# Data and Society

Beaulieu, Anne ; Leonelli, Sabina

*Submitted version of book manuscript, published by SAGE in November 2021 and available at this link: https://us.sagepub.com/en-us/nam/data-and-society/book269709*

# Data and Society: A Critical Introduction

**Anne Beaulieu**, University of Groningen

**Sabina Leonelli**, University of Exeter and Wissenschaftskolleg zu Berlin

*To Claudette and Gérard Beaulieu, and to all who also raise critical thinkers.*

## Acknowledgements

This book builds on our work on data from the past two decades. It draws on material from several research projects as well as countless interactions with generous colleagues, collaborators and audiences whose insights and engagement has informed and inspired us.

Many students from around the world contributed to our thinking about data through their astute questions, diverse insights and perspectives, and lively in-class discussions. They include the Masters students in the courses Big Data and Energy (University of Vienna) and Big Data in a Digital Society (University of Groningen), and especially the students in Introduction to Data and Data as Evidence (University of Groningen) and the postgraduate module Data Ethics and Governance at the University of Exeter.

A number of colleagues also acted as sparring partners and we are grateful for their contribution to sharpening the concepts in this book and for inspiring us through their own research. In The Netherlands, thanks go expressly to Ludo Waltman, Salome Scholtens, as well as Oskar Gstrein, Daniel Feitosa, Andrej Zwitter, and all other members of the Data Research Centre. In Exeter, thanks go to colleagues at the Exeter Centre for the Study of the Life Sciences (Egenis) and the Institute for Data Science and Artificial Intelligence (IDSAI), especially the wonderful participants to Data Crunch meetings where the book project was discussed; Lora Fleming, Gavin Shaddick, Alberto Arribas and Hywel Williams, who are making the data world a better place; John Dupré,

Rachel Ankeny, Gail Davies and Brian Rappert, who unfailing provided inspiration and support; and the brilliant PhD students at the Environmental Intelligence Centre for Doctoral Training.

As a project, this book benefited from the generous help of Andrej Zwitter who established a fruitful link with Sage. Sage editors Robert Rojek and Natalie Aguilera provided prompt and helpful steering for the project throughout, and procured six referee reports that contained extremely useful, timely and constructive feedback – many thanks to these referees too. Trish Nowak provided excellent editorial assistance in the final weeks of this project. Other colleagues provided valuable feedback on different chapters that improved the structure and clarity of the material. We want to acknowledge the contributions of Arthur Vandervoort, Marthe Stevens, Clarisse Kraamwinkel, Malcolm Campbell-Verduyn and Esther Hoorn, long-time collaborator on getting things done with data. Selen Eren and Joonas Lindeman were supportive early readers.

In an academic climate that values research activities, this book required us to make a big commitment to teaching—a commitment that was also fuelled by the work of others. In Groningen, we would like to note the contributions of Jasper Knoester and Gerard Renardel de Lavalette for championing a course on Big Data and Society in the Masters in Computer Science, of the many colleagues who helped build the minor programme on data, especially Ronald Stolk, for spearheading what became Data Wise: Data Science in Society, René Veenstra who helps steer this complex programme, and Gert Stulp who, in all ways, is an invaluable colleague and the ideal programme co-coordinator. In Exeter, we thank the College of Social Sciences and International Studies, the Department of Sociology, Philosophy and Anthropology, and the Institute for Data Studies and Artificial Intelligence - especially its director Richard Everson - for recognising the centrality of this work to data science training and supporting the expansion of our research and teaching activities. We are grateful to Chee Wong, Jill Williams and Katie Finch for their vital administrative support of data studies activities at the university through Egenis and IDSAI.

Inspiration for our teaching has come from many corners and we are lucky to be surrounded by colleagues committed to interdisciplinary teaching, to learning for both students and lecturers, and to creating optimal conditions for students' development. Anne would especially like to thank Malvina Nissim, Sepideh Yousefzadeh, Elena Cavagnaro, Berfu Unal, Tassos Sarampalis, Indira van der Zande, Ineke Visser, Amaranta Luna Arteaga, Engelien Reitsma, Piet Bouma, Bernike Pasveer, Nishant Shah and Hanny Elzinga.

Finally, the main parts of this book were written during the COVID-19 pandemic. This created considerable challenges to completing the manuscript, including health issues and bereavement for us both, and we would like to thank our husbands and children, Maarten Derksen and Félix Derksen and Michel, Leonardo and Luna Durinx, for keeping us (relatively) sane through it all. Anne would also like to thank the Post-pandemic University, participants to Wednesday Wine, the informal 'coalition of the willing' at the University of Groningen, and Jacob Veenstra; while Sabina thanks her closest friends and colleagues for all their support, strength and unfailing sense of humour.

A few passages in this book draw from the following publications:

Beaulieu, A. 2021. Data practices and SDGs: Organising Knowledge for Sustainable Futures. In Maja Hojer Bruun, Dorthe Brogaard Kristensen, Rachel Douglas-Jones, Cathrine Hasse, Klaus Høyer, Brit Ross Winthereik and Ayo Wahlberg Eds., *Handbook for the Anthropology of Technology*, Palgrave.

Gstrein, Oskar and Anne Beaulieu (under review) "How did this person get in the data? A discussion of the multiple conceptualizations of privacy in the digital age."

