Here we see that also their organizational arrangements exist in parts elsewhere. This is not to say that the data nerd should not be considered as a novel role. Data scientists combine technology roles previously seen in radical opposition to one another, their community also integrates processes the literature has previously identified in distinct contexts, and seen as part of opposing explanations for expert work. This kind of process can be conceived of as an instance of what political scientist Kathleen Thelen has described a similar process with “institutional layering” (Thelen 2003). The combination for data nerds leads to resemble a rare structure by which they gain distinct salience, potentially the beginning of further institutionalization (Berger and Luckman 1966). Capturing data science in this emergent status reveals the importance of both informal interactions and interpretations of existing institutional, technological and organizational arrangements, and the coupling with the process of utilizing these informal processes for formalizing concrete substance as a basis for an abstract stock of knowledge.

8.4.1 Just so …

With the basic results summarized, here would be the place to return to the troubling incidents in which data science work entered with private life, as well as to a summary of appropriate steps for those seeking to pursue data science work. These conclusions rest on the promise of data science’s salience independent of other forms of control and coordination. We therefore need to understand with certainty the basis of data science’s salience. From this analysis, we could infer that this has to do with data nerds undermining formal boundaries through taking advantage of technology and knowledge that is openly shared. For this purpose it was important that data nerds improvise in ways that lead to novel applications of that knowledge, regardless of the formal boundaries, whilst acknowledging the practices as a group. But there are problems with interpreting this analysis in that way. One question that arises has to do with modern technologies. Although we saw that they only begin to matter once data nerds interpret them, it is

not clear that the later argument of interpreting them in a structure that integrates different problems is necessary here as well. If this is the case, the contours we have recovered would still hold, but the argument that data science has gained salience because of distinctive principles of integrating knowledge that deviate from bureaucratic and scientific principles would not.

I have argued here that data science has gained salience through the effort of a community that applies arcane knowledge of advanced computational and analytical methods to heterogeneous and even mundane problems. This community is tied by and coordinates through the informal narratives around the implementation of formal tools and expertise. Ignoring the specific accounts, I have analyzed throughout these chapters, another argument could fit the historical sequence of events. First there was quantification and statistics. Statistics remained tied to the substantive contexts it was respectively concerned with, with little movement across. Then computational power improved and machine learning emerged. This development diversified the approaches possible in quantitative data analysis, but did not overcome the isolation across problems. And finally data collection became more universal and its storage cheaper and hence more feasible. This change could finally give rise to data science.

In that story everything was laid out for this generation of tech nerds to become the sexy rock stars the public is now recognizing them as. The technology was ready, the data came in, and without realizing it, they had the skills and so it was just natural for them to do this work.

Here the interaction of formal and informal knowledge and expertise are unimportant. It also ignores, however, that the advancement of data and technologies itself would not require data science to build appropriate applications. The work data science claims to do today could have been allocated to existing departments, such as IT, or engineering in larger or more technologically oriented organizations. Some of the accounts we have considered throughout these chapters have explicitly pointed into this direction. Instead, data science has formed around defining problems themselves. They have done so in organizations with functions that could be extended to also include data science tasks and in those without relevant functions. This effort can be best seen in the large pro-bono community of data scientists offering their services for free and during their free time.

Now from this perspective it is not so natural for data science to take over in this way. Data scientists find it surprising themselves how much attention they receive today for things they have done

for years. For sure, it is more likely to build a community with more experts around, which again results in a larger group. It is not obvious, as the following instance of a group shows, which tried to address a very similar problem but is much less salient today.

Let us consider this specific instance to better see the problems with the initial inference. We have seen Yann LeCun recount some difficult times the idea of deep learning had undergone before spreading widely in the last few years. By that time, it had begun accelerating the artificial intelligence movement with improving capabilities of interpreting images, recognizing speech, and so on. The idea of artificial intelligence is much older, and considering a small part of it from the 1970s, long before it had gained the prominence it has today, reveals important details entailed in data science's effort to generate formal abstractions of concrete, messy problems in the world.

One early pioneer in the movement was Edward Feigenbaum (Feigenbaum and McCorduck 1983). A central idea in his interpretation of artificial intelligence centered around the role of a "knowledge engineer," who would "know how to represent knowledge in a computer. They know how to create reasoning programs to utilize knowledge. And they are interdisciplinary in spirit" (Feigenbaum and McCorduck 1983, 77). It is easy to see that this role maps onto some of the task descriptions put together for data science in the last few years. Unlike the data scientists of today, the knowledge engineer had to rely on a process in which "individual computer scientists work with individual experts to explicate the experts' heuristics---to mine those jewels of knowledge out of their heads one by one" (Feigenbaum and McCorduck 1983, 79), requiring that the knowledge engineer "be able to put herself so carefully and accurately into the mind of the expert with whom she is dealing that eventually she can mimic his thought patterns with great precision. There lies her generality" (Feigenbaum and McCorduck 1983, 84).

There is great rhetorical similarity between descriptions of the data scientist of today and the knowledge engineer over thirty years ago. Today's data scientist, however, is widely recognized, while the profession of knowledge engineers is not defined on any noticeable level. The specifics are important, as they reveal fundamental differences in how each group defines their approach. Knowledge engineers were tasked with translating expert knowledge into "mimic thought patterns" of experts. Data scientists dealing with experts today are equally expected to engage with deep knowledge, as a basis for machine learning tasks, but not to translate their thought patterns directly.

This group's fate directly indicates that technological advancement does not suffice amid a singular definition of the group responsible for applying it. This gives some confidence that data science's formation is not just following from the technological advancements, and not even from the single idea to rename a young department. It also raises questions though about what constitutes this fine difference between groups that receive recognition and those that don't. This is what I turn to in part two.

II. Data science in economic and academic settings

Introduction to Part II

We have arrived here after considering many rich perspectives of data nerds defining their work at public events in New York City. While these perspective have contributed to our understanding of data science work and the basis of a distinct thought community of data nerds, they have also raised more specific questions along the way. They have shown that we can largely rule out the dominance of both technology and formal organizations over data science work, although both shape its definition at least to some degree. Yet, our understanding of the bearing of the abstract integration of distinctive data science contours and their salience, which we found instead, has remained tentative so far.

Throughout the previous chapters, I considered definitions of data science work, tasks and skills relative to those of older and more familiar technology nerds in order to gain analytical leverage amid data science's emergent status. Among the kinds of work Bill Gates, Linus Torvalds, Aaron Swartz and, as a placeholder for an otherwise anonymous role, sociologist C. Wright Mills, represented, we were able to recover some of the patterns Mills had thought were lost to perpetually strengthening bureaucracies. Data nerds acknowledge formal definitions of their work that come with Bill Gates's proprietary definition of tasks, and they relied on open source software with its heterarchical arrangements of the kind Linus Torvalds implemented. They did so, however, in a way that resembled more the contact improvisation we have associated with Aaron Swartz. This layering of different roles has offered a basis for autonomy and salience. To be sure, they don't have the institutional status of Mills's learned professions. The similarity follows from the routinization of relatively autonomous and anonymous practices. While these historical figures and organizational models they represent were useful for understanding arcane data science practices, they remain too vague as to reveal with sufficient certainty the distinctive principles underlying data science expertise and its salience.

In order to gain a more precise understanding of how it is that certain contours of arcane knowledge facilitate its public salience, while others do not, I propose a comparative approach. This entails a change of pace and of rhetoric in how I articulate the argument. Indeed, recalling the familiar tech nerds here would distract from encoding the basis of contours in formal steps that are necessary to render data science and other empirical cases comparable. It would also be unnecessary as they did what the models resulting from this formal procedure achieve here directly.

Without these familiar roles to hold on to and no practices that data nerds describe, it is much more difficult to see the relationship to the practical problems data science confronts us with. The gain in precision outbalances that loss. For instance, as data nerd told their audiences about their orientation for guidance toward other groups, organizations and nerd communities, but never mentioned managers, and instead shared their experience claiming responsibilities over data from other functions, we could directly see that if we have problems with the results of their work, managers are perhaps legally responsible, but effectively not those to address. At the same time, it remained unclear who to turn to instead. Even chapter eight, which eventually revealed most clearly that data nerds see their common basis in the solving practical problems associated with arcane data analysis, and articulated that basis through abstract references, we, as an affected public, were unable to derive ideas for how to interact with them. The chapters in this part are designed to lead to results that allow us to address this question. As part of this, the empirical chapters sacrifice the substantive richness in favor of analytical clarity. The two perspectives rejoin in the final conclusion.

Settings

The previous setting offered concrete activities, professional meetings with presenters, audience, Q&As, networking, and so on, that made it easier to follow the moments when data nerds invoked more arcane expertise. This is not possible in a comparative design that focuses on analyzing contours of expertise and of abstract stocks of knowledge. Just take law and medicine as the most canonical professions. While lawyers meet at conferences as well, the most salient activities unfold in courthouses. They have no equivalent in data science. Or take the radical contrary of informal interaction with lay patients organizing to gain access to institutionalized medical conversations. These settings neither resemble each other, nor the conversations we have observed data nerds define their relevance in for the public to an extent that would permit formal comparisons. The degree to which they share common features with familiar experts, like arcane knowledge, it would be unclear how the terms of defining it in one context compare to those of another. After all, partly overlapping with the established groups, we have seen similar processes unfold among data nerds, on an abstract level. Here we can recall their emphasis of collaborations with other organizational functions in order to specify data problems, and on the other hand how they invoked the mathematical formalism. It is not so clear how they relate to legal

formalism and medical advocacy. This goes to show that the empirical setting we have considered so far fails to resemble interactions of other experts should not imply that they are necessarily different, in principle. Instead, we need to find settings in which they emerge in more comparable terms.

In order to address the more specific questions I choose two settings in the economy and in academia. Each has features that make the respective substantive problem salient, contours of abstract knowledge from improvising, or disciplinary coordination and control. For the question of the autonomy of abstract knowledge in the otherwise bureaucratically dominated labor market, I turn to the job market. In order to understand the pertinence of disciplinary forms of coordination for data science, I turn to academic scholars in the university context and the patterns of their work, as the academic sciences have developed them most rigorously. I just consider each one briefly with respect to the ideas they develop, and leave the details to the respective chapters.

Chapter nine focuses on how skills fold together as others see them in order to analyze lay expectations of data nerds and other cases. We have seen in many instance data nerds describe that much of their work is concerned with assembling tools and process in a “tactical” way that differs from older, more specialized solutions. They often described this aspect of their work with respect to the examples they worked on, be it facilitating dates or careers, but instead of finding a systematic basis, the audiences were interested in specific experiences and data nerds themselves described hiring decisions often with references to examples of the work candidates do. On the other hand, we also learned about attempts to conceptualize this kind of work, even in course curricula and essay format. Here data nerds begin to resemble autonomous professions whose institutional status signals skills and competencies through degrees and certificates. As we saw evidence of both in the rich accounts of data nerds, once we turn to their broader salience we face the question of how others articulate those skills and requirements in a way that transcends the personal interaction such that we might understand through them how data science can gain salience as a distinct thought community. For this purpose, I consider job descriptions that target data nerds, as well as established occupations, and articulate what they mean by that. Job descriptions accommodate concise and relatively abstract titles as well as richer descriptions, as I explain in more detail below.

Chapter ten considers data science in the context of academic disciplines, identification with the sciences and even with specific scholars was common among data nerds as well. It remained unclear how such references would unfold on a routine basis, however. This focus brings us back to the meaning of "science" in the data science label, which has given reason for bewilderment since the initial Wikipedia definition of the field we considered. There we learned that although data science draws on existing academic fields, it just utilizes some of their contributions, but not others. Moreover, some of them originated in practical problems, whereas others reflected scientific motivations. As we began to consider nerds in New York City's data science scene, science was less salient. Only when we explicitly turned to discipline as a possible governing framework, in chapter eight, did we see that this was not because sciences were not important, it was because they were obvious. At the same time, observed that data nerds offering more inclusive interpretations of scientific ideas than academic scholars typically do. These more inclusive interpretations signaled an alternative way of identifying with a distinct thought community, one that emphasized improvisation. They did not clearly delineate how this different identification mechanism facilitates greater lay salience. I therefore compare such referencing patters among scholars affiliated with data science to those of other fields as a basis for specifying their differences.

Analytical design

The two settings are so specific that they require their own methodological designs, which I introduce in two separate chapters respectively. They nonetheless also draw on some similar ideas and operationalization, which I introduce briefly here. For this purpose I return to the "emergent categorization" framework from chapter three. Instead of simply repeating the general idea here, I describe it in terms of the questions part one has guided us to.

One important analytical idea in part one emerged from the empirical design. Each chapter first considered the accounts of data nerds, how they described their technology, work, skills and so on. As we considered concrete quotes, our focus was not only on what speakers said. We also considered how they said it. At the end of each chapter, I recovered the contours of their expertise by considering the rhetorical principles underlying these accounts. In chapter four, the first of this study of New York City's data events, for instance, we learned that data nerds are concerned with "log data" and "MapReduce." We got to hear about these themes in many vivid examples and stories of how big data technologies work

and why they are important. From looking at a few of these descriptions, it was clear that nerds spoke metaphorically and used analogies in order to illustrate those technical features. They never said, “let me give you an analogy,” or “metaphor.” To be sure, learning about the content of these descriptions was analytical relevant as well, in this instance we could for example understand that many of their tools impose no ownership constraints. This background facilitates organizational arrangements others would deny. In addition to this aspect, we also came to understand that data nerds experience and articulate these activities in ways that have nothing to do with their technical specificity.

What was useful in the rich qualitative setting is not necessarily useful here. Whereas in part one we were able to see those contours qualitatively, the chapters in this part aim to formalize this process in order to test their relevance systematically. In order to do so, we need activities that are commonly shared, so that we can analyze the ways in which these activities unfold. They can be found in the job market with job postings, and in the academic setting with publication and reference practices. Second, we need to associate them with data science thought communities, which must not be formalized. In the job market setting, I therefore center the analysis on job titles. While job titles are often formalized for internal purposes, in this setting posters are free to adjust them in order to signal data science affiliations, as far as they are salient to them. For the academic setting, I rely on data science programs and the instructors assigned to teach there. While there is also a formal aspect to this process, they do not have formal backgrounds in data science. Nor would such assignment formally script their researcher agendas, which this analysis focuses on. Finally, these strategies both respond to our concern with lay salience of arcane knowledge.

With this empirical design in place, we need methods that allow us to recover the contours emerging in those communities and their collective activities. I introduce community detection procedures in order to recover the patterns in which these activities unfold. This strategy recovers the guiding idea of understanding data science as a thought community in this formal orientation. Conventional implementations of such community detection would miss some important features of thought communities, however. While they are designed to identify groups on the basis of how they are interconnected, members of thought communities may have other affiliations as well (Zerubavel 1997). With respect to the technical idea of community detection, I repurpose existing implementation by

deploying it to identify the communities within the thought communities identified on the basis of the "human algorithm" operating through job posters and program committees, described above. In other words, this strategy indexes contours formally as the fragmentation and homogeneity of thought communities.

Such repurposing finally brings us back to the idea of a data science of data science. Existing research finds for both hackers (Coleman 2013) and engineers (Bechky 2003a) an emotional identification of nerds with their technical tasks. We saw indication of such attachment as well among data nerds as they spoke of "blood, sweat and tears," and excuses for deficiencies in some applications as no "scientifically cracked" solutions were available. As sociologists have not had the chance to watch a profession emerge in a century, there were no scientifically cracked solutions for studying data science. Developing the following solutions, which use "unstructured" textual data in "JSON" format and simulations "pushed" on "AWS EC2" instances with "shell scripts," and "the command line," thus lead to a more robust understanding of how data science skills and knowledge unfold. I keep the discussion and implications out of the following comparative chapters and leave it for the conclusion and final discussion.

Taken together, this analytical design is uniquely capable of capturing the rare process data science confronts us with, analyzing a thought community of arcane knowledge as it emerges.

9 Public expectations of professional expertise: Contours of Skills and knowledge in data science, law, and other occupations

In order to understand better the basis of data science's salience today, we need to shift focus and ask, in addition to what data nerds know, what others expect them to know. We have considered in chapter two that professions more broadly are seen to assist with general problems. Meanwhile today's increasingly specialized problems require informal expertise that takes into account situational details. Even where formal boundaries typical of classical professions are less profound, the public continues to recognize general expert roles, not only data nerds but also nutrition consultants, facilities managers and so on. This chapter returns to the question of how data nerds' specialized expertise gains general salience by moving on to study data science comparatively with respect to contour lines of salient and obscure expertise.

If neither formal boundaries nor informal processes are predominantly associated with expert salience, as part one where both informal peer relations as well as formalized status and knowledge guided data nerds, we need to focus on expertise directly and ask what constitutes distinct expectations. General expectations have been seen to result from abstract knowledge that integrates instances of specialized expertise such that they find distinct recognition (Berger and Luckman 1966, Abbott 1988). Other cases, especially recent ones, add expertise of specific problems to existing knowledge, whereby occupations or organizations with that knowledge absorb the novel expertise (Eyal et al. 2010, Wynne 1992). Neither mechanism directly accounts for data science, which led us to consider its status of a thought community as a more inclusive view of the contours defining data science tasks. Data science gains saliences without clear indicators of how it combines the arcane and specialized knowledge it utilizes, which moreover falls within organizational boundaries that dominated it in the past.

Understanding what constitutes expert expectations widely requires capturing their shared knowledge. To recall, much of the current literature focuses on social contexts of local and specialized expertise, asking how groups of experts interact with clients and other stakeholders. This can be seen in studies of classical and institutionalized cases, as well as of recently developing cases. For example, research on legal educational programs (Espeland and Sauder 2007) and hiring practices (Sandefur

2015) reveal more variability in the law profession than considered there in the past (Rueschemeyer 1973, Abbott 1988). Specific social contexts also shape processes developing specialized knowledge outside of institutionalized channels, such as those first moving HIV+ on the mainstream medical agenda (Epstein 1996). These findings contribute to earlier accounts that primarily focused on institutionalized occupations and their formal boundaries, showing how these groups come to add new expertise to their stock of knowledge.

Important effects of expert knowledge remain unaccounted for. The shift toward a focus on informal processes has implicitly assumed an isomorphic relationship between institutional boundaries and stocks of knowledge, leading to see the latter's effect fade with the former's demise. Several examples challenge such a view as they show how stocks of knowledge anticipate institutional boundaries and integrate locally specific problems. Formal boundaries thus index stocks of knowledge unreliably. The American legal institution developed its now familiar training channels connecting law schools and firms on top of significantly older intellectual roots (Stevens 1983). Current studies take the contemporary status for granted, explaining variation within it, without considering effects of those roots (Sandefur 2015, see Stinchcombe [2001] offers a notable exception). Psychiatrists, and other specialists who patients under mental distress turn to today, first built this distinct expertise from smaller projects that then also integrated mainstream medicine (Abbott 1988). Stocks of knowledge therefore have to be seen independent of institutional boundaries. Such a view reveals an alternative to the one in which expertise of new problems gets added to existing knowledge. This alternative emphasizes integration of areas of expertise to constitute a distinct expert role. I test these two processes below.[87]

To understand how expert knowledge of specific novel problems prevails as source for general expectations, we need to watch it unfold in a comparative context. Here I compare the emergence and structure of data science with the structure of law, software engineering, risk analysts and financial advisors. Law is seen as a classical profession with institutionalized training in a canonized stock of abstract knowledge. Its institutional boundaries and distinctive knowledge now overlap. Software engineers also undergo prolonged training, but expectations follow from their subsequent specializations. Risk analysts and financial advisors are each expected to be trained in a specific task. Data scientists

[87] These process map onto those we have considered in chapter eight, where we observed data nerds both enlarging the canon of classical scholars they turn to for guidance, as well as reinterpreting those other quantitative fields have considered for long.

combine statistical expertise with programming skills to access data and implement analytical strategies otherwise unavailable in response to specific problems from a range of areas. By analyzing the ways in which data science is expected to address different specializations, this chapter introduces a new perspective on the informal underpinning of occupations, systemic contours of expertise, and the relationship between abstract knowledge and public recognition.

This approach also aims to contribute further to our main concern with how data nerds, and experts more broadly, shape private and public life. Consensus on this relationship and its origin in experts' arcane knowledge ties two otherwise deeply divided perspectives. Disagreements follow from theoretical comparisons of selected cases that discover variation in the origin of expert knowledge. It found evidence in support of both views in this novel case. Yet, by just focusing on the data nerd case, part one remains consistent with this design choice and its limitations. The strategy in this chapter is different. Here I compare expert knowledge and skills empirically. It also responds to recent efforts that directly consider public views. This change of perspective circumvents a priori assumptions of scope and directionality.

Focusing on this specific problem we can recall the part of the literature that concerns expert groups. One camp focuses on exclusive professional groups utilizing esoteric knowledge to assist with lay problems in contrast to recent views arguing that such knowledge emerges at the level where experts interact with each other and with the lay clients whom that knowledge affects. The former bases its arguments on studies of institutionalized occupations with knowledge that transcends specific organizations such as medicine (Freidson 1988), law (Abbott 1988) and economics (Fourcade 2009, 2006). Although economists share less variable boundaries than, for instance, Bar admission imposes on lawyers, they too undergo coherent training. In contrast, the other side argues that arcane knowledge emerges from interactions with lay clients, citing a range of problems such as the discovery of autism with help of parents (Eyal et al. 2010) and understanding novel administrative structures of the EU through local interaction in key cities (Mudge and Vauchez 2012). This view has begun to directly challenge the institutionalist perspective (e.g., Sandefur 2015, Eyal 2013). In order to understand the results of part one, however, we aim to find and specify complementarities. The comparison of data science and law uncovers expectations of similarly diverse skill sets in spite of varying degrees of institutionalization.

Public salience, the foundation of professional autonomy and continuity, thus cannot be understood through either one of the dominant approaches alone.

The other part of the debate that we need to take into account here concerns the public perspective. We have turned to Dewey (1954) early on in order to consider that activities may likely gain public salience even if they have no formal boundary as long as these activities have systematic consequences. With the aim to now formally model contours of the classes and categories we took from Dewey's view initially, I here turn to a more specific debate where one side focuses on opinion patterns and public perceptions of cultural objects whereas the other camp analyzes the makeup of those objects and practices of actors who shape it. External expectations, for example, prevent law firms from diversifying into new legal specializations (Phillips, Turco, and Zuckerman 2013) and law schools from offering unconventional training (Sauder and Espeland 2009). The other side finds assumptions of homogeneous audiences, which are implicit in the first set of arguments, overly simplified, citing well-known examples of heterogeneous patterns in political opinion or cultural tastes (see Goldberg et al. [2016] for a summary).

With these directions in mind, this chapter takes one more step in response to the diverse audiences we have seen in part one, and studies instances of collective and decentralized persuasion of the public. This design is responsive to the challenges that come with thought communities. Political parties, movie productions and even professional services firms, the focus of existing literature, can cater homogeneous or variable offerings to audiences with universal or heterogeneous tastes. In the context of fading institutional boundaries in many modern contexts of work, nerds lack coordination mechanisms to do the same. Addressing both existing views, this analysis demonstrates a specific process in which publicly expected expert knowledge results from integrating specialized skills through formal abstraction.

In other words, this design here moves away from considering the degree to which observations here resemble the familiar technology nerds, the framework of part one. Instead, this analysis compares contours of data science knowledge formally to those of other cases that are similar and different relative to data science on relevant dimensions, on an empirical basis. By shifting focus, this chapter develops a formal framework for considering experts at scale without assuming formal boundaries as a prerequisite for public recognition.

Roadmap

Although we have considered the data science case from many perspectives by now, I being in section 1 with a brief note on which aspects we focus on here. I then quickly move on to a discussion of data and methods in order to operationalize areas of expertise through the skills constituting them at the intersection of lay and expert perspectives, and to analyze how their structure pertains to the problems experts are expected to address. I design samples of large numbers of job descriptions from an online database and develop an analytical strategy that leverages their informal content amid their formal structure in order to extract representative skills and model they fold into knowledge that applies across specific problems. Abstract knowledge integrates skill sets and leads some to apply across organizational and institutional contexts more easily than others without it. Section 2 presents a series of analytical steps, and results. I show that data scientists resemble the old professions from the perspective of their abstract knowledge but without relying on institutionalized channels. Software engineers are seen as experts for highly specialized tasks. Risk analysts and financial advisors draw on expertise previously associated with occupations of little autonomy. Moreover, the temporal sequence by which data science skills, originally from different areas of established expertise, fold into knowledge shows that they distinctively integrate the original areas of expertise, instead of adding one to the other. In section 3, I discuss these findings to argue that the perspective focusing on thought communities recovers a scope that offers explanatory purchase to the extent that integrated stocks of arcane knowledge facilitate expert recognition and expectations.

9.1.1 Experts skills and knowledge

The old professions were seen to operate in formally institutionalized contexts that could index the abstract stocks of knowledge they were recognized for. For example, the legal profession's knowledge was found to most significantly cluster around types of clients, ranging from individuals to large corporations (Heinz and Laumann 1982). The medical profession organized knowledge similarly along the hierarchy of general practitioners to highly specialized intensive care physicians (Menchik 2014, Freidson 1960). It follows that abstract knowledge capable of integrating different specializations, which is seen as a source of general recognition, could be indexed by the occupation's appearance. Such a

strategy offers no leverage for understanding today's expert groups that emerge increasingly without building up comparable formal boundaries.

In studies focusing on the specific interactions between members of expert movements, indexing abstract knowledge that a group of experts is collectively recognized for becomes more complex by definition. It has been shown that informal interactions produce formal codification (Collins and Evans 2007). It is less clear, however, how this process unfolds systematically throughout a group, and consequentially, how it pertains to general expectations. One feature of expertise, which a focus on informal processes reveals, operates through the direct interaction between experts and non-experts, such as parents of autistic children and medical doctors (Eyal et al. 2010), or HIV+ individuals and the medical community (Epstein 1996). The extent to which it became obvious for mentally distressed to expect relief from medical instead of theological knowledge, with the former largely relying on informal mentorship (Abbott 1988), illustrates that this mechanism of direct interaction does not account for how it is that expert groups just emerging gain salience to those not yet directly interacting with them.

Occupations that continue to utilize abstract knowledge but no longer operate within formal boundaries leave analysts without clear indicators of skills and expertise. Clients and the public still form expectations of arcane knowledge relevant for their problems. Even absent institutional machinery, experts remain associated with the labels of their specialty. Experts who provide criminal defense, write contracts and patents are lawyers. They have a JD and most likely also admission to the Bar. Experts whom we expect to build houses, report the news and design operating systems are respectively thought of as architects, journalists and software engineers. What constitutes their relevant knowledge is much less clear. In order to account for occupational groups without solid institutional foundations, we need to understand how skills, which express their expertise, are seen to fold into those labels directly.

9.2 Research design, data and methods

Among the many accounts of data science work, projects and practices, we have also seen that data nerds themselves conceptualize their work in terms of job requirements. This is seen to descriptions of their role in terms that combines expertise of a “data hacker, analyst, communicator and trusted advisor [whose] most basic, universal skill is the ability to write code” (Davenport and Patil 2012). They also acknowledge its arcane background, and none of these skills are new, as a definition we have considered

before reminds by describing the data scientist as a "[p]erson who is better at statistics than any software engineer and better at software engineering than any statistician" (Wills 2012). These descriptions, from data nerds, reveal a tension between complementary skill sets and disciplinary stocks of knowledge that keep them apart. How do skills with arcane and institutionally separate pasts become jointly expected among lay clients?