Leonelli, S. (2017) *Biomedical Knowledge Production in the Age of Big Data.* Report for the Swiss Science and Innovation Council, published online November 2017: http://www.swir.ch/images/stories/pdf/en/Exploratory_study_2_2017_Big_Data_SSIC_EN.pdf

Leonelli, S. (2018) *La Ricerca Scientifica Nell'Era Dei Big Data*. Meltemi Editore. ("Scientific Research in the Age of Big Data"). ISBN 9788883539015.

Leonelli, S. (2021) Data Science in Times of Pan(dem)ic. *Harvard Data Science Review* 3(1) https://doi.org/10.1162/99608f92.fbb1bdd6

Leonelli S, Lovell B, Fleming L, Wheeler B and Williams H. (forthcoming) From FAIR data to fair data use: Methodological data fairness in health-related social media research.

## Overview of book

In today's digital society, a critical understanding of data is essential for all. Knowledge about data is often split into areas of expertise, so that processes that span algorithms, servers, users and institutions are rarely discussed coherently and accessibly. Data and Society: A Critical Introduction presents a set of concepts to assess how data shapes science, policy, and politics, including how data is turned into metrics that are used to make decisions. It connects data as a highly technological practice to broad social questions of evidence, innovation, and knowledge.

The book provides an analytical framing to understand the role of data in contemporary society and foster good data practices. Our analysis is grounded on the following three ideas:

1. Data are not an autonomous force nor a unidimensional technical fix, and the use, valuation, circulation, and deployment of data are shaped by social and material factors, including social institutions and technologies.
2. Nearly all areas of professional and academic work involve interactions with data science, and the skills to relate, evaluate and shape data practices are necessary to be able to exercise one's expertise responsibly.
3. The ability to write code and develop algorithms contributes but is not sufficient to understanding and critically assessing data practices. Another essential

component of training in data use involves the conceptual and methodological toolkit developed within the social studies of science broadly conceived (including critical data studies, data ethics and the history, philosophy and social studies of science and technology).

Building on these key insights, the book addresses the growing attention to the social embedding of data across different settings, from business to policy and government, from sports to health and climate change. It explains the challenges that such embedding brings both for the governance data flows and for the technical management and use of data.

This book is intended as an interdisciplinary introductory textbook for advanced undergraduates or graduate students that connects the phenomenon of datafication and related technologies to social, technological and economic change. Its conceptual framework relates ideas and principles with concrete cases, to help illustrate the growing importance of data in different spheres of knowledge production and its implications for a wide variety of sectors. To this aim, the book is structured around four sets of practices around data, with a series of *data stories* used to discuss salient concepts to concrete issues. The data stories present details about a specific use of data and ask questions about different aspects and implications of that case. In doing so, they exemplify ways to question and scrutinize the broader social implications of data work and highlight how technical aspects of data practices are entwined with institutions, users, regulations, business models and cultural norms. You are invited to read the data stories and think about the questions that they pose in two stages: once before reading the related section of the book, and a second time after having read the materials in the section, which will help you to articulate your own answers to the questions being raised.

## Introduction

Let us start with an every-day story common to the world of internet users. It is April 2021 in the United Kingdom. Thirty-year-old Lara is a healthy and active individual. In the evening she occasionally watches Amazon prime shows, with a strong preference for science fiction series – but generally prefers spending her free time talking to friends and doing sports. Her normal routine is disrupted when she suddenly falls ill and has to stay in bed for two weeks to recover. During that time, she is in such bad condition (headache, fatigue, mental confusion) that all she can manage to watch are costume dramas and teenage romance films, whose pace is slower and whose soundtrack is less jarring for her headache. Once she is back in shape and able to return to her preferred lifestyle, Lara notices that the list of movies recommended to her has changed, and she is finding it harder to identify series that she might like to watch. Other parts of her Amazon account have changed too, with insistent advertising for clothes and products that she dislikes. Her Google account also seems affected by the changes, with adverts for Jane Austen-themed holidays and romantic getaways popping up every time she scrolls down. Lara is upset because the internet platforms she uses for online shopping and entertainment no longer reflect her preferences, and what used to feel like useful

shopping tips now feel like useless, intrusive clutter in her online space. Lara is also upset because she had not realised how extensively the information concerning what she watches on Prime would travel to other platforms and online services.

This everyday situation raises many questions. Is Lara right to feel upset in this situation? Is there a problem here, and what is it? Do Lara's watching preferences constitute sensitive and/or personal data? Are these preferences valuable data, and for whom? What can or should Lara do about the mismatch she experiences? Is this situation legal? Is it a necessary condition for the provision of the streaming service? Is it right that Lara should have to deal with this? Is it possible to "fix" this situation by improving the system's responsiveness to Lara's change in preferences, so that it can be updated more quickly to her changing circumstances? Or should Lara simply adapt her behaviour to the system, so that she does not get caught up in this way again in the future?

Let us now consider a different story, which happened in the early 2010s to an American scholar called Mary Ebeling – who then went on to write an important book about her experience (Ebeling, 2016). After years of attempts, Mary was delighted to be pregnant and could not wait for the birth of her baby. Tragically, however, during the eight week of her pregnancy she suffered a miscarriage and lost the baby. As so many other mothers-to-be finding themselves in that horrible situation, she came home from the hospital distraught, only to find advertisement on her doorstep that was specifically targeted to pregnant women. Despite her complaints to the companies responsible for the unwanted correspondence, her letterbox continued to fill with advertising and samples from baby products companies. On the week in which the baby should have been born, she even received congratulation notes complimenting her on the birth of a healthy child. The advertising campaign continued for years, tracking the milestones and celebrations that her daughter would have enjoyed had she survived. Mary was devastated by these constant reminders of the baby she had lost and kept pleaded with the companies responsible to stop sending her these unwanted gifts – to no avail. Exasperated, Mary did some research and discovered that while she was pregnant, the hospital sold her personal data to a private company as part of a clinical trial that she had agreed to participate in, as a way to get support for her pregnancy. The company running the trial in turn sold Mary's data to a data trading company, which sold it off again to a number of baby products companies. The data were never updated with news of the baby's death, which is why she kept receiving the merchandise. In her initial attempts to understand what has happened to her data, Mary was unsuccessful as companies did not want to release information about what data they own, and who they acquired the data from. Mary managed to elicit that information only when she revealed her ordeal as a bereaved mother.

This is clearly a situation where something went very wrong in the handling of sensitive medical data. Was the hospital justified in selling Mary's data in the first place? It turned out that she did give her consent to participate in the clinical trial and she signed an agreement that enabled the company to sell her data to third parties. Her participation in the trial may have been beneficial to the development of medical research, which was the reason why she signed up. Does this make the other ways in which her data were used ok? Is participation in medical research equivalent to consenting to data being used for advertising? Are these two things necessarily related? Can we still distinguish research from the commodification of reproduction? What does this case tell us something about the entwinement of medical care and business in the American context and the commercialisation of health? Could such a situation be avoided, and if so, how?

Let us now move to a third story, which could well be described as a data triumph. For many years, agronomic institutions around the world have worked on improving systems for sharing data about plant pathogens and diseases, so as to be able to map the spread of new pests and improve understanding of how to treat the affected crops. The decrease in disease in crops has a direct impact on hunger and malnutrition. In 2019, a consortium of public and private research institutions released an app called PlantDisease. This app could be downloaded for free on the smartphones of farmers around the world and helped them to identify diseases in their crops in a timely manner, as well as providing tips on how to treat those diseases. By providing such important information to farmers in remote areas of the globe, the app is helping to boost food production worldwide as well as supporting the livelihoods of farming communities.

We can think of this is a case of data helping to feed the world and this seems the best of what data can do for humanity. Is there any possible drawback to such a development? What happens if the information provided by the app is unreliable or faulty – who takes responsibility for the effects? Also, is the information provided though this app simply a series of "facts" about plant disease? There are actually many different approaches to understanding and treating plant diseases within agronomic research. Some privilege technology-driven interventions, like using targeted pesticides or genetic modification of seeds. Others favour environmental interventions, like choice of fertilizer or ways of selecting and managing crops. Notably, the technology-driven interventions are costly and boost the profits of corporations based in the Global North.[1] Does the app bear a responsibility to inform farmers of alternative paths, and of the extent to which different companies may profit from the ways in which agricultural production is managed? Is providing alternatives and context a way to help farmers with their decisions? Or does it generate confusion, thus defeating the very purpose of the app to provide immediate, easy-to-follow instructions? Who owns the app, and does it matter that the underlying code is not available as Open Software for others to scrutinize and re-use?

The questions raised by these cases are difficult ones and do not have easy answers. Yet, they are only a fraction of the questions associated to the technological and social life of data and the ways in which data affect human life. This book aims to provide you with instruments to identify such questions and articulate your own answers to them. It will help you understand the conditions under which data are generated, circulated, traded and used. It is not a technical book about **data science** and artificial intelligence. Rather, it is a book about the interactions between the social and technological aspects of data work.

Before we start, we should say a few words about our backgrounds and motivations for writing this book. We are both scholars working in the broad area of Science & Technology Studies (STS). Beaulieu comes from the social studies of science and Leonelli from the philosophy and history of science. This book project emerges from our experiences in creating Data Wise: Data Science in Society, a minor programme at the University of Groningen; and the training in Data Ethics and Governance for all the postgraduate and professional Data Science degrees at the University of Exeter. The book also builds on our decade-long collaborations with data practitioners working in top research programmes around the world, as well as policy-makers working on **data governance** in a variety of national and international contexts. Through this work, we realized that teaching material concerning data and society at both the undergraduate and postgraduate level was sorely missing. While the academic literature in so-called "data studies" was blossoming, there was no obvious textbook available to introduce students to this emerging field, and thus complement the technical aspects of data science teaching with an introduction to its social components. This book was conceived to fill this gap and support teaching and learning about data in context.

[1] Note that the concepts of Global North and Global South are not meant to refer to strict geographical locations but to the structural and historical inequities between countries and regions of the world in terms of the distribution of power, capital and infrastructure as well as the provision of social services.

# Section I Data in Society

## Summary

This section frames the critical exploration of the production and uses of data in society. It introduces the notions of "datafication", "data work" and "data journeys". It explains their components and illustrates where they arise across different areas of social life. A number of prominent myths around **Big Data** are discussed and the various characteristics of data are described. We introduce the cycle of knowledge production and show how this cycle helps to better define data and how to understand its roles. The section concludes by showing how data work involves many seemingly small decisions that have ethical implications.