The other aspect this chapter focuses on is that data nerds can be seen to bridge organizational and institutional boundaries and apply their analytical tools across substantive problems. These transitions lead them to encounter unfamiliar settings, or "brick walls," that undermine conventional solutions and instead challenge their pragmatism. Faculty at elite research universities interrupt their academic appointments for positions in data science departments of non–academic organizations (The New York Times and Facebook offer just two prominent examples). Such crossings are not limited to classical divide between academia and the economy. DataKind, a nonprofit project, for example, aims to facilitate "data science for the common good." It matches experts to clients who have data that pertains to public welfare but lack the analytical skills and capacities. One data scientist, for instance, reports to have used code across problems as different as modeling online click–through rates and analyses that helped a city parks office managing trees in public spaces. Most recently, the White House has followed initiatives by cities across the country to utilize data science for public management (e.g., Smith 2015, Bloomberg 2013). These moves indicate how data scientists combine their analytical and technical skills in response to the novel problems they encounter. In conceptual terms, these specific instances leave unclear whether data science additively combines quantitative expertise and programming skills, or integrates them into a distinct form of abstract knowledge of how to solve novel problems.

How do these historically separate fields of scholarship combine into a distinctive stock of knowledge that leads to collective expectations of what data scientists know across firms, nonprofits and government organizations? Answering this question requires a comparative design that also considers cases that collapse knowledge into specialized expectations.

9.2.1 Research design

Understanding the structure and emergence of stocks of data science knowledge as the basis of distinctly salient thought communities requires a comparative view. This novel case has to be considered

against familiar groups of experts that vary based on what they are collectively expected to know in order to address the initial concern with reconciling general expectations and increasing specializations. In part one we have seen how data nerds integrate the several unrelated specializations in part through references to classical scholars whose work they reinterpret, as well as through references to largely uncommon references through which they redefine the canon compared to existing quantitative fields. These references integrate data science practices for the community, but not the mundane and specific problems data nerds encounter on a routine basis. In this respect we tentatively saw in particular coding languages, which are at the same time so ubiquitous in modern technology work as to leave it unclear how they might distinctly integrate data science specializations. Lawyers are expected to be competent in multiple specializations as well. Software engineers differ. They too address a range of problems, but expectations pertain to their skills and competencies in a respective specialization either defined by the formal bureaucracy or combined in a heterarchical arrangement of the kind we found in Linus Torvalds's Linux project. Financial advisors and risk analysts are expected to consult on financial planning and risk estimation. Both are specialized and singular tasks. All utilize abstract and esoteric knowledge or expertise. Data science varies relative to these other types of expert communities with respect to the collective expectations of their knowledge. Why?

The variability of these cases' formal boundaries rejects plain vanilla institutional explanations of external expectations. Data scientists face shared lay expectations regarding their competencies in quantitative analyses whilst lacking significant institutional infrastructure. Lawyers have both. Moreover, financial advisors operate in a more scripted context compared to risk analysts, but both invoke specific expectations. Considering further cases thus reiterates the point that the formal institutional context of expert groups is thus unrelated to the collective expectations they face, which we have recovered for data science empirically. Here I analyze the stocks of knowledge and their contours directly.

For this purpose I extend the methodological considerations from the introduction. To recall, the main premises there were to consider opposing views emphasizing formal and informal features of expert work as different instances of thought communities. In order to address the specific question of how stocks of expert knowledge facilitate collective expectations I need a data structure that reconciles lay and expert views without imposing limitations on what each side may consider relevant. Several

conventional strategies address just one of those concerns. Surveys, otherwise the basis to study various lay views on expert work (see Abbott [1981] for a summary), lead to responses that may speak to one side but make less sense in the terms of the other side. Direct observations address this problem (Navon and Shwed 2012, Wynne 1992). Both direct observations and client surveys require the analyst to define a group of experts or a class of problems, which leads to ignore instance that also utilize relevant expertise without resorting to the expert designation. These considerations favor a view of stocks knowledge as arrangements of varying lay expectations. Patterns of such skill arrangements can be compared in order to shed light into the question of what constitutes general expectations of specialized experts.

The comparative design around institutionalized and emerging stocks of knowledge poses a significant operational challenge. I extend strategies that understand expert fields from their relational characteristics to identify professional skills on the basis of their structural properties. Studies of expert groups suggest skill sets as instances of professional knowledge (Freidson 1986). Skills, however, index content of knowledge and therefore do not compare easily across different fields. Moreover, data science, which is just emerging, lacks institutionalized skill sets by definition and therefore escapes operationalization strategies that rest on formally defined competencies. Both problems can be addressed with strategies from comparative studies of knowledge that analyze structures of knowledge and show that these characteristics meaningfully represent areas of expertise (Aral and Van Alstyne 2011, Shwed and Bearman 2010, Martin 2002). I implement these procedures here for identifying skill sets that constitute expert knowledge, and extend them to determine structural properties of the relations between them, allowing to analyze expertise comparatively.

Table 9.1. Sample structure

Title	Analytical role	N (in 1,000)
Attorney	Baseline	2.9
Data scientist	Focal case	1.8
Risk analyst	Occupation/quant	1.2
Software engineer	Occupation/tech	8.4
Financial advisor	Occupation/bureaucratic	1.8
Random	Common skills/ skills trajectories	~40

We need data that reflect these scope conditions.

9.2.2 Data

Job postings address the analytical concerns in several ways. First, they indicate what lay clients expect experts to know. They also articulate those expectations such that they speak to the intended group of experts. Job postings moreover accommodate the tension between formal groups of experts and their informal stocks of knowledge by indicating both a title, such as data scientist, attorney, software engineer, and so on, and a description of what skills and knowledge are expected for the specific role. I therefore collected samples of job postings that are associated with the different groups.

Analyzing how distinctive stocks of knowledge emerge from specialized expertise requires two types of samples. One sample of job descriptions needs to explicitly ask for a certain type of expert by mentioning that expert's role as a designated label. These descriptions indicate what those who post them—HR, recruiters or peers—expect from someone who applies for the designated role. A sample of random job postings needs to complement the selected samples. These postings indicate associations with the expert might also be expected where they are not associated with the respective role label. In other words, with the two sets of samples we can distinguish arcane skills from common skills.

The basic data structure is simple because the job postings initially collapse much relevant information in their textual descriptions. They are initially "unstructured," to use a term data science has introduced us to. It is textual and does not lend itself to exploratory summaries, though table 9.1 provides an overview of the sample structure, and table 9.2 shows an example of a job description for a data

Table 9.2. Job posting example

Title	Senior Data Scientist, 1 Trillion Monthly Transactions
Job description and requirements	You will handle data exploration, hypothesis creation (from both business and product goals), testing algorithms, scaling to large data-sets and validating results. We have a broad set of technologies with which the Senior Data Scientist will work: Hadoop/HDFS; Shark/Spark; NoSQL databases, and numerous charting, graphing and analysis applications such as: Gephi, Google Charts, etc. [...] We'd like to see good coding skills covering some procedural as well as statistical or data oriented languages. (Such as: Java, Scala, Python as well as R, SQL, etc.). Good communication skills and an awareness of how to communicate data effectively is a must. This individual must be comfortable working in newly forming ambiguous areas where learning and adaptability are key skills. Required Education: MS or higher in the field of Statistics or Computer Science or Applied Mathematics.

science position. Extracting relevant skills from these qualitative accounts such that they are suitable for comparisons requires a systematic approach, which I discuss next.

9.2.3 Methods

I analyze job descriptions in order to understand how abstract stocks of knowledge are seen to emerge from their specialized applications. This requires two main analytical steps. The first step identifies skill sets from job descriptions, a type of identification process. The second step analyzes the degree to which these skills form specialized clusters of expertise, and whether experts integrate knowledge from different contexts and if they do, in which ways. These clustering patterns then represent a kind of contour line. Here we can recall the conversation between an audience member and a speaker, where the audience member inquired about what the data scientist was "processing [the model] on" (incidentally, or not, also for a textual analysis). The data nerd mentioned Python and some other languages, and emphasized MapReduce. This clarification did not come immediately, however, because the terms of the question were initially not clear to the speaker. In another, similar instance we could see that this is in part because similar terms appear in machine learning culture in different ways compared to statistics. In other words, even data nerds equally familiar with languages may not immediately understand each other with respect to how they conceptualize those languages and the way they use them. Whereas settling on common terms was possible at the interactive settings in part one, it is not so easy in more fleeting interactions of identifying expert status. Table 9.3 summarizes the specific steps

Table 9.3. Analytical design

Steps	Purpose	Method	Data	Strategy
1	Skills/global	Coefficient of variation of the mean frequency and intracorpus frequency	1st 1/3 of focal data*	Extracting skill sets
2	Skills/local + classifier	Logistic classifier estimation	2nd 1/3 of focal data*	
3	Index	Logistic classifier predictions	3rd 1/3 of focal data*	
4	Discovery	Logistic classifier predictions	Random sample	Inferring specializations
5	Structural model	Newman modularity + skill density	Positive predictions of step 4 + skill sets of step 2	

Notes: Stars (*) denote sample of focal corpus matched with a random sample.

that extract these moments of overlap and confusion systematically, which I describe in detail below. In the context of job postings, the resulting data structure indexes the generality and specificity of knowledge that is expected of the respective group of experts.

Extracting skill sets

The comparative research design builds on a relational understanding of skills, and a structural definition of knowledge and specialized expertise. I propose an empirical strategy that exploits the variable content of job descriptions and their standardized format across fields of expertise. This implementation inductively operationalizes ideas that skills index expertise based on their relevance throughout an expert movement and their distinctiveness relative to other fields (Freidson 1986).

Skill sets that distinctively represent a focal expert group apply across problems which that group is expected to address, while remaining generally uncommon. I utilize strategies from textual analysis (Aral and Van Alstyne 2011) in order to operationalize distinctively expected skills. The key idea here is to describe a corpus of textual job descriptions as a set of keywords with varying frequencies. A word is representative if it is used systematically throughout descriptions of a certain type of positions. It is distinctive if it is otherwise rarely used in the overall corpus that includes job descriptions sampled at random. Technically, I implement these two intuitions by adopting the "coefficient of variation of the mean frequency" of a keyword across the two corpora, and the "intracorpus frequency" of a keyword (see step one in table 9.3).[88] Practically, I begin with combining two text corpora. The first contains job posts that

[88] See Aral and Van Alstyne (2011, 122) for details.

describe the same focal profession, seen in the assigned job title, such as data scientists, attorneys, and so on. This information is indicated by the job poster as free text, not selected from a list. The second corpus comprises a random sample of professional job posts. In the context of task descriptions, which make up the main portion of job posts, the resulting keywords constitute skills, hence indicating skill sets as job posters expect them from someone filling the role they indicated. These steps so far ensure that the skill sets represent the corpora in general, that is, they capture identification mechanisms through informal skill combinations.

These measures primarily index the aggregate set of problems a focal group of experts is seen to address, but skill expectations pertain to specific problems. I test their local relevance by considering the utility of skills mentioned in a job description for inferring the label assigned to it. I train and test a set of logistic classifiers for this purpose. The classifiers are assigned the previously identified skill sets as features, indicating presence or absence of a given skill in a given description. The outcome is represented as a binary variable that indicates whether a post has a title of interest, or not. I consider the classifiers' precision and recall to index the distinctiveness of a skill set and its representativeness (see step two in table 9.3), and as a first measure of contours. [89]

Generally representative skill sets and classifier performance reveal a tension between the local utility and global transposability of expertise that is a pivotal feature of abstract knowledge. Representative skills are derived from one part of the overall corpus for a specific role, whereas classifier performance is assessed on the basis of their success recognizing specific posts from a different part of the corpus. I leverage this tension for analytical purposes by generating a distribution of results. The distribution emerges from applying the classifier to a series of randomly composed subsamples of a given corpus. It indicates the degree to which skill sets generally representative of some the positions systematically pertain to specific other positions. This distribution thus indexes coherence of a stock of knowledge (see step three of table 9.3), or its contour lines.

Practically, the random sampling iterations unfold across three operations. Each begins with three equally sized and non-overlapping random subsets of a focal sample of job descriptions. The first subset is used in the previously described procedure for extracting skills based on their representativeness and

[89] Recall is defined as the percentage of those classified out of all those that are part of a respective class or category. Precision is defined as the percentage of the correctly classified posts, out of all those classified.

distinctiveness (recall step one in table 9.3). The second subset, matched with random descriptions, is used in this step to train the classifier to recognize descriptions associated with a focal group of experts within a larger corpus. I hold out the third subset as a test set (recall steps two and three in table 9.3). I iterate through these steps multiple times, randomly reshuffling the subsets of the focal corpus in each iteration, to account for nuances in the type of work a respective community does (see "extracting skill sets" steps one to three in table 9.3). In each iteration I record the relative importance of each skill the first step revealed in the context of the step two classifier, as well as the precision and recall of the classifier in step three.

This strategy has two main payoffs. The distributions of precision and recall scores index coherence of a stock of knowledge through the skills expected from it. Second, relevant knowledge emerges on the basis of relevant skills and their structural arrangement in the content that describes expectations of different expert groups and throughout a corpus of job descriptions. I next consider the structural arrangements induced by the relevant skill sets and the relationships between them in order to understand stocks of abstract knowledge in the context of the generality and specificity of the problems they address.

Inferring specializations

This analytical strategy allows us to consider positions beyond those explicitly declared as part of a profession, based on their assigned job label. The classifier initially indexes structural properties of expertise and reveals constituting skills sets. It can also discover expectations of skill sets resembling those of a focal profession, but not attributed to it. In data science, for instance, bureaucratic requirements may designate a title other than data science, albeit seeking someone with the skills of a data nerd rather than the more constrained skill set we have seen speakers in part one associate with other data roles. I apply the logistic classifiers previously trained and tested in steps two and three on the labeled job descriptions (see also table 9.3), to a large sample of random job descriptions. These descriptions were not selected based on any particular title and were not used in previous steps of the analysis. In step two, I train a series of classifiers for each group of experts to account for nuances in their skill sets. I account for this diversity and apply the set of classifiers iteratively to the random sample of job descriptions. This results in a distribution of likelihoods of each post to be assigned to the labeled

positions or not.[90] Here I focus on those observations that are on average seen to resemble the focal group. This procedure accounts for posts that are relatively close to most specializations, as well as those that seem very close to some and not close at all to others (see step four in table 9.3). This step produces a sample of job descriptions. Jointly, they provide the basis to analyze the distribution of skill sets as they represent a group of experts, without resorting to labels assigned to them.

Skills index knowledge as they fold together across applications. By this I mean that the patterns emerging from the ways in which skills are expected in varying combinations across positions index properties of the stocks of knowledge that integrates them. I operationalize the specialization and fragmentation of stocks of knowledge by recovering their contours from inferring clusters based on the structure of positions that expect overlapping skill sets.

In this final step, I first create a bipartite network with job positions and skills as two levels of nodes. I project this network on the level of job positions (Breiger 1974).[91] Two positions are then connected if they share one or more skills. Next, I estimate Newman modularity (Newman and Girvan 2004) to infer underlying fragmentation into groups of positions that ask for overlapping skill sets in order to detect specialized communities of positions with overlapping skill expectations (see step five in table 9.3).[92] This operationalization considers positions part of the same specialization even if overlap between skills they expect is imperfect.

The modularity-based community detection strategy is designed to reveal fragmentation. The idea that abstract knowledge encompasses distinct specializations requires that I consider the overlap between clusters revealed by the modularity estimation within the overall structure. I consider the density of skill set overlaps within and between specializations in order to operationalize the degree to which skills indexing expert knowledge address distinct and overlapping specializations. This results in a qualitative description of the underlying stock of knowledge with respect to its concentration of expertise in a singular specialization as opposed to integrating several specializations, on the basis of abstract knowledge.

[90] In the prior evaluation of the classifier, positive likelihoods would count as false positives. The purpose of this step is different, however. Here I aim to identify specific job posts that resemble those of a focal profession based on the skills expected from an applicant, but without attributing the respective professional title to it.
[91] I use a three-month time resolution, for ties.
[92] This recovers the idea of communities accommodating multiple affiliations form Zerubavel (1997).

The methodological strategy driving steps one to five (in table 9.3) generates an empirically grounded visual model of specializations underlying externally recognized expert knowledge. Network analytic studies often invite visualizations. Those visualizations require algorithms that organize the position of nodes and edges. Here I rely on a much simpler version. We are concerned with the contours of thought communities with respect to their integration of heterogeneous applications, as they index abstract knowledge.[93] This can be seen in simple adjacency matrixes in which rows and columns represent nodes, job postings in this case, and cells the edges between them, shared skills here. Contours of communities asking for singularly specialized knowledge then look different from contours of those with bureaucratically defined expertise. Those with abstract knowledge look different yet again. The empirical context of job postings considered here interacts with the effect of abstract knowledge in several ways.

Controls

Skill arrangements do not alone follow from their epistemological fit. Although formal organizations played a secondary role to data nerds on task specifications in the accounts of part one, in this comparative settings we have to return to C. Wright Mills's observation of the ubiquitous dominance of great bureaucracies over work. Organizations thus shape stocks of knowledge as well. Lawyers rose with the expansion of bureaucracies (Abbott 1988), European Union officials with the geographic concentration of political institutions (Mudge and Vauchez 2012) and today insurance companies shape medical practices (Gorman and Sandefur 2011). Here I account for the organizations advertising their interest in a certain type of expert, and their characteristics. The job descriptions data includes information on the posting organization's size (in employees), the industry or industries they are active in, the type (governmental, private, public, etc.), whether they are publicly traded and the year when they were founded. Considering these characteristics allows for testing the degree to which certain types of organizations shape stocks of knowledge.

[93] Compare this for instance to the much more nuanced interests in motifs, such as triads, fourcycles, cliques, and so on. They are irrelevant here, hence a simple adjacency format suffices.

9.3 Analysis and results

The analysis unfolds in three parts. Each part addresses the question of how specialized data science expertise meets general expectations form a different perspective. The first part develops a set of classifiers in order to consider the composition and distribution of skills as aggregate index of expert knowledge. The second part generates structural models of job positions and the combinations of skills they expect and compares the different cases in order to distinguish contours of specialized and of abstract knowledge. The third part considers the temporal sequence by which data science knowledge combines the distinct skill sets it draws on, focusing on whether one gets added to another, or whether distinct expertise integrates them. The organizational context of these sequences is also considered. This strategy produces complementary results.

The results can be briefly summarized. Overall, the analytical steps reveal data science as a thought community with its identification with distinct skills and expertise that integrates specialized skills with an abstract body of knowledge. This is thus consistent with results of part one. In the present setting and framework, this can be seen as classifiers identify data science job descriptions based on combinations of specific skills articulated across distinct samples. Moreover, comparisons of the contours emerging from structures of positions and the skills they expect reveal similarities between data science and law, and differences between those of them and specialized occupations. Finally, temporal sequences reveal evidence to show that the structure underlying data science undergoes integration through abstract knowledge that is distinct from the fields of expertise data science draws the skills from. The following sections discuss each step in detail.

9.3.1 Arcane skills

This section leads to analytical and substantive conclusions. Here I implement the first part of the analytical strategy just described (steps 1-3 in table 9.3). I first extract skill sets that are expected of expert communities in general and test their relevance throughout specific job descriptions. I end with considering quantitative measures of how well skill sets map onto expert groups as first indicator of the structure of their underlying knowledge.

This section reveals some key findings. Job posters consistently expect coherent skill sets from data scientists, lawyers, risk analysts and financial advisors, but vary in their expectations of software

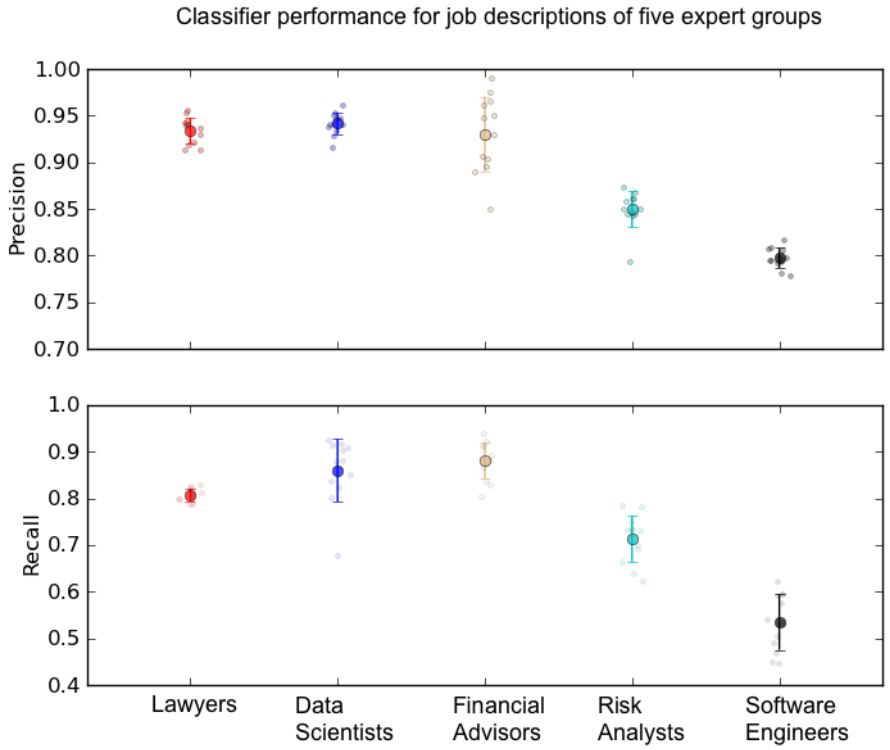

Figure 9.1. Mean, standard deviation and observations of classifier performance for five occupational groups.

engineering skills. This finding supports the merit of the analytical strategy and shows that the idea of a distinct and coherent data science thought community manifests itself in subtle agreement on their concrete skills. This analysis remains inconclusive with respect to whether their skills are primarily structured by abstract knowledge as part one suggested and typical of autonomous professions, or by specific industries as in subordinate occupations.

In the context of job postings, I find shared expectations of what just specific experts would know, for all cases. This can be seen as skills lead to exclusively identify expert labels. Specifically, figure 9.1 shows in its top panel that formally relevant skills consistently predict labels representing the legal profession, data scientists and financial advisors with a precision (i.e., the share of correct positives) of around 95%, risk analysts with a precision of 85% and software engineers with a precision of around 80%. Some groups are seen to command more exclusive skills than others. I discuss this variation below. For now it is important to note the high degree of precision for the focal data science group and three comparative cases. In order to conclude that experts utilize an integrated stock of knowledge, skill sets also need to occur throughout positions looking for a respective expert.

Moving to the bottom panel of figure 9.1, the analysis also reveals shared expectations of what skills experts throughout a group should know, for four of the five cases. This can be seen as skill sets variably represent labels of a specific group of experts. The classifiers represent data scientists and financial advisors well, with average recall rates (i.e., share of true positives out of all true observations) of around 90%. Attorneys are more fragmented; each iteration only captures 80% of the job descriptions in the test set. This is likely a result of law's institutional status, which lowers the need for detailed descriptions. No one else uses the skills (see high precision), but not all use it in sufficient richness for law. For risk analysts, with each iteration extracting skills that are only representative of roughly 70% of the remaining descriptions. Their boundaries are more penetrable. Finally, software engineering consistently classifies no more than around 55% of the positions in fact expecting a software engineer. We now see that the exclusive expectation of skill sets is independent of those skill sets' consistent expectation throughout applications of a type of expertise. Some expert groups are associated with more general knowledge than others.

Clients have clear expectations of what data scientists should know, just like they have for lawyers and other familiar groups. Closure over jurisdictions of problems and the knowledge and skills to address them lies at the center of our understanding of professions (Abbott 1988). That skill sets distinctively and broadly represent job posts for lawyers, one of the most prominent cases, therefore demonstrates the utility of this operationalization. The similarly strong results for financial advisors and risk analysts, which are also widely recognized, provide further support. Data science's emergent status leaves no basis for expecting one outcome over another, and if anything would lead to expect less agreement than the analysis finds. Yet data science skill sets represent job positions just as well as those of some familiar cases represent theirs, and better than others. It follows that although data science shows few signs of institutional status as a profession today, its clients already attribute it a distinct set of competencies.

Clients are not so clear on their expectations of software engineers. Nor would we expect them to be. Engineers have long puzzled the sociology of professions (Abbott 1988). Unlike occupations that focus on mundane tasks, such as administrative jobs, engineers rely on highly arcane knowledge. Unlike scientific disciplines they primarily apply their knowledge practically and without significant theoretical guidance. Yet they remain disintegrated as a profession. This holds for software engineers whose

expectations largely derive from their prior experience and the specific organizational context they need to apply their knowledge in, but no abstract stock of knowledge integrating those applications (Kraft 1977). This can be seen today in their division into developers, architects, interface designers and so on, defined in bureaucratic divisions of work. Software engineering neither resembles data science's integrated stock of knowledge nor its salience as a coherent expert group. Software engineering demonstrates that data science's salience does not result from its use of programming skills. This contribution exhausts its analytical leverage for the question of data science's salience and I will disregard software engineering from subsequent steps.

The consistent results for lawyers, financial advisors and risk analysts remain inconclusive as to how abstract knowledge pertains to salient expectations. Law constitutes a highly autonomous institutional component of the American society. Financial advisors and risk analysis work for a specific set of financial institutions. Yet in this analytical step all three showed similar skill expectations, or identification with similarly solid contour lines. In order to understand data science's underlying stock of abstract, or narrowly specialized, knowledge, I focus on its capacity to integrate heterogeneous problems and compare data science and the remaining three cases with respect to this next.

9.3.2 The formation of data science knowledge

This section considers the emergence and contours of data science knowledge. It focuses on relational patterns of skill expectations beyond the boundaries of the expert groups these skills index. Step four in the analytical design (recall table 9.3 above) generates the data for this analysis. As explained above, it applies the classifiers from the previous step to a longitudinal dataset of randomly selected job postings. The positive classifications of this step yields a data structure of job descriptions that resemble those attributed to the focal expert groups with respect to their distinctive skills, but not professional titles. The patterns by which skills co-occur across these positions therefore represent the contours of specific clients with overlapping or different skill expectations for problems not exclusively attributed to an expert group. The patterns can be visually represented and compared.

The patterns are modeled in several steps from the concrete job postings expecting their constituting skill sets. I begin with generating adjacency matrices of relationships between job descriptions based on their shared skill expectations. Following the methodological description above, the

rows and columns of these matrices are ordered such that those that are part of the same community detected by the modularity analysis are placed next to each other. The cells close to the diagonal of these matrices therefore represent ties between positions considered interconnected with the neighboring positions in the matrix. Cells in the off-diagonal indicate ties between positions that are part of different communities, following from the modularity estimation.

Next I consider the density of skills that connect the positions into clusters, based on their interconnected skills.[94] Here I focus on their distributions in order to compare the four cases. I consider the arrangement of clusters by viewing patterns of density distributions across clusters. Abstract knowledge implies the ability to transpose skills from one context to another with few direct connections between them. This basic property can be recognized qualitatively across cases. I produce new matrices (in the lower portion of each panel in figure 9.2) in which I code those clusters less than one standard deviation below the mean as not overlapping (blank areas in figure 9.2), and those within one standard deviation above and below the mean, and above, as somewhat overlapping (grey areas in figure 9.2). This procedure generates models of knowledge structures based on how the skill sets pertaining to that knowledge cluster positions that expect overlapping skills. The presence of empty blocks is most important for understanding the degree and diversity of integration of specialized problems. The specific location of grey and empty blocks within one case over the three periods is unimportant beyond the question of whether they are on the diagonal or not. The top row in each panel also signal particularly dense areas. They are useful points of reference when considering the role of organizations but have no bearing on the analysis of abstract and specialized knowledge, and are therefore not coded separately.

[94] This analysis combines ideas from blockmodeling (White et al. 1976) with more recent methodologies from research on scientific knowledge, consistent with the empirical categorization framework from the introduction.

Table 9.4. Distribution of relevant job descriptions across organizational types

Year-Q	prof (N_p)	Par	Non Pro	Gov Age	Pub Com*	Pri Hel	Edu	Sel Own
2010-4	fa (12)	0.08	-	-	0.58 (0.86)	0.25	0.08	-
	att (27)	0.07	0.07	-	0.52 (0.78)	0.33	-	-
	ds (33)	-	0.03	-	0.61 (0.84)	0.33	0.03	-
	ra (39)	0.05	0.08	-	0.67 (0.58)	0.18	0.03	-
2011-4	fa (72)	0.03	0.04	-	0.46 (0.97)	0.47	-	-
	att (78)	0.09	0.04	-	0.55 (0.93)	0.32	-	-
	ds (90)	-	-	0.01	0.64 (0.89)	0.34	-	-
	ra (114)	0.05	0.02	-	0.57 (0.69)	0.36	-	-
2012-4	fa (171)	0.03	0.07	-	0.63 (0.9)	0.27	-	-
	att (140)	0.1	0.02	-	0.44 (0.97)	0.4	0.03	0.01
	ds (179)	0.01	0.03	0.01	0.51 (0.85)	0.44	-	-
	ra (241)	0.05	0.06	-	0.57 (0.59)	0.33	-	-

Table 9.5. Distribution of job descriptions across organization size (number of employees) categories

Year-Q	prof (N_p)	2-10	11-50	51-200	201-500	501-1000	1001-5000	5001-10000	10001+
2010-4	fa (12)	-	0.08	0.08	0.08	0.08	0.17	0.17	0.33
	att (27)	0.07	0.19	0.07	0.04	-	0.15	0.11	0.37
	ds (33)	0.03	0.03	0.12	0.12	0.03	0.12	0.03	0.52
	ra (39)	-	-	0.1	0.05	0.03	0.13	0.13	0.56
2011-4	fa (72)	0.01	0.07	0.11	0.07	0.01	0.17	0.11	0.44
	att (78)	0.03	0.06	0.05	0.04	0.09	0.18	0.12	0.44
	ds (90)	-	0.02	0.1	0.12	0.01	0.14	0.04	0.56
	ra (114)	-	0.02	0.05	0.08	-	0.08	0.16	0.61
2012-4	fa (171)	0.01	0.02	0.04	0.05	0.02	0.13	0.14	0.59
	att (140)	0.01	0.05	0.04	0.08	0.05	0.18	0.07	0.52
	ds (179)	0.01	0.06	0.08	0.2	0.04	0.13	0.04	0.45
	ra (241)	-	0.05	0.01	0.12	0.01	0.1	0.06	0.66

Data science knowledge

These structural models reveal variation in the contours of skills folding into knowledge that the previous analysis missed. This section focuses specifically on the most recent period. The next section considers the temporal sequence. That lawyers utilize abstract knowledge is well known (Abbott 1988). Here we see skills associated with job descriptions for attorneys co-occur in positions such that they form distinct as well as overlapping clusters (figure 9.2c). Contours of financial advisors and risk analysts, which are not associated with abstract knowledge, differ. Each forms a dense core which overlaps with all other clusters but lacking distinction (figure 9.2b/d). This shows that abstract knowledge, the source of law's autonomy, is indexed by its capacity to integrate distinct specializations. This finding is consistent

Table 9.6. Table showing summary statistics of age of organizations that post relevant job descriptions

		Year Founded	
Year-Q	prof (N_o)	Mean (Std.)	Median (IQR)
2010-4	fa (11)	1955 (46)	1973 (87)
	att (24)	1974 (33)	1985 (28)
	ds (27)	1976 (48)	1995 (23)
	ra (30)	1956 (49)	1975 (60)
2011-4	fa (56)	1954 (56)	1977 (57)
	att (64)	1967 (46)	1983 (33)
	ds (50)	1981 (42)	1997 (19)
	ra (63)	1965 (54)	1989 (57)
2012-4	fa (89)	1965 (50)	1984 (38)
	att (100)	1961 (50)	1985 (74)
	ds (117)	1992 (23)	1999 (17)
	ra (95)	1972 (47)	1991 (38)

with arguments and evidence that see law as an autonomous profession, and financial advisors and risk analysts as occupations lacking that autonomy as a result of their inability to draw on abstract knowledge. Finally, focusing on the last period, the patterns emerging from skills indexing data science positions (figure 9.2a) resemble the images of law (figure 9.2c) more than those of the two other occupations (figure 9.2b/d). This suggests that although data science lacks the institutional infrastructure characteristic of canonical professions such as law and medicine, it nevertheless utilizes some form of abstract knowledge by which data scientists integrate skills also expected in distinct specializations. This brings us back to the question of how data science, emerging amid an economic setting that often scripts and thereby dominates expertise, could avoid singular organizational and institutional specialization and consistently integrate skills expected across heterogeneous problems.

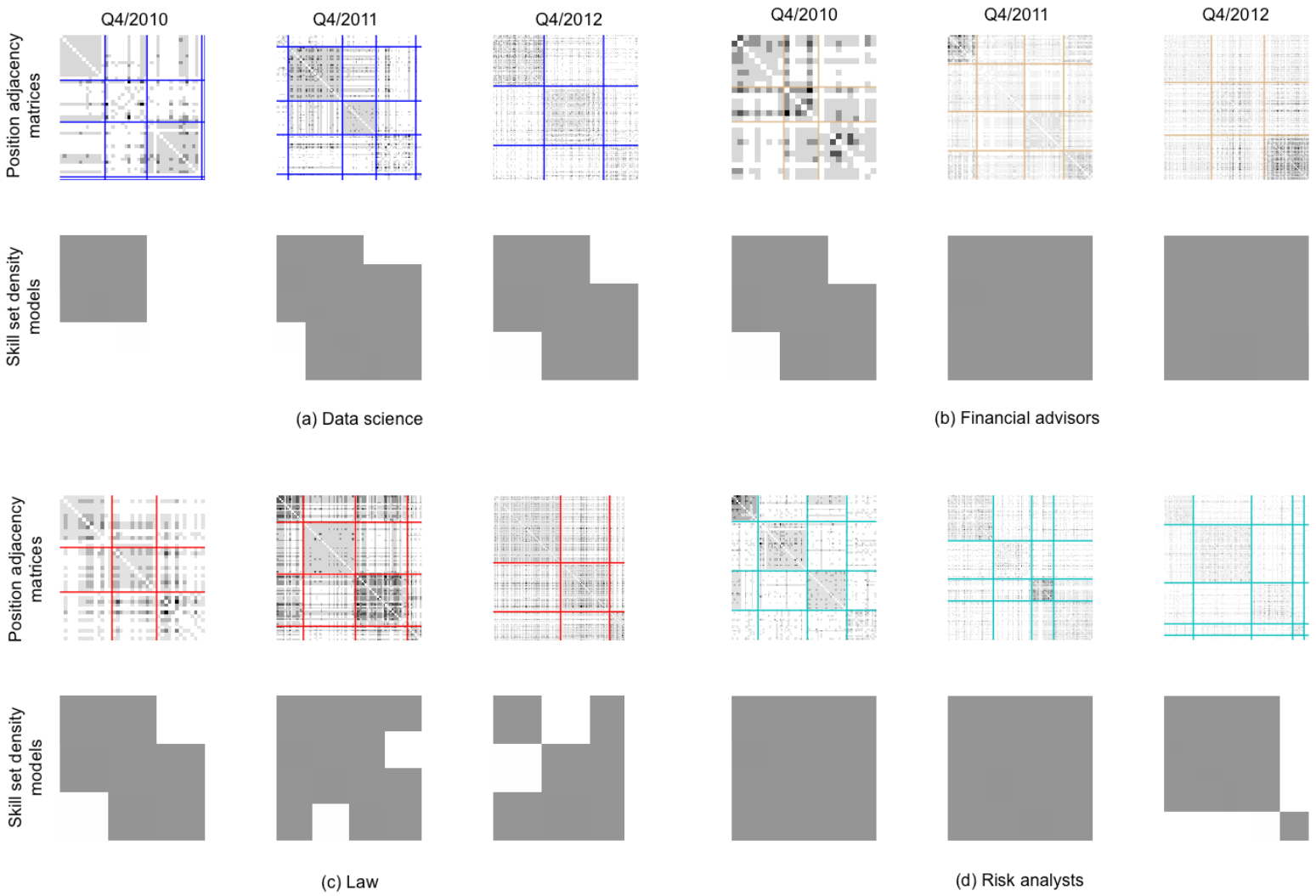

Figure 9.2. Structural representations of job descriptions based on skill co-occurrences (top row in each panel) and model of skill densities connecting these positions (bottom row in each panel) over three temporal moments (columns in each panel.

Organizations with varying characteristics expect data science skills (see tables A1.1 to A1.3 for more detailed statistics of the organizations advertising expert positions that require skills that index the four professions over the three moments shown in figure 9.2). The three cases I compare data science to utilize skills predominantly expected for positions in very large and relatively old organizations (with risk analysis in an intermediary age group). Job postings expecting skills that index law positions stand out with respect to partnerships as organizational forms they are located in. That these are the legal profession's primary organizational form is well known (Heinz and Laumann 1982). Organizations offering positions that expect skills relevant for risk analysts also show a distinct feature. They tend to be neither privately held organizations nor traded on stock exchanges (see table 9.4). This somewhat unintuitive combination is typical of divisions owned by larger corporations (which, in turn, are traded (Rubin 2006)), a common arrangement for banks and is thus consistent with the credit focus of risk analysts. Finally, organizations offering positions with skills of data scientists tend to be younger (founded in the 1990s), smaller, and less often traded on stock exchanges (see tables 9.5 and 9.6). This result is important. The analysis begins with a focus on skill sets that index expert knowledge. The types of organizations that offer positions the subsequent analysis then discovered those skill sets in map on the characteristics that also appear in accounts of those expert groups. Moreover, data science skills appear in a set of organizations with properties that most likely facilitate novel skill combinations, as we could see in the New York City presentations.

Just focusing on the last period, by which data science had become publicly recognized, data science skills are combined in ways that resemble the structure of legal expertise, which draws on abstract knowledge. Younger organizations let data scientists recombine older skills in novel ways. They are less equipped to dominate expert skills is no surprise (Saxenian 1996). As data science draws on much older fields of expertise, it remains unclear here whether the skill contours result from adding knowledge from one area to knowledge of another, or whether data science expertise integrates them in distinct ways.[95] The following section considers these processes.

[95] This question reiterates much of chapter eight in this applied setting.

Skill and knowledge sequences

The previous section has shown comparatively that skills expected of data scientists fold into contours that signal abstract knowledge. This could be seen as data science skills apply across heterogeneous contexts, similar to legal skills. Law's abstract foundation is widely established (Stinchcombe 2001, Abbott 1988). Considering historical context, the data science skill arrangements could result from several processes. To recall, the most direct processes consists of an additive effect, or ongoing specialization, where skills get added to the existing stock of knowledge. This could unfold in either of the two fields data science most significantly draws on, statistics and computer science. Earlier instances in which arcane knowledge gained salience suggest an alternative process of structural rearrangement, or integration, of expertise. Finally, formal organizations prominently compartmentalize knowledge (Abbott 1988) or facilitate its flow (Saxenian 1996). I therefore explore processes of expert knowledge construction in their respective organizational contexts in order to ask how skill sets for distinct problem areas relate to one another, such as quantitative analysis and software engineering in data science.

Statistics and computer science each address important problems, indicated by their academic status. Combining the two would plausibly enhance the joint group's salience, in quantitative terms. Yet both are known for their tendency toward specialization, which foregoes general recognition. Quantitative analysts, as discussed above, commonly focus on the organizational contexts generating the data they explore (Porter 1996). Programmers similarly specialize in their organizations' problems (Ensmenger 2010, Kraft 1977). There is thus no direct connection from their respective specialized expertise to abstract knowledge that might integrate them, considered as the foundation for a distinct role for other professions. The latter would be captured by an alternative process in which data science is associated with distinct knowledge that tells the data nerd how to integrate and apply those separate skill sets.

Here I recall examples from the conceptual discussion in order to once again clarify such an integrative process. Each reveals variation between expertise legacies and applications, that is, old and new ways of seeing problems. We can recall Brown (2000), who documents this tension in antiquity, when the Catholic Church, in its inception, took advantage of older, profane traditions. In a much more recent struggle over the treatment of mental problems, which the clergy had established itself to address,

psychology, psychiatry and related movements have convinced the lay public of a more nuanced perspective following medical and scientific principles (Abbott 1988). None of this process combines knowledge additively. They instead integrate practices and problems by revealing nuances and novel connections within established, singular views. Catholic belief deeply disagreed with the grounds of the pagan practices it nevertheless integrated. Likewise, psychologists and other movements that displaced the clergy in treating mental problems found leverage in seeking variation instead of pursuing centralized explanations, as theology offered (Abbott 1988). Law, finally, has in appellate courts institutionalized the very process of rearranging expertise (Holmes 2009, Stinchcombe 2001). These instances represent processes that instead of adding specific expertise to established knowledge arrange and integrate that knowledge such that they preserve heterogeneity of the underlying substance but sacrifice specialization.

For the data science case we can explore both additive and integrative processes, and their interaction with organizational settings, in a series of specific tests. Focusing on the most recent period, the comparisons in the previous section have revealed contours of abstract knowledge in data science skills. Observing similar arrangements in the initial period, when data science was not yet widely salient, would lend empirical support to the additive process, as it would show that the differentiation is preexisting. Because we know that data science draws on knowledge from older fields, that would indicate that the novel label maps onto additions to specializations, such as a "would-be" version of data science in part I. Shifting combinations of skill expectations, on the other hand, would support an account of data science seen to offer a distinct form of knowledge, beyond a novel skill addition. Specifically, this conclusion would find further support in evidence revealing data science skill rearrangements with associations between the earlier arrangement and specialized knowledge and abstract knowledge later on. Understating either process, finally, requires accounting for the role of organizations.

Comparing contours of legal skills and data science skills provides evidence of rearrangements of skills in data science and of utilizing abstract knowledge. The legal skills structure remains unchanged over the three moments, with respect to bridging distinct specializations (see empty blocks in all three moments in figure 2c). The arrangement of data science undergoes change toward resembling that of legal skills, in this structural respect (see empty blocks in moments 2 and 3 in figure 2a). Law's long history, and the short window considered here, lead to expect general continuity of arrangements, which

we find. The specific distributions of skill arrangements vary across the three moments, seen in the position and size of clusters (figure 2c) and distribution of organizational characteristics seeking respective skills (see appendix 1, tables A1.1 to A1.3). In the last period clustering of positions in large from medium–large organizations, and smaller organizations bridging between the two most clearly induces heterogeneity in skill arrangements. This division of specialization is consistent with the those generally observed in law (Abbott 1988, Heinz and Laumann 1982). Although counterintuitive, this combination of situational variability and general stability captures the source of professional autonomy from abstract knowledge (Merton 1968b) as it is structured along client types, but not dominated by any single one of them.

As indicated briefly, data science skills fold into different arrangements across the three periods. The first period we see in figure 9.2a shows none of the characteristic contours indicating the presence of abstract knowledge, seen in legal skills. The second and third periods do. This sequence of rearrangements and integration signals the emergence of abstract knowledge. This conclusion finds further support in the following comparisons of this trajectory to the other two cases with specialized expertise, as well as when considering its organizational context, considered thereafter.

Comparing data science skill arrangements to those of financial advisors and risk analysts reveals patterns signaling initial specialization of data science skills. Both financial advisors and risk analysts remain relatively stable in their lack of heterogeneity, if in subtly different ways. Data science skill arrangements, at the same time, move away from the shape of both. Considering these sequences in detail for financial advisors, while initially showing variation,[96] skills later on cluster in a singular core. They still apply to other problems, but all specializations overlap, without empty blocks, or meaningful heterogeneity. Financial advisor skills cluster in those centralized periods most densely in the context of large and old organizations (in both 2011 and 2012, see table A1.1 and A1.2 in the appendix). This observation is consistent with a widely reported transition in the financial advisor profession in which banks and other large organizations have taken much of the work previously provided by smaller operations. This collapse of the overall skill set into an undifferentiated structure contrasts abstract knowledge and singularly specialized expertise. Risk analysts also show relatively homogeneous overlap.

[96] A lack of observations for initial period (potentially related to the aftermath of the global financial crisis affecting that period) prevents interpretation.

Their specializations, however, preserve more variation, as the last period indicates in particular. Risk analyst skills cluster in large publicly owned organizations. Some skill sets also cluster in smaller, younger and more diverse contexts. They remain looser and sustain no noticeable overlap, relative to the dominance of the core. Although both financial advising and risk analysis skills vary in their specific composition of the stock of knowledge they instantiate, they share centralization and dominance of relatively old organizations of that skill centralization. Organizations with features that indicate rigid boundaries share skill expectations most clearly.

The comparisons so far clearly suggest that data science is seen to integrate knowledge. The details of this integration, particularly with respect to the organizations surrounding it, require clarification. The rearrangement of data science skills unfolds across varying organizational contexts, favorable and hostile. Positions in smaller organizations that are densely connected eventually (Q4/2012)[97] were only weakly connected in the first period (2010). While most positions overlap, this particular cluster hardly shares any skills with positions in large public organizations (Q4/2012). Importantly, expectations of organizations of intermediary size bridge between the two otherwise unconnected specializations. Knowledge integrating skills that are distinctively associated with data science thus transcends clusters of organizations that are otherwise seen to dominate expertise, even in risk analysis, another instance of quantitatively oriented expertise. The two differ in data science's use of programming. Although programming is often associated with specialization, as seen above, we have observed throughout part one how the integration with data analysis facilitates a different process. Open source technologies for vast data–storage that have emerged since 2005[98] together with cheap storage space have opened up resources to small organizations that were previously under the control of large bureaucracies.[99] And we have already considered how for instance Linus Torvalds's earlier Linux project created a movement around these technologies. These developments facilitate the analysis of previously inaccessible data structures. Specialization seen elsewhere is here undermined by the generally shared mathematical formalization of data analysis. This shows that the abstract knowledge data science skills fold into is

[97] All moments refer to the three columns in each panel of figure 9.2. The blocks in figure 9.2 correspond to the rows in tables A1.1 to A1.3 in the appendix.
[98] This was the year when Hadoop started, an open-source programming language for distributed data storage.
[99] The conditions under which this claim is true requires a separate discussion. Amazon, for example, offers both cloud storage and computing solutions at low costs and risk. While this facilitates smaller projects outside of bureaucratic entities, it also creates new dependencies through the market.

consistent with those observed elsewhere, although technologically based expertise is better known for its tendency to lead to specialized knowledge. Abstract knowledge here bears on the process of integrating technology and analytical methods and strategies.

These temporal comparisons reveal that data science is seen to integrate expertise in statistics and computer science such that it addresses heterogeneous problem areas, instead of merely adding skill sets from one to the other. This complements the earlier finding that data science skill expectations reveal contours of abstract knowledge by demonstrating that this abstract knowledge is distinct for data science, not the areas it draws on. Moreover, data science skill arrangements take advantage of supportive organizational contexts and integrate skills across organizations with characteristics that in both financial advising and risk analysis consolidated homogeneous skill clusters. We now have strong evidence that data science, albeit lacking formal boundaries and just emerging from existing knowledge that is known for its specialization as a thought community, integrates heterogeneous problems with distinct abstract knowledge that give rise to generally recognizable contours. These findings have implications for our understanding of expert work and occupations, specifically as expert work shifts more intensely toward technologically based problems.

9.4 Discussion

9.4.1 Summary

This analysis has shown that data science identification mechanisms combine familiar and established skills in novel ways. Its general salience can be explained by the contours of the stock of knowledge that integrates this distinct skill set. This conclusion follows from a comparison of data science to law and other expert groups. Data science and legal skills have in common that they are expected to be transferrable across distinct specializations. Financial advisors and risk analysts, on the contrary, focus on a singular class of problems. Public expectations and salience are thus not the consequence of specialized or general skills, but of the abstract knowledge to transpose them across problems.

I was able to discover this by studying large samples of job descriptions with methods from textual and network analysis. Job descriptions connect the nuances of specific positions to widely used labels for expert work. The analytical strategy moreover identifies descriptions on the basis of specific skill identification they articulate and independent of their label. This results in a dataset that directly captures

the tension between publicly salient expert groups and their constituting skills and knowledge. Data science reveals the effect of skills and knowledge independent of institutional boundaries more clearly than familiar cases do.

Three main contributions follow.

9.4.2 Contributions

Data science reveals the formation of a thought community of nerds in problems pertaining to modern technologies. Professions, and expert groups more generally, deeply impact society. This is obvious for law and medicine, and understood for many more obscure technical experts. We saw this for data science with the shopping chain and Facebook incidents in the introduction. They also remain few in number. Hence observing their emergence is difficult. The opportunity in the recent formation of data science allows viewing the effect of knowledge independent of the institutional status. Also from this perspective where we focus on organizations seeking employees, data science still undermines formal boundaries. Evidence also suggests, however, that instead data nerds navigate potentially informal paths as they utilize basic coding skills to address a broad range of problems. The knowledge they solve specialized problems with operates on the level of the group; it is abstract.

Abstract knowledge can be indexed empirically and compared across expert groups. The existing literature considers abstract knowledge as a necessary feature for achieving professional status but has presented a variety of empirical descriptions with unreliable implications. Instead of inferring abstract knowledge from the presence of some formal marker or informal process, this strategy induces it empirically by considering the transposition of skills across different specializations, and the contours resulting from this.

Public salience of arcane expertise is a question of contours, not content. By contours I mean the way in which knowledge is seen to apply to problems, as opposed to the type, content or relevance of singular problems. This one here resembles the improvisation control we found in particularly rich terms in six three. Similar to this analysis here, chapter six focused on accounts of data science skills. As discussed above, the problem of publicly salient objects is often approached as one of identifying the nuances. Such a view stops short of considering how nuanced perspectives and experiences cluster around broadly salient cultural entities. This analysis reveals one such process by showing the

relationship between skills with specific utility and knowledge of their abstract relations. Salience results from heterogeneous application.

9.4.3 Limitations

A focus on job descriptions ignores the richness we could see in part one. Descriptions from job postings directly capture the interface between clients and experts and their formal and informal features but offer limited depth. Legal knowledge is clearly more complicated than the paragraph in a job posting suggests. The question this chapter aims to address however pertains to public awareness of such arcane knowledge, for which its varying degrees of complexity are of secondary importance. The following chapter, with an analysis of academic publication patterns, is designed to capture more specificity.

The data only covers a short time period. Legal, medical and other expert knowledge is centuries-old, and even the foundations of data science expertise predates the period I consider here. The critical moment at which this knowledge has found public recognition, however, was not reached before the end of period studied here. Moreover, the main part of the analysis, which does not assume the data science label, initially found no combination of skills characteristic of it later on.

Finally, law and data science make for an asynchronous comparison. Law has long passed the moment of inception we observe data science in. This is controlled for to the extent that the analysis focuses on knowledge, not on formal entities. Expertise constituting data science skills predates the formation of this expert group. The legal institution could confound these conclusions indirectly by defining legal knowledge. Yet institutionalization does not necessarily lead to the shape legal knowledge takes, as the comparison to other cases shows.

9.4.4 Implications

Data science has a robust foundation. The main analysis revealed contours of data science knowledge that resemble those of legal knowledge, in spite of data science's lack of formal boundaries and coordination. Evidence showing that data science addresses multiple specialized problems suggests that the group of experts with that knowledge will remain relevant independent of the popular debate around the term data science. As part one has suggested as well, this group establishes itself less by

agreement on a common label and more through the arcane knowledge with which they transpose specialized skills across heterogeneous problems.

Finally, this comparative analysis points out more clearly that data science has gained distinct salience in a different setting compared to familiar professions and expert groups, but one that is likely to have more expert groups form in the near future. Lawyers and accountants flourished with the rise of commerce and large bureaucracies. Many other groups benefited from growing academic sciences. Data scientists work in a context of cheaply available, widely diffused and powerful computer processing and information technology. Access to relevant knowledge is easier but its effective composition and utility remains less clear than in formal training programs. Data science reveals a distinct effect of constructing knowledge that sacrifices depth in the interest of breadth, and shows how modern technology facilitates such structures. This finding suggests that understanding experts in problems of modern technology—such as application developers, user interface designers and the actors in the sharing economy—requires a focus on their engagement with knowledge directly, instead of resorting to formal entities or informal relationships. This strategy responds to more accessible and fluid labor markets. Required education levels initially channel its benefits a small segment of the labor force. Yet as these opportunities widen amid easier access it is critical to understand their dynamics in order for workers to navigate them by acquiring appropriate expertise.

10 Mechanisms in the Emergence of Data Science: A Comparative Study of Abstract Knowledge Construction

In order to address the most fundamental question of expert work, I now ask what kinds of problems nerds solve in the scientific setting. The academic context is critical for understanding knowledge production today. Moreover, science came up in part I for its original definition of data science and rule over its tasks. This question first came up in abstract terms when, toward the end of part one, Donoho compared data science's "collection of technical activities" to a "continually evolving, evidence-based approach," which would be scientific of lead to entitlement. In other words, there is a strong sense of what kinds of problems scientists and professionals address. By focusing on data science as a group of experts just starting to emerge, this chapter considers how it comes about that thought communities are able to define contours of knowledge such that Berger and Luckmann's (1966) observation that "I `know better' than to tell my doctor about my investment problems, my lawyer about my ulcer pains, or my accountant about my quest for religious truth" seems self-evident to the public. That the obvious knowledge of whom to consult is the product of historical struggle is well known. Less than a century ago, to recall, people concerned about their depression or other mental problems would have known better than to consult a physician. Yet today such consultation is obvious because academics with backgrounds in neurology, psychiatry, psychology, and psychotherapy, among other disciplines, came to control the treatment of mental problems by relating their contributions to existing medical knowledge thereby gaining support from established medical fields. With that support they successfully challenged the clergy's authority over this area (Abbott 1988). Similarly, the church itself first secured its status by integrating problems and strategies others had already claimed as their own (Brown 1992, 1982).

To understand how stocks of knowledge and practices are associated with sciences or professions we need to watch the process unfold in a comparative context. Here I take once again a comparative perspective and consider the emergence and contours of data science, to processes underlying law and systems biology. Law, once more offering a baseline for salient and autonomous work with arcane knowledge, is seen to address lay problems and systems biology is seen to address arcane scientific concerns that while possibly significant for non-scientists, largely look irrelevant. Data science is seen to

be both irrelevant, as systems biology, and, as law, directly salient for practical problems. By focusing on the ways in which data science has emerged, this chapter positions identification with this new thought community relative to familiar cases, sheds new light on the nature of expertise and the foundation for professional knowledge, and considers the association between technology-based knowledge and the institutional and informal processes underlying disciplines and expert groups.

With respect to this specific approach, the literature on professions and knowledge either focuses on the formal features of expert groups or on their informal specificities. The strategy undertaken in this chapter is different. Here I focus on the variation of principles of expertise—the patterns by which experts and scholars engage with existing knowledge as a way of identifying with a thought community. This enables me to account for lay recognition of arcane knowledge on the basis of the principles organizing that knowledge as potentially independent of its institutional scaffolding or informal processes. In the substantively much richer setting of New York City's data science events we observed data nerds illustrate arcane technical capabilities in vivid terms, persuade formal hierarchies of their utility, share their struggle that is part of defining this work, and finally, perform the sciences. The present chapter focuses on this last identification mechanism. Whereas at the data science events we have seen relatively crude, though not unsubstantiated, references to classical scholars and ideas, here I consider with more detail the kinds of such references scholars associate with data science make, the contours emerging from them.