## Learning Objectives

This section will help you to
1. Understand datafication processes and components
2. Analyse how knowledge, data and technology relate
3. Appreciate that different kinds of knowledge exist across time and spheres
4. Identify how and where knowledge and its trustworthiness arise as issues in everyday life
5. Understand the pervasive role of ethics in data journeys

# Chapter 1. Data in Society

## Summary

Most contemporary societies privilege ways of knowing that are grounded on data. Data have not always been so important, but now play a central role in all kinds of important expertise and decision-making processes. In order to understand and use data, it is useful to develop an awareness of what data are, what they can do and what they should do. This chapters provides a number of starting points to deepen your knowledge about data and learn to better evaluate it. It introduces the concept of datafication as a layered approach, sets out the current context in which data has come to matter and discusses the importance of considering aspects beyond technology to evaluate data and its role. The prevalence of ethical decisions across all aspects of data work is explained.

## 1.1 Introduction: Who cares about data?

A key development in recent years has been the increasingly prominent role of data in society. Most activities and interactions that any one individual has in contemporary high-tech society produce traces and generate data; whether using email, shopping online or browsing the internet. This is sometimes called the "**datafication**" of society – that is, the process through which human activities leave a digital trace that is then used. Datafication has two interrelated components: the creation of a trace that is recorded and circulated in the form of data beyond that particular moment and place,

and the further use of such a trace as a meaningful element in other processes. Importantly, datafication is not always evident in our everyday life, and many of our activities are datafied whether we wish it or not. For instance, when watching a movie on a streaming service, data are created that document our choice of movie, the time of day in which we watched it, and whether or not we watched it all in one go. These data are by-products of our activity: we are generally not watching a movie in order to produce the data, and we may not even be aware of that happening. Yet, these data are generated and used in order to understand and predict our behaviour. The label 'Big Data' has been used to refer to the assemblage and use of vast amounts of data created in a variety of different ways, but this covers only some aspects of datafication, as we will see in the next chapter.

Whereas data was long considered to be a by-product of scientific research, it has now become an output in its own right, in research and in many other spheres. Data, in other words, is now a social phenomenon. This is an important reason why data are often referred to as a singular entity in the popular press and in everyday usage of the term. Data has acquired a reputation as an uncountable, a collective noun for an undefined entity. Everybody recognizes the significance of data, but the material and substantive features of data are hard to pin down and understand. When we talk about data as a singular noun, the individual data point no longer matters. What is relevant are the emerging practices in which masses of information underpin decisions, knowledge claims and social perceptions. In this sense, 'data' has come to exist as a fruitful concept in social, political, technological and corporate settings. This is the sense in which we talk about data as 'the new oil', or the 'data deluge', or living in the 'age of data'. This is rather different from the understanding of data a set of observations or measurements to be used as evidence. When referring to data as a phenomenon, as an area of concern for particular social debates, or as an object of study for specific areas of scholarship, we will therefore use the singular in this book.

This is not, however, the only way to understand data. A major reason for scientists to think of data as plural is that there is usually little value in an isolated data point. Data are multiple, collections of objects that are assembled and – crucially – disassembled and re-organized depending on how one wishes to use them. Think of the data that Google holds about users of its email, searching or streaming services: while there may be so many data as to defy human comprehension, each and every one of those data points is important. All these data refer to someone's address, age, and music preference, for instance. Each data point might matter in its own right and can be combined with other data points to fuel different types of inferences – from the size of clothes one may be shopping for online, to the locations of events that may be of interest to a Google user. When we think of data as a plural noun, we highlight the way in which data are constituted, assembled and traded, and the judgments and intentions underpinning any one way to cluster data for use. Where data come from, how they are generated and valued, who works with which data and why: these aspects are indispensable to identify and critically evaluate the multiple roles of data in society. For this reason, we use the plural when we want to stress that data are created in specific contexts, and that they have a provenance, quality, quantity and form, and that they are handled and connected in particular ways. In sum, 'data is' refers to data as a phenomenon, 'data are' refers to a collection of data points.

Scientists, journalists, business people, politicians, policy makers, and governmental institutions all make use of 'Big Data' and 'data-driven approaches' to understand our society and to shape our daily lives. These data come from somewhere, and the means through which they are collected, circulated and analysed strongly affect how data are ultimately interpreted, and for which ends. The first aim of this book is therefore to *unearth the conditions under which data come to have social value*. This includes all the stages of data handling: from the generation of data to their dissemination through databases and infrastructures, their visualisation through models and interfaces, their combination, through to their analysis and interpretation across multiple settings.

Data also go somewhere, and in fact the more they circulate across a variety of settings, the more value they tend to acquire. Data that get stuck within the walls of one laboratory, one company or one government agency will be used much less, and interpreted in a narrower way, than data that travel beyond those walls and are scrutinized by more and more diverse people, analysed through various different tools, and integrated with data coming from other sources. At the same time, the role of data is not uniform across contexts. There are significant differences in how data come to matter. From sports to healthcare, from business to biology, from social media to education — data and data infrastructures have become a dominant force in all these spheres. Yet, what proves a tremendous opportunity in one domain may well raise significant challenges in another. Hence the second aim of this book is to *provide tools to track the diverse journeys of data, understand these differences and use that understanding to guide the management and use of data*.

Who cares about data, their provenance and their journeys? We imagine the audience for this book to include not only data scientists and curators (that is people whose main responsibilities involve the analysis and stewardship of data), but also anybody who needs to manage data as part of their work, be this in industry, policy, social services or any other profession. It is not just data experts who need to care about data and their social role. Anyone whose job includes collecting, managing, and/or interpreting data needs to worry about the implications of their data practices for their business and society at large, and to acquire skills that will empower them to make sensible decisions in that respect. The main audience for this book is what we will call "**data workers**" (See Figure 1.1). Data workers are individuals who may or may not have technical abilities and be directly involved in the development of data analytics, but who are in a position to take decisions concerning what data should be gathered, for which purposes, and in which ways. This book will also support the work of those who decide who owns the data and whether it should be shared (and with whom), and those who decide on how to reuse or repurpose data and on whether further analysis may be appropriate and justified.