Let me first revisit the debate that concerns knowledge production for that purpose. On one side are those who say scholars produce knowledge as they aim to discover universal truths about the world in contrast to those who say truth, and hence knowledge of it, is locally bound within the groups who define it. The former show how the role of the scientist has diffused through societies (Ben-David 1971), that disciplines agree on the quality of contributions (Cole and Cole 1973) and that substantive consensus manifests itself globally (Barabási and Albert 1999, Moody 2004, Merton 1968a). Others challenge these arguments with evidence, from field settings (Wynne 1992) and laboratories (Collins 2004, 1998), that demonstrates the incomprehensibility and limitations of knowledge beyond the small groups of experts producing it. With the alternative thought community framing I position myself between these sides of the debate by analyzing strategies that the emerging data science movement deploys to engage with the

existing stock of knowledge and how they compare to institutionalized practices. I model my analytical strategy on a recent effort, which the introduction touched on already, to bridge the existing gap by studying the formal patterns by which informal debates unfold (Shwed and Bearman 2010). I furthermore design simulations as part of applying existing approach to the data science puzzle. Results from those simulations that assume generally shared recognition of important contributions leave critical features of the observed data science structure unexplained. At the same time they also reject an explanation primarily based on local settings, as they provide evidence to show that scholars with different backgrounds and affiliations work toward stitching together coherent contours around a heterogeneous body of knowledge. Qualitative accounts reveal strategies, which can be both scripted and informal, for approaching problems by analogically considering how they unfold in different contexts.

As we have considered data science mostly in applied settings so far, we should also recall the other part of the debate, which concerns applied knowledge. The previous divide holds here as well, with one side are those who associate expert knowledge with institutionalized professions delivering it to the lay public, in contrast to those who do not find explanatory power in institutionalized entities and instead see knowledge develop from informal interactions among specialized experts and often together with their lay clients. The prime examples illustrating the institutionalist side come from studies of the medical and legal occupations (e.g., Freidson 2006, 1961, Abbott 1988), and more recently of economists (Fourcade 2009, 2006). Members of the relational side find support for their arguments, that to study knowledge one needs to account for informal relationships between all involved actors, in the case of the autism epidemic (Eyal et al. 2010) and the HIV movement (Epstein 1996), among others. Consistent with the observations in part one, this analysis reveals complementarities. For law it discovers stocks of knowledge that correlate with formal institutional channels. Data science experts subvert institutional contexts to build their stock of knowledge and we need to consider patterns by which they deviate. While not correlating with institutional arrangements, they nevertheless resemble a critical feature of the stock of legal knowledge in how they integrate diverse problems. Their contours differ from the purer scientific approaches we can find in systems biology.

This analysis addresses the practical problems data science confronts the public with. The previous settings have consistently revealed the limited bearing of bureaucratic control over data science

work. We discovered that data nerds enact disciplinary forms we otherwise know from academic sciences. Accordingly, I consider such a setting here directly, and those activities of scholars who would, and have, pass the test for scientific “entitlement,” by Donoho’s standards. By considering here contours of the stock of data science knowledge relative to the established fields, which have been exposed to disciplinary control much longer, we can learn forms of control for applied data science tasks where neither conventional bureaucratic, nor heterarchical or other forms of informal control fully defined tasks, and the formal academic institutions have no bearing. By studying data science in the academic context, we can most clearly consider how it relates to other fields with similarly opaque forms of control, yet with institutionalized ways of engaging them in lay concerns.

10.1 Constituent elements of professions and expert groups

Let us then return to the relevant literature in order to consider the specific principles by which groups construct arcane knowledge. On a general level we have said that while the focus on abstract knowledge leads to an emphasis on the formal boundaries of knowledge (Freidson 1986) and the focus on expertise leads to an emphasis on informal interactions underlying it (Eyal et al. 2010, Eyal 2013, Collins and Evans 2007), both features co-occur empirically. What constitutes abstract knowledge (and how to identify it) is not so easy to establish. Some scholars (e.g., Larsson 1977) have argued that abstract knowledge can be indexed by university training; yet an array of engineering occupations that have failed to professionalize require university training as well (Abbott 1988). Likewise, another group of scholars (Svensson 1990) argues that abstract knowledge is indexed by the presence of a coherent underlying theory. While this may account for the low status of social work—the case for which this argument is frequently made—it does not account for the high status of law (Abbott 1988). Abstract knowledge and theory, as source for professional autonomy, account for some important cases, but not for others.

If abstract knowledge is hard to measure, expertise by definition—seen to emerge from the interaction of actors, devices, concepts and institutions (Eyal 2013)—is even more complex. In Eyal's interpretation of expertise, for example, knowledge attainment and training are tightly coupled; physicians trained themselves in treating autism at the same moment that they developed the conceptual framework to understand it (Eyal 2010). Considering another context, Collins and Evans (2002) argue that expertise

arises from largely unscripted interactions—for example, mentor-student relationships in academic programs—in organized settings such as research laboratories. Whatever their differences with respect to the internal processes by which expertise is captured, expertise scholars insist that a crucial element of the assemblage of expertise involves the mobilization of non-expert, typically client, knowledge. Thus, for example, physician expertise on autism is seen to necessarily invoke parental observations (Eyal 2010) just as knowledge about HIV etiology and treatment arose from HIV+ individuals in the early years of the AIDS epidemic (Epstein 1996). In this context, Collins and Evans (2007) explicitly demonstrate how informally acquired expertise is translated into the formal, codified language that members of the expert community recognize. While this work is evocative, there are no clear patterns that one can distill with sufficient precision to have confidence that their presence indicates that professional recognition has been achieved.

Taken together, we can note that informal expertise and formal knowledge are present in all instances, whether they are large associations or small labs, and knowledge is always conveyed through formal training and informal mentoring. The different choices of research problems implicitly suggest that for some questions expertise offers more leverage, and for others abstract knowledge, while failing to specify their overlap and interaction. Against that background, it follows that neither abstract knowledge nor expertise, the key sociological concepts in the literature, directly help to understand why the lay public knows what a lawyer is useful for but not a systems biologist.

A useful measure needs to capture stocks of knowledge and their contours directly through the principles by which experts identify them. To recall, by identification principles I mean the ways in which new contributions are seen to follow from existing ideas. Building on the consensus that formal and informal expertise involve abstractions, which we saw in chapter eight, I ask here on a more comprehensive level how scholars arrange problems through abstract connections, and what contours stocks of knowledge subsequently take such that some become obvious.

Roadmap

In section 1, I describe the three cases with respect to how they address problems as fields of expertise by presenting examples of the ways in which members in each field engage with and thereby construct its underlying stock of knowledge. In section 2, I build an image of each field from the principles

of expertise that are implicitly and explicitly articulated as actors identify with their stocks of knowledge. In section 3, I turn to data and methods. In order to identify relevant actors and activities in each field, and to capture the institutional forces that help shape those fields and their contours, I select a sample of academic training programs in each of them, identify all the affiliated faculty, their publications and references to existing work. From these citations, I construct co-reference networks in order to identify how experts in each field identify relevant pieces in their existing body of knowledge. Some areas of expertise are less institutionally structured than others. In section 4, I argue that overall fragmentation of literature, preferential attachment in citation patterns and journal position in the citation structure are signs of institutionally scripted academic work, or expertise. In order to analyze how institutional forces shape stocks of knowledge (and therefore visibility to the lay public) relative to informal processes, I simulate the preferential attachment dynamics for each field and analyze their importance for reproducing observed fragmentation. In addition to analyzing the simulations, I draw on observational evidence in order to consider the substantive contexts in which these patterns unfold. In section 5, I discuss how the dynamics observed in the previous section help us understand how neither institutional contexts and channels, nor informal processes consistently explain salient contours of expertise, how it is instead the identification principles by which scholars use abstractions to construct their stocks of knowledge because they produce overall contours that make their utility obvious to the public, and how this approach helps us address a range of important problems that were previously not considered as questions of knowledge or expertise.

10.2 Legal scholarship, systems biology and data science

The individual scope of the three fields makes it impossible to describe them in detail. I therefore begin to approach each from a specific instance that is representative of key processes with respect to the constituting principles of abstract knowledge as interaction of formal and informal knowledge. Although we have considered law and data science before, it is still useful to consider them again specifically with respect to their appearance in this academic setting. The academic institution defines the kind of practices they engage in, such as publishing and citing work, even if the principles of how to identify relevant work, and the contours of resulting stock of knowledge, unfold differently. I consider law as it worked out the implications of the Internet for legal thinking. As data science is associated with the

formation of digital social networks, as we could see in the Facebook incidence early on, I analyze one of the earliest academic efforts studying such data. Relevant academic processes are well documented for scientific disciplines, such as systems biology. I therefore summarize them here and provide a richer illustration through the case of the Human Genome Project (HGP) in the appendix. This section describes how scholars relate new problems to the existing body of knowledge in order to discern how the principles by which they do so might shape the stock of knowledge.

Support for both sides of the debate of scientific knowledge production, outlined above, can be seen in the case of systems biology and specifically the HGP. To recall, one side emphasizes the role of general recognition of important contributions and their guidance of subsequent work (Foster, Rzhetsky, and Evans 2015, Merton 1996, Cole and Cole 1973). The other side focuses on the importance of research labs, mentorship lineages and other informal relationships (Latour and Woolgar 1986, Collins 1998). In both processes scholars tend to produce knowledge by filling gaps within recognized problems—systems biology asks what constitutes life—either defined by their field or local community (see section A2 in the appendix for a detailed description of these processes in systems biology).

How these processes unfold in legal scholarship and among data science scholars is less well documented.

10.2.1 Law

Lawyers draw on abstract knowledge to address general problems. Law's historical richness complicates a representation of its stock of knowledge through a specific instance. To begin with, the legal field varies across countries (Stinchcombe 2001).[100] Within regions it has also changed its characteristics over time, for example when adopted from antique Roman roots and again with its adjustments to sprawling commerce, both in Europe (Weber 1976) and the US (Friedman 2005). Yet there are also continuities, such as the idea that legal knowledge has to reflect actual experiences (Holmes 2009), which can be seen in specific instances.

Here I consider the legal implications of the Internet. The patterns I reveal in the following summary could also be found in legal concerns with other issues such as the war on terror, privacy and dignity or market competition (Fallon and Meltzer 2007, Whitman 2004, Brummer 2008). They also mirror, in legal

[100] This analysis strictly focuses on the US.

scholarship, the principles of analogical reasoning that lawyers apply in practice (Stinchcombe 2001, ch.4). Commenting on their own field, legal scholars have made similar observations in as weighty issues as the struggle around contract theory (Gilmore 1995), appellate courts (Llewellyn 1960) and the common law itself (Holmes 2009). I chose the Internet case for its scope, timeliness and substantive relevance.

The Internet debate reveals how legal scholars deploy analogical reasoning, in two ways. Scholars first debated whether the Internet could be regulated, and hence fall under legal jurisdiction. If it did, scholars furthermore needed to consider whether those existing regulations applied to the Internet in the same way they did elsewhere (Lessig 1999, 506). The debate rested on comparisons between problems in the digital and in the physical world.

The Internet was seen to resemble a series of physical architectures, from a legal perspective. The debate described digital and physical problems of zoning to illustrate challenges with translating existing regulation into online regulation. One example emphasized that the physical world made zoning regulation functional because it was easy to rely on visual indicators for regulating access, such as keeping minors from adult material. It pointed out that the Internet denied these intuitive mechanisms. It was possible, however, to design browsers and servers such that they recognized relevant indicators, and could block access (Lessig 1999, 510). Shopping in real-world stores and through online browsers are vastly different experiences in practice, but related in legal analyses.

Comparisons also directed the legal debate toward finding differences. The physical constraints non-digital contexts impose led law long ago to balance demands for free speech with those for private property. Internet code, however, could be designed such that it was agnostic to these conflicting demands. The legal debate invoked the imaginary of physical transitions, such as turning “airplanes into skyscrapers,” to illustrate the much less visible, yet related implications of different applications of private property to online code (Lessig 1999, 526). Hence this argument relied on analogically comparing problems that are familiar to an outsider, but rarely considered together.

Both analytical moves connected substantively unrelated problems through legal abstractions. This can be seen in the example of access to Internet code and skyscrapers, as well as in that of websites and grocery stores. Abstract concepts, such as zoning and private property, connect empirically distinct problems by revealing legal analogies. These relationships leverage old answers for new problems. They

may also uncover gaps relevant to their original applications. Filling these gaps makes the abstract codifications of property and privacy more robust (Stinchcombe 2001).

Alternative strategies would seek to understand specificities relative to similar problems and assess how both inform a larger question. In law such a strategy might consider how the problems of protection of web code and exposure to online content relate to a larger theory of justice in the digital world. This is not the strategy we have seen here, but one that is common in other academic fields, such as the systems biology case summarized above (and described in the appendix 2).

10.2.2 Data science

Like law, data science addresses general problems as we have seen in the troubling incidences of shopping coupons, but also practical ones such as internal scheduling, romantic dating and professional careers. The knowledge that facilitated all these application, however, emerges from a scientific context, as systems biology. Many applied data nerds acknowledged this but often went more into the details of the practical circumstances of those ideas, as the technical ones seemed obvious to them anyway. Focusing now on the purely academic setting, we need to consider that unlike systems biology, however, members from the established disciplines that constitute data science expertise have not split from their colleagues' academic orientation. The creation of educational programs instead reflects a move in which universities recombine expertise of scholars from separate fields who are seen to address data science problems, while not necessarily inflicting on their original disciplinary affiliations. Data science thus complicates several explanations.

Data science, and Donoho's trouble with it, shows that scientific origins need not lead to stocks of knowledge seen to focus on arcane problems. It also challenges and juxtaposes the two main views in the literature more clearly. Its imposition on existing fields signals a break with institutional forces. Yet its rapid spread throughout universities across the US simultaneously questions the importance of informal processes, often proposed as an alternative explanation. Data science therefore allows us to observe the constituting identification principles by which experts engage with and develop their stock of knowledge without immediately leading us to resort to either side of the currently proposed explanations.

Data science's academic roots reach into rich pasts, which, as we could see in part I, also applied data nerds are aware of, acknowledge, and remind each other of, as we saw in chapter eight.

Quantitative analysis, which data science most significantly draws on, has the longest academic tradition. It has also been widely studied (Donoho 2015, Porter 1986, MacKenzie 1981, 1978). One important finding in this literature has been that quantitative methods, once diffused, compartmentalize in substantive applications and specialties and experts using them there remain isolated from one another (Porter 1995). Artificial intelligence, also at the center of data science today, historically constitutes a separate lineage, closer to software engineering than to statistical modeling (Feigenbaum and McCorduck 1983). More recent technological change toward routine data collection and cheaper and more powerful processing have facilitated that experts in the respective areas more directly relate to each other. This can be seen in several events throughout the years before the design of formal programs around data science (e.g., 2006, Batagelj et al. 2006). While illuminating, these historical roots misrepresent data science's distinctiveness.

Moreover, data scientists may appear familiar because they use methods that are part of the social scientific cannon, which complicates discerning the principles at the basis of data science as well. Some of it is labeling. Variables are called "features," Type I and II error become "precision" and "recall," and so on. Some methods are used differently. Often data scientists see no need to consider coefficients or confidence intervals when estimating regression models, which might seem odd to us. It seems odd to data scientists, in turn, to forego performance in order to preserve interpretability. Most significantly, data scientists utilize a vast body of methods and data structures unfamiliar to most social scientists.

This general description situates data science in the present academic context. Now I focus on a specific instance, from quantitative network analysis, to uncover some of its principles.

Considering networks as analytical objects has a long tradition, dating back to the mathematician Leonard Euler's problem of crossing seven bridges (Dorogovtsev 2010). As other instances of quantitative work (see Porter 1995 for an overview), it has since developed in various contexts, ranging from mathematics to engineering and social sciences. Quantitative network analysis thus pursues no singular goal, which sets it apart from systems biology's efforts to understand how life unfolds. Innovative approaches nevertheless discovered that some of these different specializations related to one another (Watts and Strogatz 1998). This discovery was based on the premise that substantive contexts of social structure, biological systems or technological networks partially distort underlying relationships and

organize their randomness, and that formal comparisons could systematically reveal how these processes unfolded.

Mapping such heterogeneous substantive contexts onto each other was achieved with reference to some simple formal characteristics. Tests revealed, for example, that much of the complexity of each of these problems resulted from distances between their elements (path length) and their interconnectedness (clustering). The argument integrated otherwise unrelated problems by considering their abstract representations. Social structure is analogous to biological systems, in network terms.

This principle of leveraging abstract analogies between such different contexts reminds of practices seen before in the legal example. Social ties have as little in common with electrical power lines as cars and code. Neither comparison seeks deeper connections between the substantive problems. They contrast to systems biology, which debates technicalities around singular problems.

I now compare data science scholars to legal scholars and systems biologists with respect to how they engage with existing ideas and focusing on the shape of the stock of knowledge they thereby produce.

10.3 Conceptualization

Legal scholars, systems biologists and academics affiliated with data science follow some distinct identification principles as they construct their stocks of knowledge. Scholars in all three cases study important problems. Systems biologists go on solving remaining puzzles within them. Studying implications of the Internet for public and property law, or of small-worldness for power outages and disease spreading is counterproductive in the sense of systems biology because neither informs a larger problem. At the same time, by considering analogies scholars in law and data science discover new problems between larger concerns. I now extrapolate simplified contours of each field from the implicitly and explicitly articulated principles of expertise from the previous section in order to move from specific scholarly work to the contours of the overall stock of knowledge. As scientific work has been more accessible for the sociological research that I draw on than that of the job posters in the previous chapter, the underlying identification principles are more specific than those that emerged from skills structures in job descriptions.

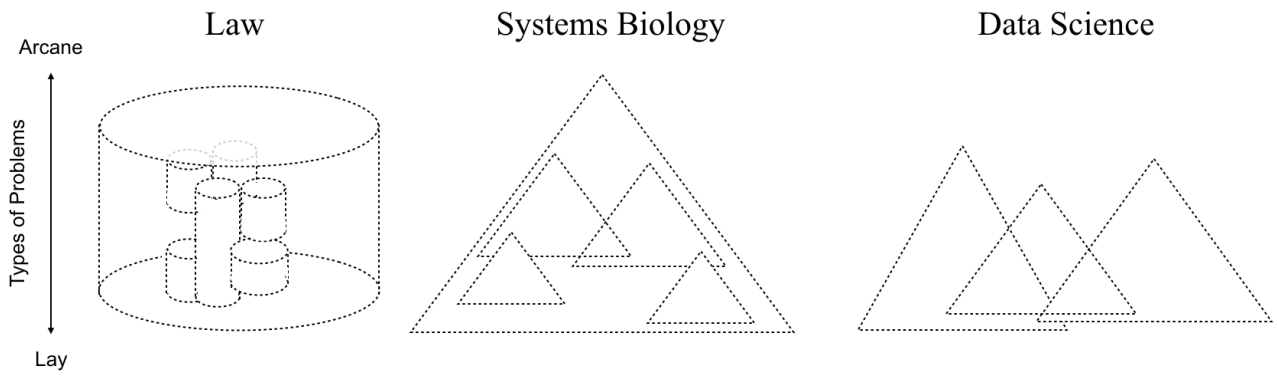

Figure 10.1. Contours of different stocks of knowledge, extrapolated from the qualitative accounts. Systems biology resembles most closely a hierarchical structure with a guiding aim (understand life) or theory (holism). Legal scholars relate different problems to one another by seeking analogical overlaps but no theoretical integration. Data science integrates multiple disciplines with scientifically organized problems.

Scientific strategies can be thought of as creating hierarchical structures that peak at the top in a guiding question or concern. Systems biology contributions aim to inform the specific question of what constitutes life. As other sciences, systems biology's larger question seeks answers in auxiliary projects, such as the HGP. The pyramid shape in the center of figure 10.1 illustrates this way of organizing knowledge. Moreover, research labs and mentor lineages frame auxiliary debates and address more detailed questions. Contributions focus on engaging with those niches and consider how their claim informs the larger interest. This generates some compartmentalization, but, importantly, different components relate to one another through hierarchical connections between the problems they inform (figure 10.1, center).

Legal contributions have no clear order. Legal concepts analogically relate practically disconnected applications to reinforce the concepts' universal status and to reveal remaining ambiguity. Cars and code have no relationship, except for the abstract concept of private property. Without guiding ideas, legal problems can be thought of as the unordered arrangement of overlapping cylinders in figure 10.1 (left).

Data science contributions, finally, uncover new questions between existing concerns. Experts analogically connect separate fields through formal characteristics of the empirical problems they are concerned with. The small-world concept roots in no singular idea and solves no particular problem, but features in many. Data science draws leverage from testing the applicability of its concepts to problems otherwise of concern in separate academic disciplines (figure 10.1, right).

Returning to the main question, these different contours suggest that legal experts and data science nerds are broadly salient because their analogy-based work leads them to address heterogeneous problems. System biologists look irrelevant from a lay perspective because their problems are seen relative to arcane ideas of what constitutes life, and less with respect to each other. It is not science's institutionalized boundaries that block arcane expertise from public view, but the principles by which scientists signal the identification with a thought community and its stocks of knowledge. This framing maps onto those then still tentative arguments in chapter eight that scientific entitlement follows from institutional status, not scientific status (although that would be one way). The comparative design of this chapter offers an analytical design to test this argument directly.

I now describe a dataset and methods to test the presence of these principles beyond the specific instances just considered, and to explore their prevalence relative to varying institutional contexts and informal agendas.

10.4 Quantitative design, data and methods

10.4.1 Analytical strategy

The conceptual framework around constituting principles of stocks of knowledge addresses both sides of the literature's main fault line. It allows considering the overall contours as either a correlate of institutional contexts or as following from informal processes. Both sides offer competing explanations, which I describe here and connect to the three cases to formulate empirical puzzles (see table 2 for a summary).

A plain institutional account leads us to expect principles of expertise that are scripted by formal entities, such as professional associations in this context. These scripts should apply across internal boundaries, such as universities here and organizations experts may work for in general. Essentially this view holds that because larger entities organize the production of expertise, they also index its distinctive characteristics (column 1 in table 10.2).

An argument based on plain informal expertise, on the contrary, would lead us to expect evidence of arrangements that deviate from scripted patterns as experts form knowledge within smaller units. This follows from the findings described above that show that developing knowledge requires trust and hence direct collaboration. Such processes are indexed by fragmentation of stocks of knowledge within the group or entity they are part of (column 2 in table 10.2).

Table 10.1. Overview over analytical strategy

Camps		Puzzles	Tests
Institutionalized Knowledge	Informal Expertise		
Scholars agree on which questions and problems constitute relevant knowledge in their field, regardless of the university or laboratory they are affiliated with.	Scholars identify relevant knowledge from their individual perspective, which is constituted by and diffuses through largely informal paths within research groups or through other personal relationships. Agreement on relevant knowledge across these settings is therefore unlikely.	*Substantive Consensus:* If relevant expertise is situated locally, members of a field should engage with different subsets of existing knowledge. If, on the other hand, relevant problems are agreed upon, this should be indexed by scholars engaging with overlapping stocks of existing knowledge.	*Modularity in observed networks* The degree to which academic fields agree on a body of knowledge, that is, its fragmentation, can be indexed by the modularity of their citation network.
Scholars also agree on how to address important problems. In science, the most frequently observed pattern for doing so is breaking up large problems into smaller puzzles. This practice generates the widely observed "Matthew Effects," in which already recognized contributions receive disproportionally more attention.	In addition to disagreement on substantive questions, scholars also vary with respect to the guiding ideas they build on.	*Conceptual Consensus:* If scholars structure knowledge according to generally shared principles, such as agreeing on important contributions, we should see a pattern as would be produced by the Matthew effect. If, on the contrary, local settings shape scholars' expertise, the overall structure should follow no systematic pattern.	*Modularity in simulated networks* The role of the Matthew effect in integrating knowledge can be indexed by estimating modularity scores of networks simulated through publications that preferentially cite existing work according to its recognition in prior citations. Partially random co-referencing patterns index unscripted citation practices.
Substantive activity of scholars is mostly organized by, or overlaps with, formal settings. In academic disciplines journals function to this effect.	Building expertise requires to engage directly with others who share similar concerns. Such activities can only be observed through actors that connect to others.	*Contextual Effects:* Institutions script interactions, for example through academic journals. Even if scholars have different perspectives on certain problems, the same set of journals should accommodate those positions. Informal expertise, on the contrary, leads scholars to construct a body of knowledge without relying on institutional infrastructure.	*Observed role of authors and journals* I consider the degree to which journals organize stocks of knowledge by constructing a network of publications citing work from the same journal, and estimating those networks' size-scaled modularity. I consider the degree to which authors address different communities, generated by the modularity estimation.

Table 10.2. Sample structure of archival records (instructor CVs and Web of Science)

Field	Institutions and centers	Instructors (without publications in %)	Publications (in 1,000)
Law	7	248 (30)	2.4
Systems Biology	15	196 (5)	12.3
Data Science	9	209 (26)	4.6

A focus on constituting principles and contours of thought communities, as I propose here, builds on both approaches. Instead of primarily focusing on formal entities, or informal deviation from them, looking directly at the way experts engage with knowledge allows us to see, for example, disagreement on guiding themes in spite of formal associations, or scripted patterns of expertise outside of institutionalized channels. This conceptual framework is designed to reconcile the historical continuity by which expert groups maintain lay salience with the extemporaneous change through which they claim it. Data science, just emerging in modern higher education from much older academic roots, reveals these processes empirically.

10.4.2 Sample scope and data

The academic setting allows juxtaposing the salient legal case and arcane sciences such as systems biology. They also accommodate both informal and institutional processes (e.g., Collins 1998, Ben-David 1971). Their main actors, scholars, furthermore leave abundant traces, in publications, from which I can directly reconstruct the principles by which they engage with existing expertise. Disciplines and training programs, such as JD education, systems biology PhD programs and data science degrees, formally organize underlying expertise. Yet variation within such formal orientations is possible and common (Lynn 2014, Abbott 2001). We can thus observe variation in constituting principles of stocks of knowledge empirically by studying what scholars cite in their contributions to them. To do this, of course, we need scholars. Hence I need to design a sample.

I first select samples of formal training programs for the three fields. I begin with the focal data science case and retrieve all programs that explicitly claim to prepare for data science careers. I identified

nine institutions with such programs in 2014.[101] Both legal scholarship and systems biology have longer traditions than data science, hence more programs and scholars. My question concerns expertise from its constituting principles and not the sheer magnitude of a field. I therefore design law and systems biology samples that approximate that of data science with respect to the number of instructors and the status of their institutions. Table 10.1 summarizes the data structure.[102]

10.4.3 Publication data

Studying stocks of knowledge from the perspective of their constituting principles, as I outlined above, entails that I consider how experts identify with a thought community as they construct their knowledge. I collected information on academic publications of scholars associated with the training programs just described. I first considered scholars' personally provided lists of published work. Next, I identified these publications in the Web of Science database and retrieved their complete lists of references.[103] The resulting dataset thus contains detailed information on the activities of scholars who are formally associated with institutional programs. These activities directly pertain to their identification mechanisms with the available stock of knowledge.

10.4.4 Methods

From this dataset I construct bipartite citation networks of publications and cited references. I project these bipartite networks such that two publications are connected to one another if they share one or more references. My main question is concerned with the ways in which academics, who are assigned to teach in a focal program, identify with the existing stock of knowledge in their own scholarly work. This data structure allows to test these processes in the context of the competing explanatory frameworks, summarized in table 2, empirically. I propose several measures to index the relative importance of institutional and informal processes and of constituting principles guiding scholarly contributions.

[101] The data was collected in the spring of 2014. Academics identified this way represent a selection of those who could legitimately teach in data science programs. Extending the group to include other scholars with relevant expertise would require coder judgment, however, and nullify the aim to study friction between formal recognition of stocks of knowledge, here through university programs, and constituting principles, observed through citation networks.

[102] JD programs have more instructors than data science programs, hence the discrepancy between the two despite the balanced sample on the level of scholars. Systems biology has the reverse relationship to data science. Also the number of instructors the program scope generated varies, but given the share of those strictly focusing on teaching, this has almost no bearing on the citation analysis. In fact, varying styles of writing papers (particularly long reviews in legal scholarship and short reports in systems biology) offset the instructor count effect. I control for the varying publication network sizes.

[103] The Web of Science database does not include books. This is unproblematic as neither data science instructors nor systems biologists tend to write books. The legal scholars who do, also write academic articles on those topics, which the database captures.

With the main focus on processes that integrate expertise, I first need to measure the diversity within a stock of knowledge. Following research on global properties of citation networks (Shwed and Bearman 2010), I estimate size-scaled modularity (Newman and Girvan 2004) to indicate fragmentation of a field.[104] Complete agreement, that is, all contributions citing overlapping previous work, would be indicated by a modularity of zero. The modularity score increases the more contributions reference work that fewer others cite such that higher modularity signals more diverse views. I compare the estimated scores between the three cases. The institutional side would lead to expect more agreement in law and systems biology than in data science, because the formers' more developed infrastructure of associations, journals, curricular and so on (see table 10.2, row 1).[105]

The institutional side also holds that a generally agreed-upon assessment of high quality work guides individual contributions. In academic disciplines, this way of organizing knowledge implies that new contributions are more likely to reference already highly cited pieces (Barabasi and Albert 1999, Cole and Cole 1973). This strategy captures the well-known Matthew Effect (Merton 1968a). I implement the preferential citation rule in simple simulations of contributions engaging with existing work based on its recognition, holding constant the number of contributions, references and citations of a given field. Estimated modularity scores of the resulting structures that differ from those of the observed structure could reflect observations of informal expertise in which knowledge is understood in its specific context (see table 10.2, row 2).

I then specify deviation from institutional processes. I reconcile the simulated with the observed structure by exposing the former to varying degrees of unscripted re-configuration, thereby specifying the magnitude of deviation from institutional processes. The degree of such additional manipulation initially captures both local activities that undermine formal institutional contexts as well as systematic processes. The relation of simulated to observed fragmentation together with the amount of manipulation indicate the

[104] This implementation here deviates slightly in that previous work focused on specific debates which emerge from directed interactions among participants. An analysis of stocks of knowledge a focal generation of scholars constructs focuses on outgoing ties (references) to existing work.

[105] This measure of the stock of knowledge is the same as the one I use in chapter nine. There I consider the communities this algorithm detects as part of estimating the modularity scores. Because we know more about academic practices than those of job posters, I rely here on the simulations, described above, in order to operationalize concrete principles of community identification, instead of inferring them from the observed structure.

importance of informal deviance relative to practices that are simultaneously unscripted and spread throughout a field (see table 10.2, row 2).

I finally view the simulation results in the context of observational evidence. Here I explicitly focus on the relative importance of individual authors and of journals. Stocks of knowledge largely tied together by individual ingenuity suggest weak institutional contexts. I index such informal processes through the share of references cited by contributions that belong to different specializations but have the same author, out of all references considered across specializations. I also analyze the role of journals as a key feature of the institutional infrastructure underlying academic disciplines. Here I code references authors made with respect to the journal that published the cited work and then estimate the modularity of this structure (see table 10.2, row 3).

These tests are designed to capture different principles by which scholars identify with their stocks of knowledge and the communities that have created them.

10.5 Analysis and results

I now present results following from the analytical steps just described. I first consider fragmentation of stocks of knowledge across the three cases throughout their recent history. Step two deployed the simulations to analyze the relationship between the observed fragmentation, Matthew Effects, and deviating practices, and step three specifies alternative processes. In step four I consider observational evidence for further clarification.

10.5.1 Fragmentation of knowledge and historical struggle

Historical struggles shape stocks of knowledge (Eyal et al. 2010, Abbott 1988). From the perspective of groups that claim and defend their jurisdictions against intruders, such struggle entails agreement on what knowledge is relevant for the group. From the perspective of informal movements, the struggle entails developing expertise that differs from the established stock of knowledge. The top graph of figure 2 shows annual size–scaled modularity scores of networks to which nodes, that is publications, have been cumulatively added according to the date of their publication. The bottom panel shows the change of a focal year relative to the previous year. To control for new members entering the field, the

vertical dashed lines indicate the year in which the average instructor published her first article. The three cases differ with respect to their internal variation and relative to one another.

Systems biology has consolidated following a phase of systematic fragmentation. This interplay of general agreement and compartmentalization as well as its position relative to law are consistent with expectations from the literature on scientific knowledge production and with its historical trajectory. Because these processes are well known, I now focus on law and data science (see appendix A2a for a discussion of the systems biology results).

Law remains stable and active. The network shows annual fluctuation in its modularity score (ticks in the lower graph), but no trend toward either more or less agreement (stable line in upper graph). We would not expect to see a trend within one generation of a field where changes unfold over centuries. The field remains active as legal scholars continue to integrate their contributions into the existing body of knowledge and reveal ambiguity in legal concepts. We have considered the debate over Internet law above as one such issue. The consistently low degree of fragmentation over relevant literature corresponds to the historical legacy of legal scholarship that by now goes back over a century. It does not indicate how the overall shape relates to the way scholars make their contributions (recall table 10.2, rows 1 and 2). I turn to this aspect after considering the data science case.

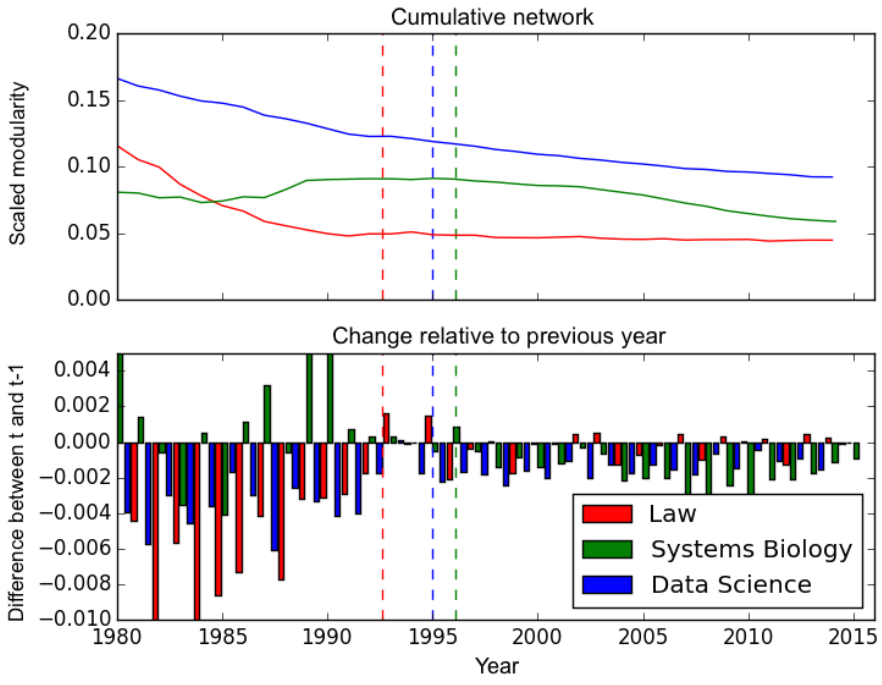

Figure 10.2. Size-scaled modularity scores for cumulative co-reference networks in legal scholarship, systems biology and data science. The vertical dashed lines indicate the year in which the average instructor published her first article.

Data science scholars increasingly engage with overlapping literatures. This can be seen as the modularity continuously decreases. Although the rate of change slows down, it does not stabilize before the end of the observed period (bottom panel of figure 10.2). Given that data science was only conceived of as a formal program by the end of this period, the observed trend shows that scholars from across the constituting disciplines and departments had begun to form agreement before they were assigned to teach in data science programs. At the same time, since both systems biology and data science originate in academia, it is noteworthy that data science forms around deeper gaps, and considers them worth closing, than we see in systems biology. Because we also know that data science lacks an institutional infrastructure with integrative forces, it remains unclear at this point which processes generate the observed structure instead. I consider the different candidates next.

In summary, these temporal dynamics follow expected patterns given the cases' historical backgrounds, but they fail to explain why the stock of knowledge some groups engage with and produce appear obvious, in terms of the lay applications they facilitate, and others arcane. Their differences begin to address the tension between stocks of knowledge as both institutional correlates and informal activities

of scholars on the basis of the principles they follow. Law, the canonical case especially for the institutionalist side, resembles what would be expected of knowledge generated by institutionalized consolidation; although scholars remain active, likely in part through the informal activities some research has turned to (e.g., Sandefur 2015), the overall order changes little. Conversely, systems biology, with its organization in laboratories as well-documented contexts for the development of informal expertise (Collins 1998), has undergone some substantial disruption and remains more contentious than law.

Inferring that agreed-upon knowledge gains lay visibility, while contentious knowledge does not, is troubled by the data scientists who have gained public salience despite their disagreement surpassing that of systems biology.

10.5.2 Constituting principles of stocks of knowledge

How do global contours emerge from scholarly practices? The structure of data science, relative to that of law and systems biology, showed that experts' capability to address generally salient problems is neither contingent on consolidating knowledge nor on retaining fragmentation in it. We therefore need to ask whether similar strategies of engaging with existing knowledge could be associated with different degrees of fragmentation (row 2 in table 10.2). I test the relationship between observed structures and institutional Matthew Effects, which would be the default baseline in the academic setting, and whether deviance is more likely to result from informal practices or alternative principles that integrate heterogeneous problems. I summarize results of these tests in figure 3. It reads from top to bottom as follows: It begins with the scaled estimated modularity scores for the observed structures (these correspond to the final value reported in figure 10.2), the subsequent rows analyze processes by which scholars integrate their contributions into the existing stock of knowledge. The dashed lines indicate where one data structure was used as starting condition for another simulation design. I review the main results before entering detailed discussions.

The main findings are as follows. Systems biology, beyond accommodating some informal activity, which is well documented in lab-based fields, shares an understanding of important ideas. Legal scholars and academics in data science do not. Although we have just seen that the legal field agrees on overlapping sets of literature, this overlap is not a result of publications preferentially citing already highly cited pieces. This does not mean that all sources are considered equally important, but that they are

combined in no particular order that has to do with solving smaller problems within larger ideas. Data science considers multiple different sets of important ideas and thereby diverges from both systems biology and law. Scholars leverage knowledge from diverse disciplines. Although the overall integration is marginal, scholars nevertheless seem to relatively systematically cite existing work that their immediate peers do not find relevant. Law and data science have therefore in common that they integrate heterogeneous problems. They differ in that legal scholars operate in the context of institutional channels while data science follows informal yet systematic scripts. I now discuss the findings for legal scholarship and data science in detail. Because this summary addresses the key results for systems biology, I leave a more detailed account of it for the appendix (see appendix section A2).

Legal scholarship

Legal scholarship integrates diversity as it routinely undermines scientific discipline. And of course, it is not a science even if it has found its way into universities. I turn to this in a moment. Meanwhile, the red square marker in figure 10.3, row 2, indicating the estimated modularity score of the Matthew Effect process, reveals considerably more fragmentation than we see originally. Recovering the observed order in the legal field from this formally simulated structure requires rewiring of over 20 percent of the underlying references (see size and position of red hexagon marker on row 3 in figure 3; see also figure A1b in the appendix). This is considerable compared to rewiring just 11 percent of the references to reconcile the two structures for systems biology (green hexagon marker; also figure A3.1a in appendix 3).[106] If legal scholars studied problems in the way system biologists and other scientists answer their questions, we would expect that the legal field considered each in more isolation than it does, or as Donoho put it, a continuously evolving approach. That the field deviates from the shared assessment of relevant ideas, one of the indicators of institutionalization, seems surprising given its canonical status.

[106] Appendix 2 discusses the systems biology results in detail.

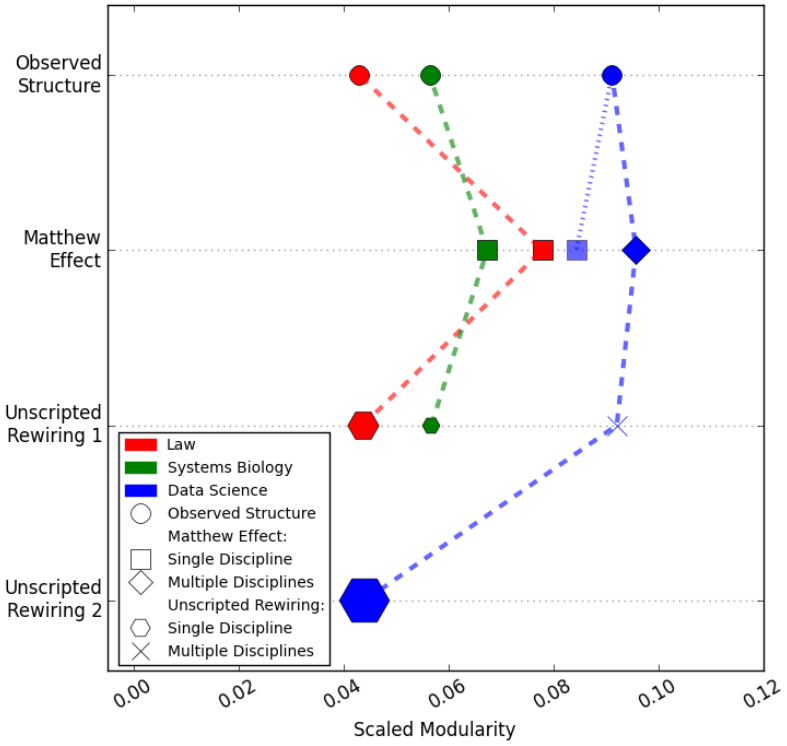

Figure 10.3. Estimated size-scaled modularity scores. The first row indicates values for observed networks, subsequent rows for simulations. In row 3, marker sizes for law and systems biology indicate the percentage of rewired ties, as for data science in row 4.

Law became part of universities late and is still an outsider today. While legal expertise was central for the development of the US into a formal entity, training in university affiliated programs considered standard today only formed almost a century later (Goody 1986, Stevens 1983). In this context, the result indicates how legal scholarship goes along with some disciplinary practices, such as journal publishing, but otherwise maintains its relation to the legal institution instead of adopting a scientific agenda. With this marginal position there is little reason to expect law to follow otherwise typical practices of organizing knowledge.

Legal scholarship moreover lacks a guiding theoretical orientation (Freidson 1986, Holmes 2009). It follows other principles. The considerable amount of unscripted rewiring, relative to systems biology, captures the earlier insight that legal scholars consider multiple problems without deeper connections. We have seen this qualitatively in the Internet debate, which juxtaposed widely different topics to reveal conceptual ambiguity and complementarity. Here we see these principles unfold systematically.

Law's consolidated stock of knowledge results from integrating conceptually unrelated problems through abstract analogies. System biology's stock of knowledge results from a combination of general agreement on important ideas and occasional deviance. Against this background, the quantitative results corroborate the tentative argument that entitlement comes less from being scientific, and rather from institutionalization. While today behaving scientifically would be one way to achieve this, as the systems biology result shows, it is not the only one, as we see here. Which principles integrate data science knowledge?

Data science

Despite its external salience, in which it qualitatively resembles law, data science roots in academic work, where it contrives knowledge from separate disciplines. Systems biology also emerged within academia, but split from biology, chemistry and related fields. It preserved the underlying strategy for engaging with existing work that is considered important by others. The systems biology simulation closely reproduces the observed structure, with just slightly more fragmentation. Contrary, the data science simulation initially shows considerable consolidation (see blue square marker on row 2 of figure 10.3). This suggests that if data science was also a singular, historically developing discipline and we thought that scholars shared a view of guiding ideas, we should have found less fragmentation in the observed data, not more. As data science appears like a "collection of technical activities," this result comes as no surprise. Scholars assigned to data science programs seem to systematically ignore some of the otherwise widely cited work, focusing instead on specific niches. While this would indeed be a familiar process of specialization within a single discipline and reflective of informal sources of knowledge production, here we know that these scholars have affiliations with several distinct disciplines, including statistics, mathematics, computer science and various substantive fields. This background leads to a different interpretation. Instead of specializing internally, data science scholars seem to find overlaps between guiding ideas in separate fields. That may still follow from locally informal practices, but here they unfold across many local settings. I now ask how they make these connections

Understanding data science requires accounting for its separate roots. I take one more step to specify the process by which data science scholars integrate the separate disciplines the field was superimposed on. I run the preferential attachment simulation, capturing Matthew Effect processes, on

distinct subsets of the overall citation network, again preserving the observed degree distributions. The third value reported in row 2 of figure 10.3 (diamond marker shape) indicates the estimated modularity score. Even the most conservative partition with a split in just two fields (these could be data science's computer science and statistics components) produces structures that reveal more disagreement on where to find new knowledge compared to the observed case. With this background I now analyze how scholars connect separate disciplines.

Data science scholars integrate disciplines without coordinating. To recognize this, we need to recall the side of the debate that finds how informal relations facilitate such deviance. Such processes, for example alliances across factions from the existing disciplines, imply groups sufficiently small to sustain interactions. The degree of rewiring offers an indicator of how concentrated the integration is. For strategic bridging, reconciling the observed and simulated structures takes rewiring of around 12% of the references in about 14% of the publications associated with data science expertise such that scholars find references colleagues from data science's other initiating fields consider theirs ("X"-marker row 3 of figure 10.3; see figure A3.1a in the appendix for a distribution of combinations).[107] Overall, the intensity of rewiring remains below that of systems biology and law. Yet the magnitude of cross-references is sufficiently large to expose purely local and informal efforts to integrate the field as unlikely. This evidence is consistent with systematic applications of the strategies seen qualitatively in the small-worlds instance above, or many of the richer accounts of part one, which analogically related biological networks to social structures through abstract formalism.

Data science and law therefore qualitatively resemble each other. Both integrate heterogeneous knowledge in unscripted patterns. At the same time, we know that law is much more institutionalized than data science, which the preceding temporal analysis reflected. Considering this relationship, we can ask how much data science scholars would be expected to reconstruct their stock of knowledge such that it resembles the integration of legal knowledge. This step narrows the range of trajectories data science

[107] The "X"-marker (row 3 of figure 10.3) reports the estimated modularity of a simulated structure after rewiring it strategically such that contributions seek to connect distinct fields. Simplified, two processes lead to strategic integration: Either few contributions integrate much diverse substance, or many publications integrating some. Disciplines make scholars focus on their main concerns (Foster, Rzhetsky, and Evans 2015, Abbott 2001), which can be expressed by relatively self-contained citation patterns. I therefore focus on results from configurations in which scholars aim to mainly contribute to their fields, instead of mainly bridging to others. The rewiring process is designed such that it reconciles the simulated with the observed structure while minimizing the number of publication from one specialization and references in them citing work from another. See figure A2 for configurations of rule deviating publications and their reference choices.

may plausibly follow. Here it takes rewiring, on average, of 44% of the references in three quarters of the publications (hexagon marker in figure 10.3, row 4; see also figure A3.2b in appendix 3). Such integration seems unlikely in an institutional context keeping data science's constituting fields apart through disciplinary politics and legacies. While data science addresses broadly salient problems, its institutional infrastructure remains underdeveloped. Scholars addressing data science problems will for some time at least continue to work without the orientation legal scholars find. One instance of this could be seen in the applied setting of New York City, where several academics joined the many nerds at the events. There we could see also that knowledge integration remains systematic even without more formal guidance.

Synthesis

Scholars engage with stocks of knowledge in ways that make their expertise appear arcane or obvious. Law and data science's stocks of knowledge take different contours than they would had the scholars constituting those fields followed strategies seen in scientific fields and related specific questions to larger aims. Instead they seek analogies and integrate heterogeneous problems. The qualitative analysis revealed formal abstractions as a process facilitating those connections. These scholarly strategies prevail in spite of varying institutional contexts of law and data science, one representing a stronghold and the other a novelty in society.

I expose these findings to one more set of observational evidence.

10.5.3 Practical contexts

Law and data science follow similar principles for integrating diverse knowledge. How do these principles unfold amid varying historical legacies? I now account for the empirical contexts of these processes. Table 3 reports two sets of observational results. The first column reports the estimated modularity scores when constructing the citation network through the journals from which publications cite pieces. The second column indicates the relative importance of authors integrating heterogeneous sets of ideas. It shows the percentage of shared citations between publications in different specializations (induced as part of the modularity estimation) by the same author. The results corroborate and contextualize the preceding simulations.

Law relies on institutionalized channels. The low modularity score for law indicates that legal scholarship widely agrees on which journals publish relevant contributions (see table 10.3). In data

Table 10.3. Role of authors and journals in integrating co-reference networks.

Field	Journal-reference modularity	Author bridging
Law	0.01	21%
Data Science	0.05	58%

science, on the contrary, journals only generate a slightly more integrated contour compared to the actual citations. That legal scholars draw on the same journals seems, retrospectively at least, obvious. The key formats in this field are review pieces, and all the major schools have their own law review. The collapse of any fragmentation must be seen against the backdrop of their local organization, however. We also have to bear in mind the organizational setup of these journals. Students serve as editors. They find guidance from their faculty, who, in turn, write the review pieces, strongly supports the argument that local expertise contributes to deviating from the general practice of citing already significant work. This process would lead to expect fragmentation at least on the level of the major schools. The low modularity score indicates, on the contrary, that the processes we observed qualitatively in the Internet law debate hold systematically. Scholars agree on abstract principles that apply across specialized areas.

Data science rests on individual, yet systematic efforts. The second column of table 10.3 indicates that data science relies substantially on individual scholars to integrate different specializations. Authors connect the heterogeneous stock of knowledge underlying data sciences by publishing across specializations, instead of simply citing material from other areas. This suggests that scholars actively relate heterogeneous problems to one another as part of their research agendas. Although drawn from distinct disciplines, data science scholars share a form of contributory expertise (Collins and Evans 2007). The instance of the small-world problem has shown how the quantitative orientation of data science facilitates formal abstractions of various contexts that thereby connect. Law operates through formal language, built around ideas such as contracts and zoning, that pertains to various problems. Quantitative analysis combined with computer science also works with formalisms, some of which apply indiscriminately across several substantive areas.

10.6 Discussion

10.6.1 Summary

This chapter set out to ask what problems nerds solve so that the stocks of knowledge they utilize seem obvious to the public. I found that law and data science are more broadly seen as practically relevant because they construct their stocks of knowledge from making contributions in the context of different problems. Systems biology foregoes public recognition because its scholars primarily contribute to a specificity problem their field is concerned with. Data science and legal scholars connect problems by transposing their analytical concepts, or schemas (Sewell 1992, 17), from one context to another to the degree to which they recognize analogies. For example, we have seen how the principles by which legal scholars reveal gaps in public and private law follow from juxtaposing cars and computer code. Similarly, we have seen how data science scholars demonstrate the utility of structural measures as they compare social networks to electricity grids. These principles operate in formal and in informal settings.

I was able to reveal these principles by viewing the three cases comparatively. The chief analytical leverage came from generating the contours stocks of knowledge take from the traces scholars leave behind. Reviews of qualitative instances gave insight into the principles by which scholars in each field engage with existing work in order to address new problems. I then designed a dataset of detailed citation records by a large number of scholars whom educational institutions assigned to teach in a respective field. In this data I analyzed the prevalence and effect of the constituting principles systematically and relative to the existing explanations, which tend to focus either on institutional or on informal processes. Both applied for data science as well, but we found the analogy-based strategies to be most relevant.

This chapter has added specificity to ideas from previous chapters. Similar to chapter nine, it formally compared contours of different stocks of knowledge. As this academic setting also scripts relevant activities more so than many other settings of daily work activity, while allowing for the variation for how that activities bears on the stock of knowledge, which is central here, we were able to distill with greater specificity the identification mechanisms of different thought communities. Consistent with previous findings, this analysis has recovered the improvisation strategies here among scholars associated with data science programs. Such improvisation draws on available material but integrates it

in ways that differ from the patterns usually underlying such knowledge production. I develop these points further in the conclusion.

10.6.2 Limitations and scope conditions

These conclusions are based on an analysis of practices specific to the institutional and historical context of higher education in the US today. There is still evidence to indicate that the ideas derived from this analysis apply more broadly and that opportunities will increase to test them in future research.

This chapter's network-based analytical strategy hinges on scholars' aim to leave traces and the academic publication conventions by which they do so. I argue that as more traces are recorded digitally and can be digitized more easily and accurately, as is evidenced by the growing amount of research based on such data (Evans 2013, Lazer et al. 2009), also studies of the knowledge encoded in them, and how it is constructed and applied in other social settings, will become feasible.

Moreover, the academic context underlying this analysis offers opportunities for engaging with stocks of knowledge few applied contexts do. Existing work on law and systems biology (Owen-Smith and Powell 2004, Stinchcombe 2001), and the previous chapters on data science, strongly suggest that lawyers and data scientists analogically integrate heterogeneous problems at least in effectively similar ways outside the academic context as we have found inside of it. The details of how they do so, in contexts with less permeable boundaries than disciplines maintain, remain subject to future work.

Finally, emerging data science and contemporary law make an asynchronous comparison, which raises the question of law's institutional and intellectual shape during its own beginning. For similar reasons as the ones just mentioned, it is not yet feasible to replicate the formal comparisons for this question. The historical evidence for law however shows that, similar to data science today, legal training and scholarship was contested in universities. I already noted that legal training as it is known today only formed in the second half of the nineteenth century. Earlier efforts were initiated by practitioners (Stevens 1983), as are many private offerings for data science seminars now. Moreover, we can recover the strategy of integrating heterogeneous problems, which emerged throughout this analysis, from central work that defined legal scholarship's inception (e.g., Holmes 2009).

10.6.3 Implications

Given the findings of this chapter it is unlikely that data science will take the shape of law in the near future, despite similarities and shared differences with sciences. Data science rather remains institutionally more fragmented with its integration still ongoing, obscuring paths toward identifying and attaining appropriate expertise for those seeking it. The findings also reveal, however, underlying principles of organizing knowledge that share with legal scholarship the quality of integrating heterogeneous problems. Both utilize analogies and thereby vary from theoretically sound sciences. The resulting stocks of knowledge, and their utility, become broadly salient. Whereas legal scholarship operates through bureaucratic and administrative channels, data science relies on technological infrastructures. This aspect opens important directions as modern technologies are likely to become more often the basis of professional work.

With a focus on the academic context, this chapter has considered more clearly considered a setting that more systematically relies on subtle and flexible mechanisms of control than those in previous settings. With journals and conferences, facilitated through the open discussion of knowledge, the academic context has mechanisms in place that control the overall body of knowledge without requiring elaborate formal infrastructures. Without them, to be sure, deviance in harmful ways is still easy. But as far as such consequences are accidental and not deliberate, the academic disciplinary framework offers a model at least for rudimentary control of data science work. I discuss the implications of this for data science specifically in the conclusion.