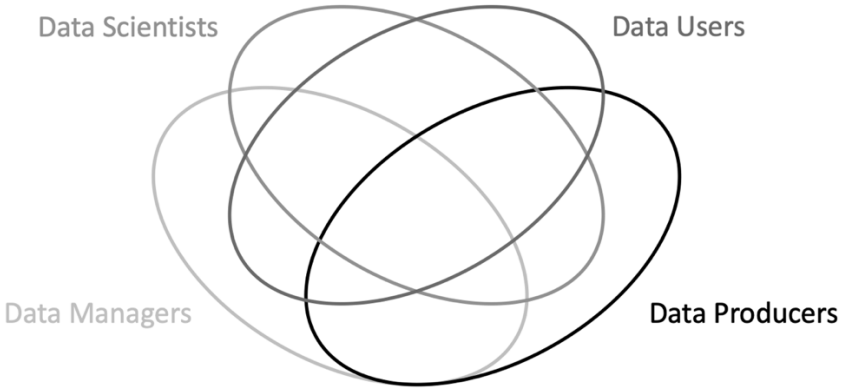



*Figure 1.1 Engagement with data: different roles of data workers that often intersect.*

## 1.2 Datafication and its components

Many Big Data advocates have discussed the datafication of society as centered on the acquisition and technical handling of data (e.g., Mayer-Schönberger and Cukier, 2013). These data have several features, typically including the increased volume of data at hand, the velocity with which they are produced and analysed, and the variety of data types and sources. These discussions often use the labels of the 3Vs (volume, velocity, variety), or modified versions that add further Vs (venue, vocabulary, vagueness, validity, veracity, etc).[2] This focus on data and its features misses necessary aspects of datafication as a process. In other words, it is not just about the data, it is also essential to understand how data comes about, what is done with it and why, and to whom this might matter.

Datafication is in a first instance, the turning of objects and processes into data (Mayer-Schönberger and Cukier, 2013). As Van Dijck and colleagues (Dijck, Poell, and Waal, 2018) show, this 'turning into data' has a number of dimensions. Think of how networked **platform**s render into data many aspects of the world (and our behaviour in it) that had not been formalized in this way before. For example, social networks have become formalized on social media platforms, via the rendering of social ties as digital traces. Our 'friends' on Facebook or the accounts we follow on Twitter are recorded as digital data. Datafication is also the process of rendering activity as quantifiable traces in which patterns can be discovered. For example, the platform LinkedIn makes it possible to create an individual profile, including a photo, and to list one's employment history. LinkedIn keeps track of when users update their profiles and when they change their employment data. The company has identified patterns of activity on users' profiles (such as updating your profile photo) that indicate that an account holder is likely looking for a job and is therefore a good target for job ads. Datafication is also the transformation of interactions into data that can be valued and used for predictive activities. Examples of this are the analysis of public sentiment from tweets and the prediction of electoral outcomes; or the tracking of population movements via

[2] https://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data

localisation of smartphones, and its use to predict and prevent the spread of disease as in the case of COVID-19. Datafication is finally also the extension of the collection of traces for every interaction, even seemingly trivial ones like the direction and speed of a cursor moving over a webpage or the number of corrections in a draft message before it was posted or published. Some examples of this are often based on surprising **correlations**, such as the relation between a user's clicking speed and depression.

What makes these processes possible? The extension of automation, with the proliferation of digital technologies, the willing production of massive amounts of data, and the combination and mobilization of data sets are all important (Rieder and Simon, 2016; Dijck, Poell, and Waal, 2018). But technological possibilities are not sufficient to explain the scope of datafication. As soon as we consider the various environments and practices involved in making and interpreting data, it becomes clear that datafication is not only about data and related computational techniques. For this reason, we propose a layered model that put neither data nor technology at its centre. Datafication is a practice that links at least four necessary elements:

(1) the community of actors (and related institutions) who engage with the data, for example because they handle the data on an everyday basis. Many forms of engagement with data are possible: some may use it, innovate with it or even oppose it. This community could be constituted by one or many social groups and have various degrees of cohesion depending on whether or not those involved know each other, have shared values, backgrounds and goals, and work in similar conditions. For instance, a research group working within a small academic field may be highly cohesive, since all its members are likely to have received similar training by the same mentors, and share an interest in - and understanding of – a narrow and well-specified range of topics. By contrast, users of a fitness app may vary considerably in their values, interests, skillset and background. Given how widely some data tend to travel, and how differently they can be perceived in different parts of society, the community relevant to a particular dataset could be so dispersed and unbounded as to be difficult to identify. Without actors, however, datafication will not be a dynamic practice. When we talk of solutions looking for problems, it often the case that there is a lack of engagement on the part of actors.

(2) the forms of care to which data are subjected, which include specific ways of attributing meaning to data, regulations and laws aimed at preventing data abuse, as well as incentives to value data in particular ways (from the affective, for instance if the data concern a loved one or a cherished project, to the financial, if data are the result of a big investment and/or promise to deliver significant profit). This also includes the care work needed (maintenance, repair, back-ups, etc).

(3) the capacities of those who handle data, whether they are humans (in which case it is a question of different skills, training and experiential baggage) or machines (in which case, we are looking at statistical methods, computational tools and **machine learning algorithm**s, as well as hardware such as storage and dissemination systems, **network**s).

(4) data themselves in their many forms (ranging from numbers to images, text, symbols), whose significance and value depend both on their physical characteristics and on the care and meaning bestowed upon them by those who use them.

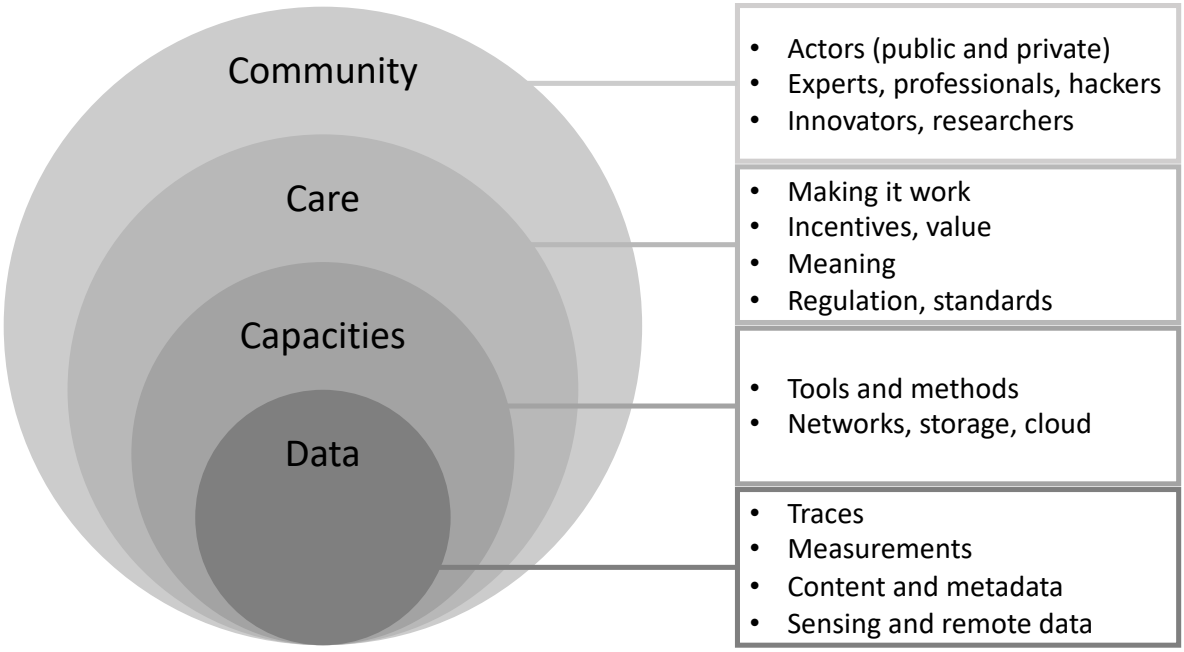



*Figure 1.2 A model of the spheres of datafication, adapted from the ecosystem of Big Data (Letouzé, 2015) through the addition of the care sphere and explanation of the meaning of each sphere.*

Datafication links community, care and capacities to data. This link highlights how data, big or small, never stand on their own. Data only matter if someone cares about them and takes care of them. Care includes all the ways in which we value, regulate, curate, and give meaning to data. As we will illustrate in Section 2, without users who are interested, who have relevant expertise and who engage meaningfully and creatively, there can be no data in the first place. Because care is so complex, it requires an entire community of data workers.

To illustrate how data is not enough and how datafication requires these four elements, think of recreational running and how that sport has changed in the past decade. Many runners now use a fitness watch that measures and tracks various aspects of their performance such as distance, speed, heart rate, and geographical path. These are the data, the measurements whose value relies on a broader landscape of capacities – such as wifi networks, servers, applications for displaying the data to the runner, digital platforms where the data is shared. In turn, these capacities require care. The manufacturers of the watch need to coordinate their efforts with platform providers, so that their services are smoothly integrated. The runner needs to charge their digital tools appropriately, maintain their accounts on the platforms, and check the quality and

reliability of data against their own experience. They may also increasingly care about what the data can and cannot say about their running, and train in ways that accommodate the parameters used by the fitness watch (for instance, by training according to heart rate zones rather than perceived effort). In this, they may be assisted by a community, made up of a coach who uses the data to evaluate the quality of their training, engage with other runners who socialize on the platform because they share data via the same app, and even be targeted by marketers who may focus on particular types of runners based on their data profiles.

It is useful to think about how these components of datafication intersect in the case of other social practices and experiences. To focus only on data and their features is to miss a huge part of the way data comes to matter. Weight loss, child rearing, wayfinding, driving, political debate, medical treatment, opinion formation, entertainment, travel: all have been reshaped by datafication. The stronger the ties between the four components of datafication, the more thorough the process of datafication.

Datafication shapes everyday lives in a variety of ways. It shapes self-presentation to others, as evidenced by the attention paid in managing our social media profiles. For example, someone might post information in particular ways that are likely to yield the kind of profile and garner the kind of attention she or he considers desirable. To do this means taking into account what we know about how algorithms and platforms work or how other users react. Datafication and the increased entwinement of various digital settings also means that different audiences might merge (Beaulieu and Estalella, 2012; Pitcan, Marwick, and Boyd, 2018), and the extent to which profiles and presentation can be managed is also limited. Datafication can also create new vulnerabilities; when undocumented immigrants become more visible because they use digital platforms to find work, they may be more easily exposed to surveillance and detection by immigration authorities (Ticona, 2016).