Summary of Part II

This part has addressed with more specificity principles of data nerd work that part one begun to reveal. Previous chapters had left our understanding of how the contours those principles produce relate to data science's salience tentative. More precise results have followed from comparative designs in economic and academic settings, the two most relevant institutional contexts from part one. I limit myself to brief summaries of the results, and consider the main contributions and implications in the overall conclusion that immediately follows.

Chapter nine has studied emerging contours of data science knowledge and skill identification in comparison to that of law and occupations that work without significant autonomy from bureaucratic control. It formally analyzed contours of expert skills on the basis of a large corpus of textual job descriptions as a way to consider the role of abstract knowledge, relative to organizational and institutional processes, as source of public recognition. We were able to recover how data scientists are seen to have combined expertise form distinct specializations, even absent the rich references to classical scholars and their biographies of chapter eight and from accounts of how others see their skills. On the basis of these contours they resemble the canonical legal profession in encountering expectations that require them to transpose knowledge from one context to another. Moreover, we saw that unlike occupations focused on singular industries, data science skills apply across organizational forms. We could also see evidence of distinct integration, compared to a simpler addition of skills from existing areas of expertise, most importantly quantitative and computational skills. If data science definitions in the job market setting had simply added skills from one of the existing areas to the other, this would have indicated the bearing of the respective older field's institutional framework as a source scripting data science expertise as well. That we instead discovered the restructuring of expertise associated with data science challenges to specify the principles through which data science integrates distinct fields. We can recall here conclusions of chapter six, which was also considered on work and skills, where we also found integrative practice. Whereas there we considered local experiences, which revealed improvising, here we have seen those contours on a systematic level across a large number of specific problems where data nerd skills are sought.

As in part one, here I have turned again to discipline as a form of coordinating such practices. As with the applied setting here, also in a simultaneously more granular and systematic, though less austere perspective.

Chapter ten analyzed data science relative to familiar academic disciplines in order to scrutinize the specific identification principles by which contours of abstract knowledge becomes publicly salient. Here I compared citation networks of scholars in data science, law and systems biology, and found variation reflecting their historical trajectories. Further simulations have shown however that both law and data science depart from principles of theoretically guided knowledge construction by integrating otherwise distinct specializations through analogies. Law relies on institutionalized channels while data science scholars systematically undermine disciplinary boundaries. We can interpret these results to capture how the principles by which data scientists identify existing expertise are associated with stocks of knowledge that accommodate heterogeneous specializations. Data scientists utilize computer scripts to format data of empirical problems for mathematical analyses, and they manipulate the math to suit specific problems. They draw these components from the institutionalized contexts of academic disciplines and integrate them informally as they analyze empirically and theoretically unconnected problems. This improvisation practice induces relations from analogies. The history of the tools data scientists use invokes the premise of research on professions as a way to conceive of continuity of groups tied by abstract knowledge. Data science is also just beginning to emerge and draws on unscripted processes to construct its stock of knowledge. These are familiar markers that are consistent with the informal improvisation data nerds described in part one.

III. Conclusions

11 Data science: Contours, chances and consequences

We have come a long way from a single-sentence Wikipedia definition of data science to considering data nerds and data science expertise in the rich and vivid setting of New York City's technology scene, the mundane professional job market and elite sciences. Where has this left us with respect to our understanding of this emerging profession and its consequence? I remain brief in addressing this question. I take five specific perspectives: (1) What is data science; (2) why is it distinctively salient; (3) who has control; (4) how can one become a data scientist; and (5) what is its sociological value in terms of improving our understanding of expert work in modern technology, and more broadly?

(1) What is data science?

We began with a Wikipedia definition that besides the obvious reference to different data formats related the novel field of data science to existing fields from statistics and computer sciences. It also raised some puzzles as some of the components it supposedly drew on reflected commercial motivations, whereas others had scientific backgrounds, and as some components of larger and established stocks of knowledge were chosen, and not others. Moreover, Wikipedia, which almost anyone can contribute to, reminded us of the possible tensions of the different camps that define data science expertise while drawing on the definitions of other areas of expertise, and arranging them in way that potentially differed from what those experts who defined them originally had in mind. Finally, it was not clear how the kind of general definition we found there would unfold across problems as different as shopping chains and online social networks, where data science has received public scrutiny. What did the subsequent empirical investigations reveal that the initial definition left unclear?

We have learned how data science works in a fair amount of detail, first in accounts of data nerds reporting on their experiences applying it, and then in practice of both using data science and directly observing it in natural settings without the curation of public events, or rather a different kind of curation that more clearly revealed variation with which we could address the new questions that had emerged. Interpreting results from data science practice was much more tedious than data nerds' reports thereof. But even in the reports of experiences data nerds shared with their audiences, we could get the clear sense that connecting the different fields an abstract definition of data science combines in a sentence

and through hyperlinks, involves much more uncertainty and struggle in practice. Moreover, the distinction between practical applications and scientific backgrounds turned out to be less clear than abstract summaries indicate. The transition invokes a sense of “recovery” and “brick walls,” although from other perspectives there is a sense of entitlement that one has, or not. To be sure, we have learned also about the differences of computer science and statistics “cultures,” the integration of data processing and analyzes through an intact stack, and the need to speak to others about appropriate questions and applications. That the integration is not so clear was also rejected in the formal studies of part two, where we could see redefinition of positions in across industries over just a few years in a way that came to resemble skill contours from the legal field. As we turned to the sciences, we found that the dignified halls of elite universities accommodate more diversity than their institutional status might suggest, and that that diversity, once again, leads to different kinds of contours.

Indeed, instead of understanding data science in terms of its relationship to arcane fields, we have come to understand data science in terms of the distinct kinds of contours of an autonomous thought community. The abstract idea from the introductory definition of combining different areas of knowledge involve for data nerds to illustrate the arcane technical underpinnings and more immediate persuasion that lifts their status precisely by not directly attacking others and instead attracting action through that passivity. For data nerds pursuing their own terms involves systematic improvising in order to solve problems lacking scripted solutions, as well as the intimacy of admitting struggle that is involved in defining such solutions. All these contours escape the familiar bureaucratic control most occupations experiences. Instead, data nerds integrate their heterogeneous practices on the basis of articulating them in more abstract and formalized references of the role models and scholars that have defined them initially. While this practice reminds of scientific citations, data nerds perform their own interpretation that through the informal settings leads to a more inclusive and diversified stock of knowledge. None of this could be seen from the singular definition.

Most concisely, we learned that data science is about turning concrete substance, or problems, into abstract representations, on the basis of often improvised arrangement of formal procedures.

(2) Why is data science distinctively salient to the lay public although it draws on arcane expertise?

We considered early on research that found for historical moments in quantification significant specialization in listing institutional settings, even where that quantitative expertise addressed practical and highly relevant problems in the census, insurance, civil engineering and individual credit worthiness. Indeed, even that the Wikipedia definition mostly referred existing areas of expertise gave no indication for how data science is seen to be distinctively salient. We can once again turn to the empirical material we have considered since that initial definition.

One immediate explanation, based on those observations, could be that data scientists focus explicitly on illustrating their arcane competencies in ways that lay audiences can follow. We can recall the references to the Obama campaign, oceanographers and Frankenstein, which represent web-based survey infrastructures, integration of data streams, and layering of storage and analytical capabilities. Data nerds spoke in lay terms instead of technical terms. But this conclusion would ignore all the practical challenges of applying these techniques to practical problems. Here we could see the improvisation, though we were not able to tell with certainty on what basis such activities generate salience while institutionally or otherwise formally guided specializations do not.

Answers to these questions emerged more clearly from part two and the comparative design introduced. Principles of deploying and constructing data science expertise, and the contours emerging from them, were more similar to those underlying legal expertise than those of both occupations with bureaucratically defined tasks, as well as scientifically specialized disciplines. In the introduction we considered Dewey's view to understand data science as a class or category salient to the public, instead of the independently viewed activities earlier studies of quantitative expertise have recovered. We could see that this takes a break with existing knowledge, defined by the sciences, and in addition to this collective effort of integrating that new combination of knowledge, recombining its formal components in informal arrangements. In other words, data science is distinctively salient precisely because it is not a science, despite its title.

(3) If data science undermines not only overt bureaucratic control but also the more subtle academic institutions, who has control?

The results so far are both concerning and comforting. Data nerds are no hackers, which although not harmful in the case of Aaron Swartz, with their unpredictable style of work sometimes are. They are

more disciplined, and besides the contact improvisation, also rely on more durable relationships by engaging with the open source movement. Nor have they clearly questioned the relevance of traditional hierarchies, even if they ridiculed those applications that “make people buy shit,” as long as the formal arrangements did not try to define data science tasks specifically. That is precisely the problem, however. Even if data nerds have no negative intention, their work may have consequences others find alarming, and if they escape formal bureaucratic oversight, the systems we are used to rely on lose their efficacy.

Let us just recall the breadth of data science applications. With the advance of digital computers, “artificial intelligence,” that is codified expert knowledge, has taken over decision-making tasks, such as in health diagnostics (Collins 1992), transportation (Bilger 2013), warfare (Singer 2011), and has elsewhere replace human labor altogether (Collins 1992). Recently, data scientists have taken a more subtle turn in developing ways for inferring personal and intimate information (Duhigg 2012) and manipulating emotions of those who use their products, in the interest of inducing behavioral change (Kramer, Guillory and Hancock 2014). Picking just one more example of many, consider quantitative models and their role in credit lending decisions where they threatened to reinforce discrimination on the basis of social constructs (Poon 2015). As noted in the beginning, we routinely recognize that lawyers and doctors impact our lives in numerous ways. Nerds impact our lives in multiple ways as well, but it was not so clear who to turn to, initially.

That we were able to analyze their principles relative to other academic fields offers significant direction. As data science undermines the traditional academic channels and transcends disciplinary boundaries, it also foregoes established sources of formal control, such as journals and associations. Against this background, the results nonetheless lead to practical implications for the problems we have seen data science confront us with. Having often sidestepped the question of whether data science is good or bad for us in the analyses, the findings show that the movement emerges from an intellectual underpinning that predates the formal organization in programs and centers, yet that the way nerds and scholars integrate that underpinning disconnects them from existing principles of control. It follows that because of the contours of this knowledge and underlying principles, addressing many concrete projects, but no larger aim, it seems ill-advised to treat data science as a corporate entity capable of administering procedures that make it conform with what we consider morally right and socially just.

While this is consistent with earlier conclusions, the academic setting of chapter ten has led to a more specific understanding of what appropriate and effective steps may look like. Rules apply to specific contexts but data scientists address general problems. Regulating which analytical models are legitimate, and which not, which information can be included, and which not, although sometimes necessary, will more often lead to a conundrum of exceptions and qualifications. It will be more productive, the results suggested, to iteratively engage with data science; it is most likely more responsive to discipline than to regulation. In law, courts mediate between private interests and public values (Lessig 1999). Universities have installed a similar function with IRBs. While data science's academic side is also subject to IRB review, the field addresses a range of problems beyond IRB jurisdiction. Prominent members of the movement have recognized this challenge, but the process of implementing similar infrastructures is still ongoing (Watts 2014). That President Obama's administration has hired data scientists and directly engages with their work supports this kind of development. After all, that we were able to recover data science's contours partly by considering emerging thought communities in other contexts, ranging from early forms of global trade, late socialism and antiquity's Catholicism have reminded also that discipline is not limited to academic setting we primarily associate it with today, and can therefore help us to take advantage of data science's applied contributes whiteout surrendering control.

(4) How can one become a data scientist?

There are countless blog posts, books, training initiatives and of by now also university courses describing and teaching how to become a data scientist. Ironically, the nerds who have defined data science in part I, and the scholars in part II, didn't follow any of them, because they were not yet available. By focusing on the organizational arrangements underlying data science work, we could see that this experience of defining their role might not have been insignificant for solving the problems they routinely encounter. What can we learn from this perspective in practical respects?

We began considering this question in terms of the work we are more familiar with in the technology context. We could clearly reject the proprietary type of work Bill Gates famously advocated. At the same time, there was also just limited support for his chief opposition, "hobbyists" like Linus Torvalds and his Linux community. Data nerds themselves like to introduce themselves as hackers, which,

because they often work in obscurity, we considered Aaron Swartz as a model of for the sporadic recognition he received for that work. We found that they differ from those hackers in important ways, as much as they endorse this view. In the work setting, data nerds replace the white collar that came to describe the new kind of office work over half a century ago with hoodies as their wear of choice. It is not only a change of fashion, but also a change of work, to be sure. They struggle deeply as part of solving the problems they confront. But they rarely choose the problems themselves. For better understanding this combination of individual autonomy and uncertainty as a collective routine, we have to go as far back as 1950s sociologist C. Wright Mills, who then observed the demise of another type of work that is autonomous, yet disciplined. Indeed, besides the professions of law and medicine, which I have more focused on here, Mills also emphasized the independent shopkeeper. There is not much data nerds and shopkeepers have in common, except the central feature of control over their tasks.

In other words, while becoming a data nerd does not require to challenge the existence of present institutional arrangement, it neither suffices to rely on them. The reference to hackers is useful in as far as data science work requires the kind of contact improvisation we associate with them. But that is not all it takes. Data science also entails to work in a continuously operating system of such improvisation practices, for which data nerds maintain more durable relationships. In a more technical summary we would seek to understand such a system through abstract terms instead of concrete historical figures. Data nerds have chosen “science” as such a term. They do not follow scientific principles, as they said in their own accounts and which we could see comparatively. The imagery of science as an institutional system for arcane work is useful nonetheless, precisely because the mechanisms of control disciplines offer promise to apply to data science as well. This requires from data nerds to take those mechanisms and make them accommodate the kind of principles they solve problems with and accept that they differ from those sciences typically follow.

(5) Where can we find implications for our sociological understanding of expert work in modern technology, and more broadly?

Integrating the literature on professions with the literature on expertise through the idea of thought communities, I have proposed a view that focuses on identification principles and “contours” of arcane knowledge and its applications. This view has several advantages. As a conceptual idea, it

accommodates and leverages substantively rich observations, as part one and the contours I outlined there demonstrated, as much as formal interpretations of the kind part two relied on. This view offers a modern interpretation of existing research, most significantly the idea of empirical categorization, and that of thought communities and constructivist views more broadly. Analyzing contours and the principles of their emergence thus leads to the advantage that it directly connects qualitative and informal to formal interpretations of social activities. It thereby facilitates an understanding of social processes that analysts can interpret, and that social actors experience. As a conceptual idea, it is not tied to a specific methodological framework or type of data. Analyzing contours of thought communities provides a context in which actors define and understand their activities, such as solving quantitative problems by improvising although that departs from earlier principles of solving quantitative problems. In this specific application, the "data science of data science" approach as pushed this point to the greatest extent as it showed that, like data science, this is an integrated, not additive view.[108]

The findings resulting from this approach shed new light on problems in the sociology of professions, knowledge and expertise. Viewing expert groups from the perspective of constituting principles of their knowledge appreciates the leverage of the concept of professions, their continuity, and the dynamism studies of expertise account for. Stocks of knowledge more easily transcend organizational boundaries than individual experts could, and penetrate locally emerging problems more quickly than formal groups. Research on professions aims to explain how experts construct and defend their authority over general lay problems, in contrast to occupations that are subject to hierarchical control. Expertise helps understand how specialists address sporadic and often local problems. A focus on constituting principles of expertise and the contours they form integrates the analytical scopes of professions and informal expertise, and hence their explanatory value.

An analytical perspective that focuses on constituting principles and contours knowledge is thus useful because it facilitates comparison and equips us to study a broad range of problems including economic investments, cultural differentiation and international relations, through the lens of the expert groups addressing them. Musicians, comedians and chefs are seen to address public concerns broadly and in varying ways. They are also marked by important differences, including their substantive

[108] That what might have looked like data science in part two was not data science is also clear now as it addressed no problem that was provided.

orientations and organizational contexts. Yet a focus on contours of their knowledge base offers purchase by revealing identification principles that are not necessarily bound by their respective settings. Diplomats, another instance, represent American interests globally. At the same time, they undergo periodic rotations as bureaucratic procedures for preventing standardization. Diplomats clearly rely on the US's global recognition, but need to reconcile this with local conditions. These processes complicate explanations that emphasize institutional status or training and raise questions of how diplomats transpose relevant knowledge from one context to another. In just one more context, the economy, we can again focus on knowledge in order to address questions such as how venture capitalists, seen to facilitate many promising economic developments, compare to the classical, hierarchically integrated investment banks, seen to destroy much value. Some of these groups are so different that comparing them seems unintuitive, at the same time this distance offers analytical leverage (Stinchcombe 1978), and thought community contours a lever to exploit this.[109]

Existing approaches underutilize relationships and differences between these cases. A professions perspective might focus on their respective institutional forms, while an expertise focus might lead to address their informal interactions internally and with outsiders. The focus on principles of constructing knowledge captures a different level of variation as it generates the contours from concrete activities, formal or informal, instead of relying on formal boundaries, classes or categories. Considering members of these occupations as actors skilled in transposing their knowledge amplifies these cases' analytical leverage. This perspective links easily overlooked areas of work to cases that are known to derive continuity from their expertise. By focusing on how these occupations transpose knowledge it captures the key resource facilitating work outside of conventional forms of employment at an analytical moment that also accounts for organizational constraints.

[109] Note that chefs, diplomats, musicians, venture capitalists and comedians all leave different types of traces in different contexts, such as recipes and menus, speeches, treaties and contracts, and scores, scripts and recordings. All of these index their respective stocks of knowledge. These types of data fit no traditional structures but lend themselves to the flexibility of modern computational strategies that extract comparable measures from them.

12 Reflections

Data science and I have matured over the course of this project. These experiences are interrelated. Although it is clear that data science shaped me much more than the other way around, distinguishing its own change from its effect on my view of it is not so easy. Today both data science and I have a clearer basis that allows revisiting how our relationship unfolded as a way to reveal my biases as well as features of emergence itself. As a younger student with a poorer of understanding of the social world, it is likely that I overlooked important processes. At the same time the aspects I have overlooked here may also not have existed when data science only began to gain recognition. Revisiting some of my design ideas and decisions as well as their background and reactions along the way helps to untangle the two processes.

The following paragraphs consider three moments of our shared history, their background and verdict. The first revolves around my attempts to define data science as a sociological problem, the second my ideas for operationalizing it, and the third some discoveries of how it works. The specific instances do not describe the decisions I eventually settled on, but they do reflect critical moments in the research process.

I first encountered data science in a classroom setting. And while the experience of seeing so much interest in such an arcane topic sparked my initial interest, I was looking for evidence that data science is real and consequential in order to accept it as a sociological research project. To me that required that it had tools at its disposal with which it could impact society more broadly. This intuition resonated with reports at the time of how data analysts invaded family privacy by revealing a teenage pregnancy. It also came from the thought that this is something sociology should have an answer to. As part of the process of developing ideas to research these consequences and their drivers systematically, I remember a meeting with Mitch Duneier where I described a project in which I wanted to ethnographically study both the methods and algorithms and their consequences among those they affect. Although Duneier encouraged my intuition and recommended existing work on these issues, I soon realized that such a scope stretched my ability. My early research designs reduced these ambitions to study the algorithm development directly. Because the focus on isolated groups of data scientists alone captured few activities that seemed relevant, as Shamus Khan helped me recognize, I saw little promise in

pursuing this direction further as well. It was through more conversations, workshops and catching up with some other literature on the formation of professions that I considered new directions.

I began to accept that what had I thought of as the less interesting question—that data science constitutes a group of experts—could nevertheless address some relevant problems. Settling on data science as an expert problem resonated with that time when data scientists articulated their claim on the title of the sexiest job of the twenty-first century in the media. This suggested to me job descriptions as a direct opportunity for considering how data scientists define their skills, as I had learned then to be a central question in the formation of professions. This approach also recovered some of my early questions in that experts rely on skills in order to build the tools and models with public and private consequences, which made me pay attention to the case initially. Moreover, in order to reveal distinct skills from the vast amount of job postings in data science that had started to pour out by that time, I resorted to doing data science myself. Conference presentations and informal discussions over this direction, however, reminded me that getting a better understanding of applied skills does not address how arcane knowledge unfolds itself.

Skills rely on knowledge but in order to understand their roots in established disciplines we would want to understand their knowledge directly. By that time data science had systematically spread across universities throughout the country. The class I had encountered data science in initially, for instance, was followed by another data science class in the next semester. Both were offered by external instructors and turned out as prelude to a large and well-funded institute with many faculty affiliates from several disciplines and that launched an entire data science program, conferences and so on. Similar developments occurred at universities across the country. Their pace and scope made it impossible to capture this transformation in its richness. Academics, at the same time, aim to leave traces through publications. My intuition was that if data science rests on distinct principles, they should emerge from the intellectual traces of those scholars that came to constitute the new institutes. Without going into too many details, I thought that these processes should emerge on a broad level and I demonstrated variation across the specializations indexed by journal homogeneity across data science and established disciplines. That was also the time of my proposal defense and the background of this preliminary finding did not convince my committee.

There had been plenty of citation research around that could have pointed me into the direction I initially though too nuanced for the questions data science raised. Although part one has none of the citation analyses, through them I started considering that varying skill patterns are associated with existing knowledge, and that these patterns pertain to their salience. After others helped me see that journals index forms of knowledge poorly (just thinking of AJS or ASR for us) my own data science exercises helped me understand how more detailed arrangements specifically in informal shortcuts matter for generating distinct stocks of knowledge. I could see this in my own data science exercises. For instance, as I analyzed the textual jobs data mentioned above, I was interested in applying a method rarely used by sociologists at that time. In order to ensure that my application seems reasonable I emailed its inventor, David Blei, directly and received an answer. There was no relationship (except that he was about to join Columbia, which perhaps contributed to his motivation to respond to my email), our entire conversation took about 200 words. It was enough for me to implement Blei's method.[110] With these experiences in mind I recovered a set of distinct principles that vary across salient and arcane instances of knowledge production. I still failed to establish a clear link to the data science scene I had been following throughout this time.

It was only after considering skills in organizations and knowledge in research that I recognized the community I had become part of. That this came so late although I had participated in the same events all along suggests that I missed it for a while. Indeed, my initial fascination with the magnitude and anonymity in which I first experienced data science, as well as the subsequent focus on its constituting scientific knowledge, likely diverted my attention from seeing the relevance of the processes unfolding right in front of me. It took a visit to Europe late in the project, where I met a friend from my undergraduate studies, for me to recognize some cohesion. We had learned statistics together and worked as TAs in methods courses but had not been in touch since that time. I knew that he had received a master's degree in statistics and then worked for a large bank. When we met again, however, he introduced himself as a data scientist. As we talked some more my approaches to the job description analyses resonated with him and his work sounded like the data science I knew. As the conversation went on we realized moreover that I had just been to events with speakers whose papers and blog posts my friend

[110] I did not include the result in my final analysis because it was too powerful for my relatively simple job descriptions.

follows from Europe as well. Talking about data science on another continent, not only about the kinds of practices but even the individuals who help defining them, led me to shift my focus. It also showed, however, that the sense of a community can be consistent with anonymous relationships.

On the other hand, perhaps I didn't miss the community aspect quite as much as part of this suggests. The core community of data scientists doubtlessly debated the principles of the field well before I was aware it existed. But they alone do not produce systematic salience. The first time crowds gathered to discuss these ideas, they did not constitute a distinct community. The events, which remained stable, saw the community emerge over time, both in numbers and in imagination. I have already mentioned instances that demonstrate this. We can recall the speaker who reflected on experiences as a data entrepreneur and who had met his cofounder at a prior event, or all those speakers who get cited for their talks elsewhere by audience members at those events who become speakers at others. Similarly, at an event I attended in the final days of this project I met a former Columbia student who had taken the data science class with me three years earlier. The last I knew of him was that he had started to work in the financial sector. As we caught up now at this event, I learned that in the meantime he had left that job and turned to data science, consulting startups and giving public presentations. In other words, he had moved on to make data science broadly salient.

The reflexive concern here can be expressed in terms of the structure of the previous chapters. How come discipline came in the end? Did I find it stylistically more appealing, is it part of the data science formation at a later point, or did I not see it before? I in fact almost missed it. I was on my way toward understanding the anonymous principles of data science work until my committee stopped me. I was not told that I missed the community, in fact that direction never made it into the initial project design in any developed forms. Instead, I did not know what else to do after finishing the other approaches, with at best mixed successes in convincing audiences. My initial motivation for spending time in the data science scene was to confirm that the field was still active.[111] The discipline imposed on me helped that I recognize community features. But what seems most important of all is that I had the time to see this grow and experience this process as a participant. Discipline takes time to unfold.

[111] Another strategy involved singing up with career websites to receive updates for data science openings. Seeing new posts every morning offered some relief against the most basic anxieties regarding the choice of a case that did not exist yet for a dissertation topic.

References

112th Congress. 2011. *H.R.3261 - Stop Online Piracy Act*. 2011-2012.

United States. Cong. House of Representatives. 2011b. Committee on the Judiciary. *H.R. 3261, Stop Online Piracy Act*.

Abbott, Andrew. 1981. "Status and Status Strain in the Professions." *American Journal of Sociology* 86 (4):819-835.

Abbott, Andrew. 1988. *The system of professions: An essay on the division of expert labor*: University of Chicago Press.

Abbott, Andrew. 1991. "The Order of Professionalization: An Empirical Analysis." *Work and Occupations* 18 (4):355-384. doi: 10.1177/0730888491018004001.

Abbott, Andrew. 1992. "From Causes to Events Notes on Narrative Positivism." *Sociological Methods & Research* 20 (4):428-455. doi: 10.1177/0049124192020004002.

Abbott, Andrew. 1993. "The sociology of work and occupations." *Annual review of sociology*:187-209.

Abbott, Andrew. 1983. "Professional Ethics." *American Journal of Sociology* 88 (5):855-885. doi: 10.1086/227762.

Abbott, Andrew. 2001. *Chaos of disciplines*. Chicago: University of Chicago Press.

Abbott, Andrew, and John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *Journal of Interdisciplinary History* 16 (3):471-494. doi: 10.2307/204500.

Allert, Tilman. 2009. *The Hitler salute : on the meaning of a gesture*. New York: Metropolitan Books/Henry Holt and Co.

Anderson, Benedict R. O'G. 1983. *Imagined communities : reflections on the origin and spread of nationalism*. London: Verso.

Aral, Sinan, and Marshall Van Alstyne. 2011. "The Diversity-Bandwidth Trade-off." *American Journal of Sociology* 117 (1):90-171. doi: 10.1086/661238.

Ayres, Ian. 2008. *Super crunchers: why thinking-by-numbers is the new way to be smart*. New York: Bantam Books.

Baker, Stephen. 2008. *The numerati*. Boston: Houghton Mifflin Co.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439):509-512.

Batagelj, Vladimir, Hans-Hermann Bock, Anuska Ferligoj, and Ales Ziberna. 2006. *Data science and classification*, *Studies in classification, data analysis, and knowledge organization.* Berlin: Springer-Verlag.

Bechky, B. A. 2003a. "Object lessons: Workplace artifacts as representations of occupational jurisdiction." *American Journal of Sociology* 109 (3):720-752. doi: 10.1086/379527.

Bechky, B. A. 2003b. "Sharing meaning across occupational communities: The transformation of understanding on a production floor." *Organization Science* 14 (3):312-330. doi: 10.1287/orsc.14.3.312.15162.