Indeed, it is important to keep in mind that not all practices undergo datafication in the same way, nor to the same extent. As the examples discussed in this chapter show, datafication ranges from the personal (the datafied runner) to the global (IoT agriculture). Datafication becomes well established, even entrenched to specific ways of life, only when it is strongly established across data, capacities, care and communities. Later on, in Chapter 3, we will also consider how the qualities associated with data tend to make us value **metrics** and indicators in many areas of knowledge production and policy-making.

## 1.3 Data, ethics and knowledge production

When you use social media, you trust the company that maintains the specific app you are using with all sorts of personal information. You may be sharing contact details of friends and family, your favourite places to hang out, or the events you attend. Is it ok to use those data to inform traffic control, so that you can more easily go to those places? And what about using data about your friends to devise whether you are an outgoing person or not, or work out what political party you are likely to vote for? Besides this kind of data that we explicitly share, much more data is gathered about us, from our patterns of logging on to a platform, to the frequency of our likes or the length of our replies. Using your data to acquire insights about your behaviour and preferences can

lead to what seem to be unambiguously good social outcomes – such as better traffic flow, easy access to services and increased safety. And yet, even in that case, the release of location data from smartphone users can have unpleasant, unexpected side effects. For instance, it can highlight places and times in which vulnerable people find themselves alone and isolated, thus facilitating stalking and attacks. Other outcomes, such as more accurate polling before an election, are more ambiguous in their social effects. As exemplified by the scandal surrounding how Cambridge Analytica used Facebook data to launch aggressive political advertising, better predictions of voting behaviour can help politicians to tailor their policies. It can also interfere with public discourse in ways that may be perceived as dishonest and ill-intentioned. These are cases of concern to **data ethics**, which informs the evaluation of what constitutes right and wrong actions in relation to data handling. Data ethics is complex and ever-present. It is part of all design and selection decisions and practices around data, and it typically involves choosing between different options in the absence of obvious solutions or even without knowing what the results of such choices will be. In fact, data workers have to make choices even in situations where all options are problematic, and/or there is no clarity over what constitutes the "right" choice. Crucially, data ethics is not just a conversation happening on the sidelines of data work, but rather underpins many actions taken in the course of such work (even when the action is to do nothing) – a fundamental characteristic to which we return in Chapter 9.

A key concern in data ethics is the open-endedness of data use. It is not always possible to say ahead of time exactly how data will be used and with what effects. Technologies for the production, dissemination and analysis of data keep evolving at great speed. This is partly due to the scale of investments by governments and corporations and to the lack of regulation of these activities. These technologies are often used in unpredictable ways. The internet of things is increasingly dominating human existence, and data-driven systems are entering spheres as diverse as policing, immigration, healthcare and energy consumption. It is ever more difficult to know which data are held, by whom and where. All these elements add to the difficulty of knowing how data could be used in the future, and what the social, economic and political implications could be. How we deal with this complexity is an ethical question. While it is, by definition, impossible to deal with all implications of data present and future, it is possible to reflect on the best ways to address dilemmas, and on which values should prevail when faced with these. This book aims to show the complexity of such processes and to help deal with effects in more responsible ways.

For example, in 2008, EU policy decreed that every household in Europe should be equipped with a smart meter that can transmit data about energy use and production. Armies of engineers subsequently installed new devices in millions of homes. And yet this massive change was not accompanied by a discussion of the different roles that households that now have more data about their energy use could play in new energy regimes. Nor has there been a public debate over the trade-offs between the related loss of privacy due to the smart meter, and a better managed power transmission grid. Fostering such debates does not require full knowledge of the possible consequences of datafication, nor does it involve reaching consensus on ways forward. What it does require is providing a space to: consider possible implications from a variety of viewpoints; and ensure that potential concerns are explicitly acknowledged and taken into account in the development of technology-focused social interventions.

A related concern is the opacity of the technical work surrounding data processing, analysis and interpretation. As we will see in this book, many decisions with significant ethical implications are taken in the course of developing systems for data analysis. This can be through simple decisions, such as whether to accept a given data format or source, or how to label particular datasets. The extent to which data analysis depends on statistical, mathematical and computational expertise make such work daunting for anybody who is not trained in these fields. In turn, this creates the fear that technicians may implement ethically dubious decisions without any oversight or consultation. And yet, the idea that a statistician or a computer scientist would be able to fully comprehend data systems is also misleading: many different forms of specialised work are required to process and analyse data. This means that even very technically savvy data workers may not be able to understand the data system as a whole, and much less to predict its implications.

Does this mean that we cannot do anything about ethical issues? We think that being aware of this complexity actually encourages us to have regular and wide-raging consultations on the possible implications of technical changes to data systems. Consider for instance the growing tendency to encourage users to make use of a Facebook, Twitter or Google account to access other digital platforms. Think of using a Facebook account to sign onto a library account. While accepting this service may seem harmless and even convenient, it has big implications for data flows. By using your Facebook profile to log onto a different service, you are linking two sets of data. Different databases operated by different platforms become linked. This enhances the value of data held by the corporate owners of the platforms – in this case, social networks and book and media use can be correlated. In addition, reusing profiles across different platforms makes you more vulnerable to identity theft or other privacy breaches. This practice of using profiles to log onto other platforms also makes it more difficult *not* to use particular platforms. If the expectation is that one will use an existing profile, accessing the platform in an alternative way increases the cost (in time, in attention, in amount of work required) of *not* participating in a given platform.