Becker, Howard S., Blanche Geer, Everett C. Hughes, and Anselm L. Strauss. 1976. *Boys in white: student culture in medical school*. New Brunswick, N.J.: Transaction Books.

Ben-David, Joseph. 1971. *The scientist's role in society: a comparative study*, *Foundations of modern sociology series*. Englewood Cliffs, N.J: Prentice-Hall.

Ben-David, Joseph, and Randall Collins. 1966. "Social Factors in the Origins of a New Science: The Case of Psychology." *American Sociological Review* 31 (4):451-465. doi: 10.2307/2090769.

Berger, Peter L., and Thomas Luckmann. 1966. *The social construction of reality: a treatise in the sociology of knowledge*. Garden City, N.Y.: Doubleday.

Bilger, Burkhard. 2013. "Auto correct: Has the self-driving car at last arrived?" *The New Yorker*, 2013/11/25/, 96-109.

Blau, Peter Michael, and Otis Dudley Duncan. 1967. *The American occupational structure*. New York: Wiley.

Bourdieu, Pierre. 2008. *The bachelors' ball: The crisis of peasant society in Béarn*. Cambridge: Polity.

Breiger, Ronald L. 1974. "The Duality of Persons and Groups." *Social Forces* 53 (2):181-190. doi: 10.2307/2576011.

Brekhus, Wayne. 2003. "Peacocks, chameleons, centaurs: Gay suburbia and the grammar of social identity." Based on the author's PhD thesis, Rutgers University, University of Chicago Press.
Brekhus, Wayne. 2007. "The Rutgers School: A Zerubavelian Culturalist Cognitive Sociology." *European Journal of Social Theory* 10 (3):448-464. doi: 10.1177/1368431007080705.
Brown, Peter. 1981. *The cult of the saints: its rise and function in Latin Christianity*. Chicago: University of Chicago Press.
Brown, Peter. 1982. *The Cult of the Saints: Its Rise and Function in Latin Christianity*: University of Chicago Press.
Brown, Peter. 1992. *Power and Persuasion in Late Antiquity: Towards a Christian Empire*: Univ of Wisconsin Press.
Brown, Peter. 2000. *Augustine of Hippo : a biography*. Berkeley: University of California Press.
Brummer, Chris. 2008. "Stock Exchanges and the New Markets for Securities Laws." *The University of Chicago Law Review* 75 (4):1435-1491.
Callon, Michel. 2008. "Economic Markets and the Rise of Interactive Agencements: From Prosthetic Agencies to "Habilitated" Agencies." In *Living in a material world: Economic sociology meets science and technology studies*, edited by Trevor Pinch and Richard Swedberg, 29-56. Cambridge, Mass.: MIT Press.
Carr-Saunders, A. M., and P. A. Wilson. 1933. *The professions*. Oxford: The Clarendon Press.
Carruthers, Bruce, and Wendy Nelson Espeland. 1991. "Accounting for rationality: Double-entry bookkeeping and the rhetoric of economic rationality." *American Journal of Sociology* 97 (1):31-69. doi: 10.1086/229739.
Cerulo, Karen A. 1998. *Deciphering violence : the cognitive structure of right and wrong*. New York: Routledge.
Ceruzzi, Paul E. 2003. *A history of modern computing*. London, Eng.; Cambridge, Mass.: MIT Press.
Chuang, Han-Yu, Matan Hofree, and Trey Ideker. 2010. "A Decade of Systems Biology." *Annual Review of Cell and Developmental Biology* 26 (1):721-744. doi: 10.1146/annurev-cellbio-100109-104122.
Cohain, Judy Slome. 2009. "Learning to be a midwife." *Midwifery today with international midwife* (90):44-6.
Cole, Jonathan R., and Stephen Cole. 1973. *Social stratification in science*: University of Chicago Press.
Coleman, E. Gabriella. 2013. *Coding freedom: The ethics and aesthetics of hacking*. Princeton: Princeton University Press.
Collins, Harry M. 1992. *Artificial experts: social knowledge and intelligent machines*. Cambridge, Mass.: MIT Press.
Collins, Harry M. 1998. "The Meaning of Data: Open and Closed Evidential Cultures in the Search for Gravitational Waves." *American Journal of Sociology* 104 (2):293-338. doi: 10.1086/ajs.1998.104.issue-2.
Collins, Harry M. 2004. *Gravity's Shadow: The Search for Gravitational Waves*. Chicago: University of Chicago Press.
Collins, Harry M., and Robert Evans. 2002. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32 (2):235-296.
Collins, Harry M., and Robert Evans. 2007. *Rethinking Expertise*: Chicago: University of Chicago Press.
Collins, Randall. 2000. *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Cambridge, Mass. [u.a.]: Belknap Press of Harvard University Press.
Conk, M. A. 1978. "Occupational Classification in the United States Census - 1870-1940." *Journal of Interdisciplinary History* 9 (1):111-130. doi: 10.2307/203672.
Conk, Margo Anderson. 1980. *The United States census and labor force change : a history of occupation statistics, 1870-1940*, *Studies in American history and culture*. Ann Arbor, Mich.: UMI Research Press.
Contributors, Wikipedia. 2016. Data science--Wikipedia, The Free Encyclopedia.
Crawford, Kate. 2016. "Artificial Intelligence's White Guy Problem." *The New York Times*. Accessed July 16, 2016. http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.
Davenport, Thomas H., and D. J. Patil. 2012. "Data Scientist: The Sexiest Job Of the 21st Century." *Harvard business review* 90 (10):70-76.
DellaPosta, D., Y. Shi, and M. Macy. 2015. "Why Do Liberals Drink Lattes?" *American Journal of Sociology* 120 (5):1473-1511. doi: 10.1086/681254.

Desrosiáeres, Alain. 1998. *The politics of large numbers: A history of statistical reasoning*. Cambridge, Mass.: Harvard University Press.
Dewey, John. 1954. *The public and its problems*. Chicago: Swallow Press.
Didier, Emmanuel. 2007. "Do Statistics "Perform" the Economy?" In *Do economists make markets? On the performativity of economics*, edited by Donald A. MacKenzie, Fabian Muniesa and Lucia Siu, 276-310. Princeton: Princeton University Press.
DiMaggio, Paul J. 1991. "Constructing an Organizational Field as a Professional Project: U.S. Art Museums, 1920-1940." In *The New Institutionalism in Organizational Analysis*, edited by Walter Powell and Paul DiMaggio, 267-292. Chicago: University of Chicago Press.
DiMaggio, Paul J., and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48 (2):147-160. doi: 10.2307/2095101.
Dorogovtsev, Sergey. 2010. *Lectures on Complex Networks*. Oxford: Oxford University Press.
Duhigg, Charles. 2012. "How Companies Learn Your Secrets." *The New York Times Magazine*, 2012/02/16/.
Elias, Norbert. 1985. *Über den Prozess der Zivilisation. 2, Wandlungen der Gesellschaft: Entwurf zu einer Theorie der Zivilisation*. "10" ed, *Suhrkamp-Taschenbuch Wissenschaft 159,10 159 (DE-576)016264207*: Suhrkamp.
Ensmenger, Nathan. 2010. *The computer boys take over computers, programmers, and the politics of technical expertise*. Cambridge, Massachusetts; London, England: The MIT Press.
Epstein, Steven. 1995. "The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials." *Science, Technology, & Human Values* 20 (4):408-437.
Epstein, Steven. 1996. *Impure science: AIDS, activism, and the politics of knowledge*. Berkeley: University of California Press.
Erikson, Emily, and Peter Bearman. 2006. "Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833." *American Journal of Sociology* 112 (1):195-230. doi: 10.1086/502694.
Espeland, Wendy Nelson, and Michael Sauder. 2007. "Rankings and reactivity: How public measures recreate social worlds." *American Journal of Sociology* 113 (1):1-40.
Etzioni, Amitai. 1969. *The semi-professions and their organization: Teachers, nurses, social workers*. New York: Free Press.
Evans, James A. 2010. "Industry Induces Academic Science to Know Less about More." *American Journal of Sociology* 116 (2):389-452. doi: 10.1086/653834.
Evans, James A. 2013. "Future Science." *Science* 342 (6154):44-45. doi: 10.1126/science.1245218.
Evans, James A., and Jacob G. Foster. 2011. "Metaknowledge." *Science* 331 (6018):721-725. doi: 10.1126/science.1201765.
Evans-Pritchard, Edward. E. 1978. *Hexerei, Orakel und Magie bei den Zande*. Frankfurt am Main, Germany: Suhrkamp.
Eyal, Gil. 2013. "For a Sociology of Expertise: The Social Origins of the Autism Epidemic." *American Journal of Sociology* 118 (4):863-907. doi: 10.1086/668448.
Eyal, Gil, Brendan Hart, Emine Onculer, Neta Oren, and Natasha Rossi. 2010. *The autism matrix: The social origins of the autism epidemic*. Cambridge, UK ; Malden, MA: Polity.
Fallon, Richard H., and Daniel J. Meltzer. 2007. "Habeas Corpus Jurisdiction, Substantive Rights, and the War on Terror." *Harvard Law Review* 120 (8):2029-2112.
Feigenbaum, Edward A., and Pamela McCorduck. 1983. *The fifth generation: Artificial intelligence and Japan's computer challenge to the world*. Reading, Mass.: Addison-Wesley.
Fernandez, James W. 1972. "Persuasions and Performances: Of the Beast in Every Body... And the Metaphors of Everyman." *Daedalus* 101 (1):39-60.
Ferraro, Fabrizio, and Siobhan O'Mahony. 2012. "Managing the Boundaries of an "Open" Project." In, edited by John F. Padgett and Walter W. Powell, 545-565. Princeton [N.J.]: Princeton University Press.
Foreman, John W. 2014. "Data smart: Using data science to transform information into insight." In. Indianapolis, Ind.: John Wiley & Sons.
Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. "Tradition and Innovation in Scientists' Research Strategies." *American Sociological Review* 80 (5):875-908. doi: 10.1177/0003122415601618.

Foucault, Michel. 1995. *Discipline & punish: The birth of the prison*: Vintage.
Fourcade, Marion. 2006. "The Construction of a Global Profession: The Transnationalization of Economics." *American Journal of Sociology* 112 (1):145-194. doi: 10.1086/ajs.2006.112.issue-1.
Fourcade, Marion. 2009. *Economists and Societies: Discipline and Profession in the United States, Britain, and France, 1890s to 1990s*: Princeton University Press.
Freidson, Eliot. 1960. "Client Control and Medical Practice." *American Journal of Sociology* 65 (4):374-382.
Freidson, Eliot. 1961. *Patients' views of medical practice: a study of subscribers to a prepaid medical plan in the Bronx*. New York: Russell Sage Foundation.
Freidson, Eliot. 1973. *The professions and their prospects*. Beverly Hills [Calif.]: Sage Publications.
Freidson, Eliot. 1986. *Professional powers: a study of the institutionalization of formal knowledge*. Chicago: University of Chicago Press.
Freidson, Eliot. 1988. *Profession of medicine: a study of the sociology of applied knowledge*. Chicago: University of Chicago Press.
Freidson, Eliot. 2001. *Professionalism: The Third Logic*. Chicago: University of Chicago Press.
Freudenthal, Gad. 1991. "General Introduction: Joseph Ben-David, An Outline of His Life and Work." In *Scientific growth: essays on the social organization and ethos of science*, edited by Joseph Ben-David and Gad Freudenthal, 1, 25. Berkeley: University of California Press.
Friedman, Lawrence M. 2005. *A history of American law*. 3rd ed ed. New York: Simon & Schuster.
Gieryn, Thomas F. 1983. "Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists." *American Sociological Review*:781-795.
Giles, J. 2005. "Internet encyclopaedias go head to head." *Nature* 438 (7070):900-901. doi: 10.1038/438900a.
Gill, Matthew. 2009. *Accountants' Truth: Knowledge and Ethics in the Financial World*. Oxford, U.K.: Oxford University Press.
Gilmore, Grant. 1995. *The death of contract*. Columbus: Ohio State University Press.
Ginzburg, Carlo. 1992. *The cheese and the worms : the cosmos of a sixteenth-century miller*. Baltimore: Johns Hopkins University Press.
Goffman, Erving. 1990. *Asylums: Essays on the social situation of mental patients and other inmates*. New York: Doubleday.
Goldberg, Amir, Sameer B. Srivastava, V. Govind Manian, and Christopher Potts. 2016.
Goode, William J. 1957. "Community Within a Community: The Professions." *American Sociological Review* 22 (2):194-200. doi: 10.2307/2088857.
Goode, William J. 1960. "Encroachment, Charlatanism, and the Emerging Profession: Psychology, Sociology, and Medicine." *American Sociological Review* 25 (6):902-914.
Goode, William J. 1961. "The Librarian: From Occupation to Profession?" *The Library Quarterly* 31 (4):306-320.
Goody, Jack. 1986. *The logic of writing and the organization of society*, *Studies in literacy, family, culture, and the state*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.
Gorman, Elizabeth. H. 2005. "Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms." *American Sociological Review* 70 (4):702-728.
Gorman, Elizabeth. H., and Rebecca L. Sandefur. 2011. ""Golden Age," Quiescence, and Revival: How the Sociology of Professions Became the Study of Knowledge-Based Work." *Work and Occupations* 38 (3):275-302. doi: 10.1177/0730888411417565.
Granovetter, Mark S. 1974. *Getting a job: A study of contacts and careers*. Cambridge, Mass.: Harvard University Press.
Green, P. 1997. "Against a whole-genome shotgun." *Genome Research* 7 (5):410-417.
Gregan-Paxton, J., and D. R. John. 1997. "Consumer learning by analogy: A model of internal knowledge transfer." *Journal of Consumer Research* 24 (3):266-284. doi: 10.1086/209509.
Hardy, Quentin. 2016. "Tech Companies, New and Old, Clamor to Entice Cloud Computing Experts." *The New York Times*, March 7. http://www.nytimes.com/2016/03/07/technology/tech-companies-new-and-old-clamor-to-entice-cloud-computing-experts.html.
Healy, Kieran. 2015. "The Performativity of Networks." *ArcHives Europeennes De Sociologie* 56 (2):175-205. doi: 10.1017/s0003975615000107.

Heinz, John P., and Edward O. Laumann. 1982. *Chicago lawyers: The social structure of the bar*. New York; Chicago: Russell Sage Foundation; American Bar Foundation.
Hoeffler, S. 2003. "Measuring preferences for really new products." *Journal of Marketing Research* 40 (4):406-420. doi: 10.1509/jmkr.40.4.406.19394.
Holmes, Oliver Wendell. 2009. *The common law*. Cambridge, Mass: The Belknap Press of Harvard University Press.
Hood, Leroy. 2008. "A Personal Journey of Discovery: Developing Technology and Changing Biology." *Annual Review of Analytical Chemistry* 1 (1):1-43. doi: 10.1146/annurev.anchem.1.031207.113113.
Howarth, David. 2013. *Law as engineering: thinking about what lawyers do*. Cheltenham, UK; Northampton, MA, USA: Edward Elgar.
Hughes, E. C. 1963. "Professions." *Daedalus* 92 (4):655-668.
Ideker, Trey, Timothy Galitski, and Leroy Hood. 2001. "A new approach to decoding life: Systems biology." *Annual Review of Genomics and Human Genetics* 2 (1):343-372. doi: 10.1146/annurev.genom.2.1.343.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*." SpringerLink.
Janssens, Jeroen. 2014. *Data science at the command line*. United States of America: O'Reilly Media.
Kaufman, Herbert. 1986. *The forest ranger: A study in administrative behavior*. Washington, D.C.: Resources for the Future.
Keller, Evelyn Fox. 2002. *Making sense of life explaining biological development with models, metaphors, and machines*. Cambridge, Mass.: Harvard University Press.
Kelty, Christopher M. 2008. *Two bits: The cultural significance of free software*, *Experimental futures*. Durham: Duke University Press.
Khan, Shamus Rahman. 2011. *Privilege : the making of an adolescent elite at St. Paul's School*, *Princeton studies in cultural sociology*. Princeton, N.J.: Princeton University Press.
Klegon, D. 1978. "The Sociology of Professions: An Emerging Perspective." *Sociology of Work and Occupations* 5 (3):259-283.
Knappenberger, Brian. 2014. The Internet's Own Boy: The Story of Aaron Swartz.
Knoke, David, and Song Yang. 2008. *Social Network Analysis*. Los Angeles, London, New Delhi, Singapore: Sage Publications.
Kraft, Philip. 1977. *Programmers and managers: the routinization of computer programming in the United States*. New York: Springer-Verlag.
Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental evidence of massive-scale emotional contagion through social networks." *Proceedings of the National Academy of Sciences* 111 (24):8788-8790. doi: 10.1073/pnas.1320040111.
Kreiner, Kristian. 2012. "Organizational decision mechanisms in an architectural competition." In *The Garbage Can Model of Organizational Choice: Looking forward at Forty*, 399-429. Emergald.
Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Lander, Eric S. 1996. "The New Genomics: Global Views of Biology." *Science* 274 (5287):536-539.
Lander, Eric S. et al. 2001. "Initial sequencing and analysis of the human genome." *Nature* 409 (6822):860-921. doi: 10.1038/35057062.
Larson, Magali Sarfatti. 1977. *The rise of professionalism: A sociological analysis*. Berkeley: University of California Press.
Larson, Magali Sarfatti. 1980. "Proletarianization and Educated Labor." *Theory and Society* 9 (1):131-175.
Latour, Bruno, and Steve Woolgar. 1986. *Laboratory life: The construction of scientific facts*. Princeton, N.J.: Princeton University Press.
Lave, Jean. 1988. *Cognition in practice: Mind, mathematics, and culture in everyday life*. Cambridge; New York: Cambridge University Press.
Lazega, Emmanuel. 2001. *The collegial phenomenon : the social mechanisms of cooperation among peers in a corporate law partnership*. Oxford: Oxford University Press.
Lazer, D., A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. "SOCIAL SCIENCE: Computational Social Science." *Science* 323 (5915):721-723. doi: 10.1126/science.1167742.

Leifer, Eric M. 1988. "Interaction Preludes to Role Setting: Exploratory Local Action." *American Sociological Review* 53 (6):865-878.
Leifer, Eric, and Valli Rajah. 2000. "Getting Observations: Strategic Ambiguities in Social Interaction." *Soziale Systeme* 6:251-267.
Liu, Ka-Yuet, Marissa King, and Peter S. Bearman. 2010. "Social Influence and the Autism Epidemic." *American Journal of Sociology* 115 (5):1387-1434.
Llewellyn, Karl N. 1960. *The common law tradition: Deciding appeals*. Boston: Little, Brown.
Lynn, Freda B. 2014. "Diffusing through Disciplines: Insiders, Outsiders, and Socially Influenced Citation Behavior." *Social Forces* 93 (1):355-382. doi: 10.1093/sf/sou069.
Lévi-Strauss, Claude. 1963. *Structural anthropology*. New York: Basic Books.
MacKenzie, Donald. 1978. "Statistical Theory and Social Interests: A Case-Study." *Social Studies of Science* 8 (1):35-83.
MacKenzie, Donald. 2011. "The Credit Crisis as a Problem in the Sociology of Knowledge." *American Journal of Sociology* 116 (6):1778-1841.
MacKenzie, Donald. 2014. “A Sociology of Algorithms: High-Frequency Trading and the Shaping of Markets.“
MacKenzie, Donald. 2016. “How Algorithms Interact: Goffman’s ‘Interaction Order’ in Automated Trading.“
MacKenzie, Donald A. 1981. *Statistics in Britain, 1865-1930: the social construction of scientific knowledge*. Edinburgh: Edinburgh University Press.
Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.
Marshall, Emily A. 2013. "Defining population problems: Using topic models for cross-national comparison of disciplinary development." *Poetics* 41 (6):701-724. doi: 10.1016/j.poetic.2013.08.001.
Martin, John Levi. 2000. "What do animals do all day? The division of labor, class bodies, and totemic thinking in the popular imagination." *Poetics* 27 (2-3):195-231. doi: 10.1016/s0304-422x(99)00025-x.
Martin, John Levi. 2002. "Power, Authority, and the Constraint of Belief Systems." *American Journal of Sociology* 107 (4):861-904. doi: 10.1086/ajs.2002.107.issue-4.
McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology* 27:415-444. doi: 10.1146/annurev.soc.27.1.415.
Meadow, Tey. 2011. "'Deep down where the music plays': How parents account for childhood gender variance." *Sexualities* 14 (6):725-747. doi: 10.1177/1363460711420463.
Menchik, Daniel A. 2014. "Decisions about Knowledge in Medical Practice: The Effect of Temporal Features of a Task." *American Journal of Sociology* 120 (3):701-749. doi: 10.1086/679105.
Merton, Robert K. 1968a. "The Matthew Effect in Science." *Science* 159 (3810):56-63.
Merton, Robert K. 1968b. *Social Theory and Social Structure*. 1968 enl. ed. ed. New York: Free Press.
Merton, Robert K. 1996. *On social structure and science*, *Heritage of sociology*. Chicago: University of Chicago Press.
Meyer, J. W., and B. Rowan. 1977. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83 (2):340-363. doi: 10.1086/226550.
Miller, Claire Cain. 2013. "Data Science: The Numbers of Our Lives." *The New York Times*.
Mills, C. Wright. 1951. *White Collar: The American Middle Classes*. New York: Oxford University Press.
Monk, E. P. 2015. "The Cost of Color: Skin Color, Discrimination, and Health among African-Americans." *American Journal of Sociology* 121 (2):396-444. doi: 10.1086/682162.
Moody, J. 2004. "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999." *American Sociological Review* 69 (2):213-238.
Moser, Petra. 2016. Patents and Innovation in Economic History.
Mudge, Stephanie Lee, and Antoine Vauchez. 2012. "Building Europe on a Weak Field: Law, Economics, and Scholarly Avatars in Transnational Politics." *American Journal of Sociology* 118 (2):449-492. doi: 10.1086/666382.
Muniesa, Fabian. 2014. *The Provoked Economy: Economic reality and the performative turn*. New York, NY: Routledge.
Muñoz, Cecilia, Megan Smith, and DJ Patil. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Executive Office of the President.

Navon, Daniel, and Uri Shwed. 2012. "The chromosome 22q11.2 deletion: From the unification of biomedical fields to a new kind of genetic condition." *Social science & medicine* 75 (9):1633-1641.
Neff, Gina, and David Stark. 2004. "Permanently Beta." In *Society Online: The Internet in Context*, edited by Philip N. Howard and Steve Jones. Thousand Oaks, CA [u.a.]: Sage.
Newman, M. E. J., and M. Girvan. 2004. "Finding and evaluating community structure in networks." *Phys. Rev. E* 69 (2):026113. doi: 10.1103/PhysRevE.69.026113.
Nippert-Eng, Christena E. 1996. *Home and work: Negotiating boundaries through everyday life*. Chicago, IL: University of Chicago Press.
Owen-Smith, Jason, and Walter W. Powell. 2004. "Careers and contradictions: Faculty responses to the transformation of knowledge and its uses in the life sciences." *Sociologie du Travail* 46 (3):347-377. doi: 10.1016/j.soctra.2004.07.001.
Parigi, Paolo. 2012. *The rationalization of miracles*. New York: Cambridge University Press.
Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. New York: Penguin Press.
Parsons, Talcott. 1939. "The Professions and Social Structure." *Social Forces* 17 (4):457-467. doi: 10.2307/2570695.
Phillips, Damon J., Catherine J. Turco, and Ezra W. Zuckerman. 2013. "Betrayal as Market Barrier: Identity-Based Limits to Diversification among High-Status Corporate Law Firms." *American Journal of Sociology* 118 (4):1023-1054. doi: 10.1086/668412.
Piore, M. J. 2011. "Beyond Markets: Sociology, street-level bureaucracy, and the management of the public sector." *Regulation & Governance* 5 (1):145-164. doi: 10.1111/j.1748-5991.2010.01098.x.
Poon, Martha. 2015. "Statistically Discriminating Without Discrimination: The history of data analytics in consumer credit." history of data / data in history conference, 2015/04/18/.
Porter, Theodore M. 1986. *The rise of statistical thinking, 1820-1900*. Princeton, N.J: Princeton University Press.
Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J: Princeton University Press.
Powell, W. W., and K. Snellman. 2004. "The knowledge economy." *Annual Review of Sociology* 30:199-220. doi: 10.1146/annurev.soc.29.010202.100037.
Rubin, Jeffrey. 2006. "When Parent and Subsidiary Are Public; News; Overlapping responsibilities require coordination." *New York Law Journal*.
Rueschemeyer, Dietrich. 1973. *Lawyers and their society; a comparative study of the legal profession in Germany and in the United States*. Cambridge: Harvard University Press.
Sampson, Robert J. 2011. *Great American city : Chicago and the enduring neighborhood effect*. Chicago: The University of Chicago Press.
Sandefur, Rebecca. L. 2015. "Elements of Professional Expertise: Understanding Relational and Substantive Expertise through Lawyers' Impact." *American Sociological Review* 80 (5):909-933. doi: 10.1177/0003122415601157.
Sauder, Michael, and Wendy Nelson Espeland. 2009. "The Discipline of Rankings: Tight Coupling and Organizational Change." *American Sociological Review* 74 (1):63-82.
Saxenian, Anna Lee. 1996. *Regional advantage: Culture and Competition in Silicon Valley and Route 128*. 1st Harvard University Press pbk. ed. ed. Cambridge, Mass. :: Harvard University Press.
Schneider, Todd W. 2016. "Taxi Use Patterns Can Tell Us How Good the Super Bowl and Halftime Show Were." *The New York Times*, Feb. 3, The Upshot. http://www.nytimes.com/2016/02/04/upshot/taxi-use-patterns-can-tell-us-how-good-the-super-bowl-was.html.
Schutt, Rachel, and Cathy O'Neil. 2013. *Doing data science*. United States of America: O'Reilly Media, Inc.
Seabrooke, Leonard. 2014. "Epistemic arbitrage: Transnational professional knowledge in action." *Journal of Professions and Organization* 1 (1):49-64.
Seabrooke, Leonard, and Eleni Tsingou. 2014. "Distinctions, affiliations, and professional knowledge in financial reform expert groups." *Journal of European Public Policy* 21 (3):389-407. doi: 10.1080/13501763.2014.882967.
Sewell, William H., Jr. 1992. "A Theory of Structure: Duality, Agency, and Transformation." *American Journal of Sociology* 98 (1):1-29.

Shan, Carl, Henry Wang, William Chen, and Max Song. 2015. *The data science handbook: Advice and insights from 25 amazing data scientists*.
Shwed, Uri, and Peter S. Bearman. 2010. "The Temporal Structure of Scientific Consensus Formation." *American Sociological Review* 75 (6):817-840.
Singer, P. W. 2011. "Military robotics and ethics: A world of killer apps." *Nature* 477 (7365):399-401. doi: 10.1038/477399a.
Smith, Megan. 2015. "Title." *The White House Blog*. https://www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist.
Stark, David. 2009. *The sense of dissonance: Accounts of worth in economic life*. Princeton: Princeton University Press.
Stevens, Robert Bocking. 1983. *Law school: legal education in America from the 1850s to the 1980s*, *Studies in legal history*. Chapel Hill: University of North Carolina Press.
Stinchcombe, A. L. 1982. "Should Sociologists Forget Their Mothers and Fathers?" *American Sociologist* 17 (1):2-11.
Stinchcombe, Arthur L. 1978. *Theoretical methods in social history*. New York: Academic Press.
Stinchcombe, Arthur L. 1997. "On the virtues of the old institutionalism." *Annual Review of Sociology* 23:1-18.
Stinchcombe, Arthur L. 2001. *When formality works: authority and abstraction in law and organizations*. Chicago: University of Chicago Press.
Stinchcombe, Arthur L. 2005. *The logic of social research*. Chicago: University of Chicago Press.
Svensson, Lennart G. 1990. "Knowledge as a professional resource: case studies of architects and psychologists at work." In *The formation of professions: Knowledge, state and strategy*, edited by Rolf Torstendahl and Michael Burrage, 51-70. London: SAGE Publications.
Swidler, A. 1986. "Culture in Action: Symbols and Strategies." *American Sociological Review* 51 (2):273-286. doi: 10.2307/2095521.
Thelen, Kathleen. 2003. "How Institutions Evolve: Insights from Comparative Historical Analysis." In, edited by James Mahoney and Dietrich Rueschemeyer, 208. Cambridge, UK: Cambridge University Press.
Turco, Catherine J. 2016. *The Conversational Firm: Rethinking Bureaucracy in the Age of Social Media*: Columbia University Press.
Uzzi, B., and R. Lancaster. 2004. "Embeddedness and price formation in the corporate law market." *American Sociological Review* 69 (3):319-344.
Vaughan, Diane. 1996. *The Challenger launch decision: risky technology, culture, and deviance at NASA*. Chicago: University of Chicago Press.
Venter, J. Craig, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507):1304-1351.
Vollmer, Hendrik, Andrea Mennicken, and Alex Preda. 2009. "Tracking the numbers: Across accounting and finance, organizations and markets." *Accounting Organizations and Society* 34 (5):619-637. doi: 10.1016/j.aos.2008.06.007.
von Roten, Fabienne Crettaz, and Yves de Roten. 2013. "Statistics in science and in society: From a state-of-the-art to a new research agenda." *Public Understanding of Science* 22 (7):768-784. doi: 10.1177/0963662513495769.
Wacquant, Loïc J. D. 2004. *Body & soul: Notebooks of an apprentice boxer*. New York: Oxford University Press.
Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective dynamics of 'small-world' networks." *Nature* 393 (6684):440-2. doi: 10.1038/30918.
Weber, J. L., and E. W. Myers. 1997. "Human whole-genome shotgun sequencing." *Genome Research* 7 (5):401-409.
Weber, Max. 1976. *Writschaft und Gesellschaft*. 5 ed. Tübingen: Mohr Siebeck.
Weber, Max. 1988. *Gesammelte Aufsätze zur Religionssoziologie 1*. Tübingen: Mohr.
Weber, Steve. 2004. *The success of open source*. Cambridge, MA: Harvard University Press.
Weeden, K. A. 2002. "Why do some occupations pay more than others? Social closure and earnings inequality in the United States." *American Journal of Sociology* 108 (1):55-101. doi: 10.1086/344121.
Weick, K. E. 1990. "The Vulnerable System: An analysis of the Tenerife Air Disaster." *Journal of Management* 16 (3):571-593. doi: 10.1177/014920639001600304.

Weiss, Linda. 2014. *America inc.?: Innovation and enterprise in the national security state.* Cornell University Press.
Westerhoff, Hans V., and Bernhard O. Palsson. 2004. "The evolution of molecular biology into systems biology." *Nature Biotechnology* 22 (10):1249-1252. doi: 10.1038/nbt1020.
White, Harrison C., Scott A. Boorman, and Ronald L. Breiger. 1976. "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions." *American Journal of Sociology* 81 (4):730-780.
Whitford, Josh. 2002. "Pragmatism and the untenable dualism of means and ends: Why rational choice theory does not deserve paradigmatic privilege." *Theory and Society* 31 (3):325-363.
Whitford, Josh. 2005. *The new old economy: Networks, institutions, and the organizational transformation of American manufacturing*. Oxford; New York: Oxford University Press.
Whitman, James Q. 2004. "The Two Western Cultures of Privacy: Dignity versus Liberty." *The Yale Law Journal* 113 (6):1151-1221.
Widdicombe, Lizzie. 2014. "The Programmer's Price." *The New Yorker*, 11.
Wilensky, H. L. 1964. "The professionalization of everyone." *American Journal of Sociology* 70 (2):137-158. doi: 10.1086/223790.
Wimmer, Andreas. 2013. *Waves of war: Nationalism, state formation, and ethnic exclusion in the modern world*, *Cambridge studies in comparative politics*. Cambridge, England; New York: Cambridge University Press.
Wu, Tim. 2014. "The new Aaron Swartz documentary at Sundance." *The New Yorker*.
Wynne, Brian. 1992. "Misunderstood misunderstanding: Social identities and public uptake of science." *Public Understanding of Science* 1 (3):281-304. doi: 10.1088/0963-6625/1/3/004.
Zelizer, Viviana A. Rotman. 2005. "The purchase of intimacy." In. Princeton, N.J.: Princeton University Press,. http://www.columbia.edu/cgi-bin/cul/resolve?clio9370656.
Zerubavel, Eviatar. 1979. *Patterns of time in hospital life: A sociological perspective*. Chicago: University of Chicago Press.
Zerubavel, Eviatar. 1982. "Easter and Passover: On Calendars and Group Identity." *American Sociological Review* 47 (2):284-289. doi: 10.2307/2094969.
Zerubavel, Eviatar. 1992. *Terra cognita : the mental discovery of America*. New Brunswick, N.J.: Rutgers University Press.
Zhou, X. G. 2005. "The institutional logic of occupational prestige ranking: Reconceptualization and reanalyses." *American Journal of Sociology* 111 (1):90-140. doi: 10.1086/428687.
Zhou, Xueguang. 1993. "Occupational Power, State Capacities, and the Diffusion of Licensing in the American States, 1890 to 1950." *American Sociological Review* 58 (4):536-552. doi: 10.2307/2096075.

Appendix 1 Detailed distribution of organizational characteristics

Year-Q	prof	block(N_o)	Year Founded	
			Mean (Std.)	Median (IQR)
2010-4	fa	0 (6)	1956 (44)	1956 (44)
		1 (4)	1988 (14)	1987 (14)
		2 (1)	1889 (0)	1889 (0)
	att	0 (9)	1977 (32)	1990 (25)
		1 (7)	1967 (36)	1982 (21)
		2 (9)	1984 (26)	1996 (9)
	ds	0 (8)	1956 (54)	1974 (6)
		1 (8)	1998 (8)	1998 (10)
		2 (14)	1975 (50)	1996 (17)
	ra	0 (4)	1970 (28)	1966 (32)
		1 (12)	1954 (57)	1983 (61)
		2 (13)	1941 (38)	1942 (54)
		3 (5)	1956 (51)	1982 (58)
2011-4	fa	0 (6)	1962 (33)	1960 (44)
		1 (22)	1953 (55)	1975 (51)
		2 (17)	1962 (52)	1979 (32)
		3 (14)	1938 (64)	1946 (122)
	att	0 (16)	1954 (54)	1975 (40)
		1 (17)	1977 (36)	1989 (16)
		2 (27)	1967 (47)	1983 (46)
		3 (9)	1977 (13)	1977 (16)
	ds	0 (6)	1982 (25)	1988 (19)
		1 (16)	1949 (62)	1975 (111)
		2 (20)	1990 (29)	2000 (14)
		3 (16)	1997 (9)	1999 (9)
		4 (3)	1985 (10)	1980 (12)
	ra	0 (27)	1948 (61)	1979 (74)
		1 (26)	1961 (62)	1988 (45)
		2 (14)	1966 (56)	1990 (46)
		3 (13)	1967 (65)	2000 (38)
2012-4	fa	0 (46)	1962 (55)	1982 (32)
		1 (45)	1964 (48)	1987 (53)
		2 (16)	1952 (60)	1979 (36)
	att	0 (42)	1963 (47)	1989 (81)
		1 (38)	1960 (51)	1985 (72)
		2 (33)	1965 (47)	1985 (44)
	ds	0 (38)	2000 (13)	2003 (10)
		1 (57)	1992 (13)	1996 (16)
		2 (36)	1983 (35)	2000 (23)
	ra	0 (21)	1950 (70)	1984 (56)
		1 (23)	1963 (53)	1973 (48)
		2 (68)	1976 (44)	1994 (18)
		3 (10)	1936 (77)	1944 (88)
		4 (2)	1999 (0)	1999 (0)
-	-	-	-	-
-	-	-	-	-

Table A1.1. Table showing summary statistics of age of organizations that post relevant job descriptions in each period and cluster

Year-Q	prof	block(N_p)	2-10	11-50	51-200	201-500	501-1000	1001-5000	5001-10000	10001+
2010-4	fa	0 (6)	-	0.17	0.17	-	0.17	0.33	0.17	-
		1 (5)	-	-	-	0.2	-	-	0.2	0.6
		2 (1)	-	-	-	-	-	-	-	1.0
	att	0 (9)	0.22	0.22	0.11	-	-	0.33	-	0.11
		1 (7)	-	0.29	0.14	-	-	0.14	-	0.43
		2 (11)	-	0.09	-	0.09	-	-	0.27	0.55
	ds	0 (8)	-	-	0.12	0.12	-	0.25	-	0.5
		1 (9)	0.11	-	0.22	-	0.11	-	-	0.56
		2 (16)	-	0.06	0.06	0.19	-	0.12	0.06	0.5
	ra	0 (8)	-	-	-	-	-	0.12	0.5	0.38
		1 (12)	-	-	0.25	0.08	-	0.08	0.08	0.5
		2 (13)	-	-	0.08	-	0.08	0.23	-	0.62
		3 (6)	-	-	-	0.17	-	-	-	0.83
2011-4	fa	0 (14)	-	0.07	-	-	-	0.07	0.29	0.57
		1 (24)	-	-	0.12	0.12	-	0.25	0.08	0.42
		2 (18)	-	0.17	0.17	-	0.06	0.17	-	0.44
		3 (16)	0.06	0.06	0.12	0.12	-	0.12	0.12	0.38
	att	0 (16)	-	0.12	0.06	-	0.06	0.25	0.06	0.44
		1 (23)	0.04	-	0.04	0.04	0.17	0.17	-	0.52
		2 (29)	-	0.07	0.07	0.07	0.07	0.21	0.1	0.41
		3 (10)	0.1	0.1	-	-	-	-	0.5	0.3
	ds	0 (7)	-	-	-	-	-	0.71	-	0.29
		1 (34)	-	-	0.09	-	-	0.06	0.06	0.79
		2 (20)	-	0.1	0.2	0.05	0.05	-	0.1	0.5
		3 (26)	-	-	0.08	0.35	-	0.23	-	0.35
		4 (3)	-	-	-	0.33	-	-	-	0.67
	ra	0 (40)	-	-	-	0.08	-	0.08	0.08	0.78
		1 (28)	-	-	0.11	0.07	-	0.11	0.11	0.61
		2 (25)	-	0.08	0.08	-	-	0.12	0.44	0.28
		3 (21)	-	-	0.05	0.19	-	-	0.05	0.71
2012-4	fa	0 (57)	-	0.05	0.07	0.09	-	0.18	0.14	0.47
		1 (58)	0.03	-	0.05	0.02	0.05	0.17	0.16	0.52
		2 (56)	-	-	-	0.04	0.02	0.04	0.12	0.79
	att	0 (53)	-	0.06	0.04	0.06	0.08	0.13	0.04	0.6
		1 (41)	-	0.02	0.02	0.12	0.07	0.27	0.05	0.44
		2 (46)	0.04	0.07	0.04	0.07	-	0.15	0.13	0.5
	ds	0 (66)	0.02	0.06	0.14	0.29	0.06	0.05	0.05	0.35
		1 (70)	-	0.06	0.04	0.2	0.01	0.19	0.04	0.46
		2 (43)	0.02	0.05	0.07	0.05	0.05	0.16	0.02	0.58
	ra	0 (43)	-	-	-	0.02	0.02	0.07	0.07	0.81
		1 (81)	-	0.01	-	0.07	-	0.04	0.11	0.77
		2 (89)	-	0.11	0.03	0.22	0.01	0.18	0.02	0.42
		3 (21)	-	-	-	-	-	0.1	0.05	0.86
		4 (7)	-	-	-	0.14	-	-	-	0.86
-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-

Table A1.2. Distribution of job descriptions across organization size (number of employees) categories by period and cluster

Year-Q	prof	block (N_p)	Par	Non Pro	Gov Age	Pub Com*	Pri Hel	Edu	Sel Own
2010-4	fa	0 (6)	0.17	-	-	0.33 (0.51)	0.33	0.17	-
		1 (5)	-	-	-	0.8 (1.0)	0.2	-	-
		2 (1)	-	-	-	1.0 (1.0)	-	-	-
	att	0 (9)	0.22	-	-	0.22 (1.01)	0.56	-	-
		1 (7)	-	-	-	0.57 (0.75)	0.43	-	-
		2 (11)	-	0.18	-	0.73 (0.75)	0.09	-	-
	ds	0 (8)	-	0.12	-	0.5 (0.75)	0.38	-	-
		1 (9)	-	-	-	0.67 (0.83)	0.33	-	-
		2 (16)	-	-	-	0.62 (0.91)	0.31	0.06	-
	ra	0 (8)	-	0.12	-	0.88 (0.8)	-	-	-
		1 (12)	-	-	-	0.67 (0.4)	0.33	-	-
		2 (13)	-	0.15	-	0.62 (0.99)	0.15	0.08	-
		3 (6)	0.33	-	-	0.5 (0.4)	0.17	-	-
2011-4	fa	0 (14)	-	-	-	0.5 (1.0)	0.5	-	-
		1 (24)	-	0.04	-	0.46 (1.0)	0.5	-	-
		2 (18)	-	-	-	0.56 (0.89)	0.44	-	-
		3 (16)	0.12	0.12	-	0.31 (1.01)	0.44	-	-
	att	0 (16)	0.12	-	-	0.56 (1.0)	0.31	-	-
		1 (23)	0.22	0.04	-	0.39 (0.78)	0.35	-	-
		2 (29)	-	0.03	-	0.59 (0.99)	0.38	-	-
		3 (10)	-	0.1	-	0.8 (0.87)	0.1	-	-
	ds	0 (7)	-	-	0.14	0.57 (0.75)	0.29	-	-
		1 (34)	-	-	-	0.79 (0.97)	0.21	-	-
		2 (20)	-	-	-	0.65 (0.81)	0.35	-	-
		3 (26)	-	-	-	0.46 (0.92)	0.54	-	-
		4 (3)	-	-	-	0.67 (0.6)	0.33	-	-
	ra	0 (40)	0.1	0.02	-	0.52 (0.78)	0.35	-	-
		1 (28)	-	-	-	0.5 (0.63)	0.5	-	-
		2 (25)	-	0.04	-	0.72 (0.94)	0.24	-	-
		3 (21)	0.1	-	-	0.57 (0.5)	0.33	-	-
2012-4	fa	0 (57)	0.04	-	-	0.61 (0.95)	0.35	-	-
		1 (58)	0.05	0.12	-	0.48 (0.9)	0.34	-	-
		2 (56)	-	0.09	-	0.79 (0.88)	0.12	-	-
	att	0 (53)	0.13	0.02	-	0.42 (0.94)	0.38	0.06	-
		1 (41)	0.05	0.05	-	0.54 (0.95)	0.34	0.02	-
		2 (46)	0.11	-	-	0.39 (1.0)	0.48	-	0.02
	ds	0 (66)	0.02	-	-	0.45 (0.81)	0.53	-	-
		1 (70)	-	0.03	0.03	0.56 (0.87)	0.39	-	-
		2 (43)	0.02	0.07	-	0.53 (0.88)	0.37	-	-
	ra	0 (43)	0.14	0.09	-	0.4 (0.63)	0.37	-	-
		1 (81)	-	0.04	-	0.79 (0.52)	0.17	-	-
		2 (89)	0.04	0.07	-	0.39 (0.72)	0.49	-	-
		3 (21)	0.05	0.05	-	0.71 (0.35)	0.19	-	-
		4 (7)	-	-	-	0.86 (0.7)	0.14	-	-
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-

Table A1.3. Distribution of relevant job descriptions across organizational types by period and cluster

Appendix 2 Systems biology and the HGP case

I analyze systems biology to consider a stock of knowledge that addresses arcane scientific problems. Systems biology directly emerges from branches of academic biology (Westerhoff and Palsson 2004). It pursues a singular goal: Understanding how life unfolds (Hood 2008, 21).[112] Organizationally, systems biologists work in laboratories and research groups. While law has been invoked as a signal case in several institutionalist arguments (Stinchcombe 2001, Abbott 1988), system biologists' work in laboratories makes this a context resembling those in which the significance of trust and other forms of unscripted interactions for constructing knowledge have been demonstrated (e.g., Collins 1998).

Systems biology studies how life unfolds in its various forms. It operates through projects that focus on a specific problem for which they organize the "capture, validation, storage, analysis, integration, visualization, and graphical or mathematical modeling of data sets" (Hood 2008, 22). Explaining life in this way produces arcane knowledge, such as understanding "cis– and trans–regulatory networks in sea urchins" (Ideker, Galitski, and Hood 2001) or "immune response in mammals" (Chuang, Hofree, and Ideker 2010). Another project, also characteristic of the field in that it is conducted by a research group and contributes to the larger goal of understanding life, aimed to decode the human genome (HGP). Because the HGP has also been widely written about, I utilize this instance to recover some key patterns by which system biologists construct their stock of specialized knowledge.

The HGP's main questions sought answers to specific problems. Its designers saw it as a strategy to develop new scientific knowledge addressing several important problems, including common diseases such as cancer and schizophrenia (Lander 1996). Answering such large questions required a thorough division of labor that organized tasks in more specific questions (Lander et al. 2001).[113] These organizational challenges gave rise to, but also facilitated solutions of problems not immediately addressing the guiding question of the project. When debating how to take the HGP from testing on model organisms to the phase of human genome sequencing, for example, the conventional practice to clone regions of the genome, sequence them and then generate the human genome from them, was

[112] The question of what constitutes life is not unique to systems biology. It has guided biology in general at least through the last century (Keller 2002). My argument, however, is not about unique or novel ideas but relations between ideas.

[113] The HGP distributed work across academic institutions in different countries. A privately funded effort remained within one facility (Venter et al. 2001). The organizational problem of integrating distributed labor remains.

challenged. An alternative strategy proposed to avoid arguably unnecessary redundancy of sequencing overlapping regions and instead suggested to target the entire genome at once (Weber and Myers 1997). Arguments brought up against this “monolithic” approach emphasized the possibility of consequential bottlenecks when proceeding from the sequencing to the finishing stage and its hostility toward dividing the work among members of the group (Lander et al. 2001, 864). It was seen to introduce the risk of being incompatible with the existing work, without promising significant impact (Green 1997). Viewing the technical details jointly reveals scholarly strategies that generate increasing specialization in problems that collectively inform the composition of the human genome as a key step toward understanding life.

System biologists, here viewed through the HGP, engage with their stock of knowledge as they contribute details to abstract ideas. They develop these abstractions, such as the sequencing code just described, in projects that relate to each other through the questions they jointly aim to answer. Understanding a general problem, such as cancer, narrows into increasingly detailed and inaccessible discussions of auxiliary questions, which are still critical. This strategy noticeably differs from the more vivid illustrations common in legal scholarship (described in the main text). The division of labor across HGP contributors facilitates that tasks become so specialized that experts lose sight of the heterogeneity of the field. The specialization of research narrows further within the research laboratories contributing to the HGP. Recalling evidence from the literature, we should expect that informal relationships between principle investigators, postdocs, students, technicians and other lab affiliates facilitate deviation from institutionally scripted patterns of knowledge production. The singular aim of the field is thus seen relative to, and potentially in competition with, more immediate concerns.

A2.1 History and composition

Systems biology has consolidated following a phase of systematic fragmentation. Figure 10.2 (in the main text) shows increasing fragmentation beginning in 1990, and a decrease of modularity again in the early 2000s. These trends mirror its documented history. The period of increasing modularity corresponds to the most active period of the HGP, described above, a formative moment of the systems biology field in general. The project, which was completed in 2001, has raised new questions and has led systems biology to discover and integrate new tools to address familiar problems. That we see this trend in the sample of systems biology instructors, which is not restricted to HGP contributions, is still consistent

with the reported formative impact the HGP had on systems biology (Hood 2008). The informal expertise argument suggests that research groups rely on internal trust to work out these innovations. These findings imply that the groups consider knowledge that their field does not recognize to be relevant more broadly (recall table 10.2 in the main text).

A2.2 Scientific discipline

Systems biology expertise is disciplined, with exceptions. Simulations show that if publications preferentially cite already popular work this generates a structure that closely resembles the observed order (see the green square marker on row 2 in figure 10.3 in the main text, indicating a similar modularity score of the simulation compared to the observed structure, reported in row 1). Divergence from this practice can be specified quantitatively as unscripted reshuffling of between 11 and 12% of the simulated references (figure A3.1a, also compare to the 20% necessary to recover law, figure A3.1b). That generally shared assessment of contributions reproduces much of the observed structure is consistent with those accounts in the literature that find institutional scripts guiding academic work. The amount of rewiring still indicates some activities not following the scripted practice in which contributions focus on some main ideas. Systems biology utilizes new technologies, which could induce such informal integration. This finding reflects previous observations that building such technologies requires informal expertise to produce valid data (Collins 1998). The relatively small amount of rewiring, however, reveals this unscripted knowledge in its institutional context, marked also by consistent agreement on important pieces.

Appendix 3 Supplementary analyses of citation networks

Reconciling observed and simulated structures of law and systems biology

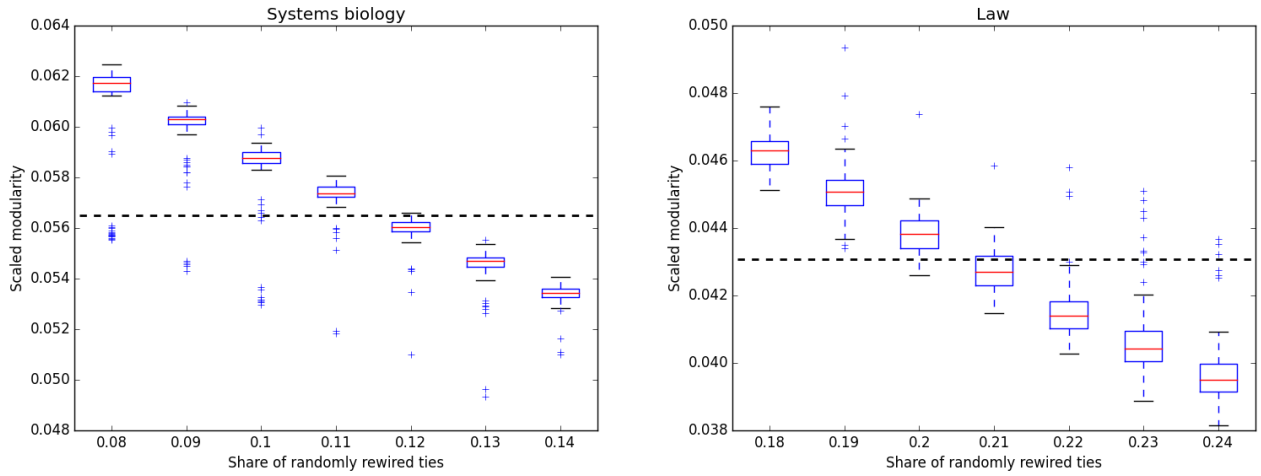

Figure A3.1. Figure showing distribution of scaled modularity scores of unscripted rewiring iterations at varying shares of ties for (a) systems biology and (b) legal scholarship and the scaled modularity of the observed networks (dashed line).

Figure 3 in the main text shows that the preferential citation rule reproduces systems biology more closely than law. Here I analyze the specific degree of unscripted rewiring that reconciles the simulated with the observed structure in each case. I test this by randomly rewiring the networks simulated from the Matthew Effect process of preferential citation. Figures A3.1a and A3.1b show results for systems biology and law, respectively. The box plot shows the distribution of scaled modularity scores (Y-axis) over 100 rewiring iterations. The X-axis indicates the percentage of rewired reference ties. The dashed line indicates the scaled modularity score of the observed networks. Reconciling the observed and simulated citation structures underlying legal scholarship requires rewiring of between 20 and 21% of the simulated structures. For systems biology it suffices to rewire between 11 and 12% of the reference ties.

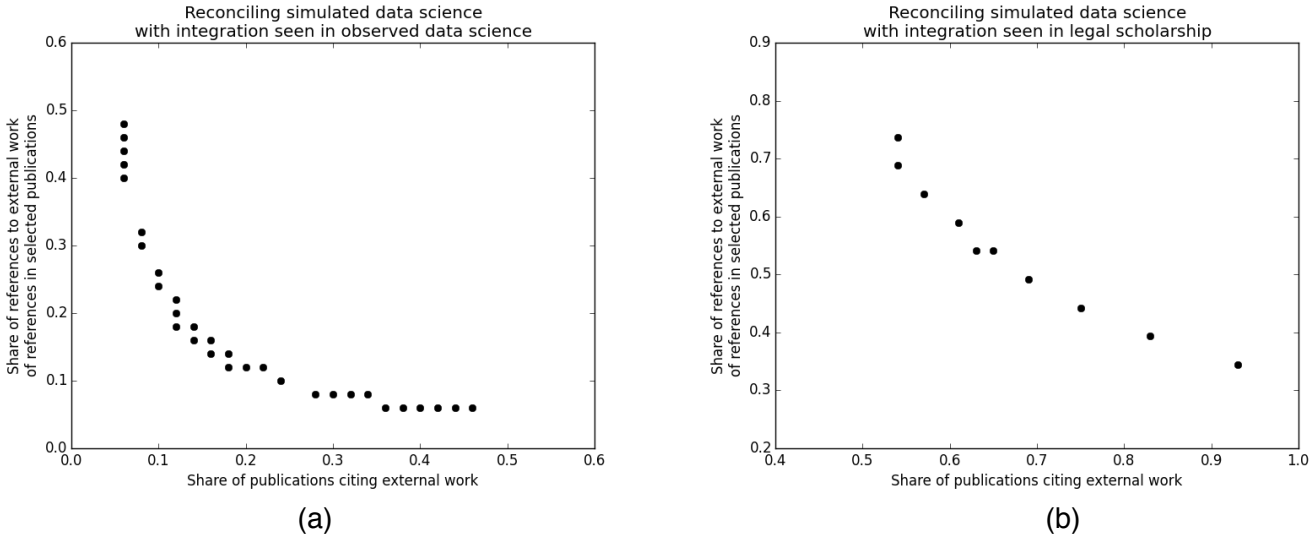

Figure A3.2: Distribution of combinations of publications and references in them that consider work from outside the field such that they reconcile (a) the simulated and observed structure of data science and (b) that of simulated data science and observed legal scholarship.

Reconciling observed and simulated structures of data science

Universities partially reorganize their existing departmental structures in order to design data science programs from them. Instructors often remain affiliated with their disciplinary background in statistics, computer sciences, and so on, as they train data science students. As the main part describes, the data science simulations in this analysis accordingly generate the Matthew Effects by constraining the preferential citation rule to two distinct subsets of the citation network underlying the observed data science structure. Two different aspects need to be considered. Integration could follow from few publications if each devotes substantial attention to a different field, or from many publications that marginally consider other fields. Figures A3.2a and b consider combinations of the two. The X-axis shows the share of publications considering work from another discipline and the Y-axis shows the share of references in these publications that invoke work from another field. In figure A3.2a, the markers indicate combinations at which rewiring iterations generate a distribution of structures that reconciles the simulated with the observed scaled data science modularity. Figure A3.2b shows the results when extending the analysis to reconcile simulated data science networks with the observed integration of legal scholarship.