To realise how these seemingly mundane practices have an ethical dimension is not based on in-depth technical knowledge. Seeing the ethical dimension requires a basic understanding of the system, coupled with the opportunity and time to think about its implications. In a world where technological development (or innovation) is considered both a good in itself and a competitive advantage, such opportunities and time are seldom created. This increases the perception of technological choices as impenetrable. So how can we ensure that there is attention to ethical aspects of data? The field of data ethics focuses on this question. The most important goal of data ethics is to promote responsible and sustainable data work in ways that may contribute to human flourishing (Floridi, 2014; Floridi et al., 2018). And yet in the context of datafication, where intricate flows of data and multiple sets of algorithms are the rule rather than the exception, it can be very difficult to establish who is responsible for ensuring that data is used ethically. This can mean not knowing who should take the blame when things go wrong, or who should be responsible for fixing errors or solving problems when they arise. Incorrect data can be difficult to remove or correct once it has moved beyond the context where it originated. Imagine an error in a school record, that is shared with an employer, and then travels to a person's file at an insurance company. It can be very

difficult to trace where the error occurred and nearly impossible to find out where the incorrect data ended up. This is one of the reasons why governments are currently working very hard to articulate new laws that clarify responsibilities in relation to data. In this example, personal data is at the forefront, but as we will see in Chapter 10, attributing responsibility is even more difficult in cases where data about groups get built into data infrastructures or where machine learning tools are trained on historical data sets.

In nearly every aspect of data work, ethical decisions will be required, and there will be no ready-made answer as to which course of action is the best. This book will treat data ethics as an integral part of data management, and point you towards tools, principles and guidelines that can help to identify, address and resolve ethical concerns as they arise in data handling and use. This will be the main focus of Section III. Furthermore, across the chapters in this book, we create opportunities to stop and think about the small moments of design and decision and about the effects of large-scale implementation. By discussing ethical consideration across many different situations in relation to data, we hope to enhance awareness that **ethics** is not a one-off or separate kind of concern. Data ethics contributes substantively to the effectiveness and positive impact of data solutions.

## 1.4 Conclusion: The Impact of Datafication

What is happening to knowledge in contemporary society is not simply that we have more or 'bigger' data: the whole system of knowledge production is changing. Datafication involves not only data, but also community, care and capacities – all of which relies on material conditions, values, preferences and norms for acceptable behaviour. When knowledge practices integrate data, they also align across all components. For data to matter and become evidence for a specific claim or course of action, all four have to be in place.

Data-intensive practices connect to contemporary ideas about what is good knowledge. Put more concretely, we are living in a world where data is part of what we feel we need to know in order to parent, police, govern, be healthy or put together a good soccer team. These changes in knowledge are related to issues of trust in knowledge and truth. As data becomes more central to how we produce knowledge, data also become more central to how we evaluate knowledge. This holds for everyday knowledge queries (we google to find out), for researching life-changing decisions about health or real estate (we look for data to inform our decisions; we compare our data to averages/profiles/percentiles), and for participation in public life (we discriminate between real and fake news). In all these activities, we use our data skills and insights – what we know about how data is generated, what is good data, how to analyse data responsibly, and how data might be tampered with. All these ways of evaluating data are far from self-evident, if you look back in time just a couple of decades. In the chapters that follow, we will explore different aspects of data work and provide a set of concepts and tools to further think about, analyse and evaluate data.

## Additional Reading

Ebeling, M., 2016. *Healthcare and Big Data: Digital Spectres and Phantom Objects*. Palgrave.

Floridi, L., 2014. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Kitchin, R., 2014. *The data revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Van Dijck, J., Poell, T. and De Waal, M., 2018. *The platform society: Public values in a connective world*. Oxford University Press.

# Section 2: Data Creation

## Summary

This section focuses on the creation of data. It discusses our expectations of what data is and what data can do. We start by discussing the promises of big data and the historical development of data in Chapter 2. Chapter 3 reviews the characteristics of data, the importance of context of creation and of data journeys in shaping the meaning and characteristics of data. By discussing concrete examples of data creation and data journeys, we show how taking these elements into account puts us in a better position to evaluate the suitability and reliability of data. In Chapter 4, we turn to different ways of conceptualizing data, and contrast the representational and relational view of data. A fixed, representational view of data positions data as a foundation on which to build knowledge and emphasizes the need to remove bias and noise as the most important data work. A relational view positions data as changeable and contextual, and emphasizes that many kinds of data work are needed across all steps of knowledge production. Finally, we consider the changing role of data, as it becomes more central to how we evaluate knowledge and to the broader knowledge production cycle.

This section will help you to

1. understand the historical roots of data science and big data
2. understand what data are and how they relate to knowledge production
3. critically evaluate hyperbolic claims on the power of Big Data
4. identify the various technical, epistemological, social, legal, institutional and economic dimensions of data journeys and of how they are entwined
5. evaluate data according to their provenance, their merits and disadvantages, and critically assess their quality and limitations

# Data Story 1: Big Data on Consumer Habits

The introduction of credit card payments and loyalty cards for customers of supermarkets has created a vast amount of digital information on customers' preferences and spending habits. These technologies are typically advertised as means to facilitate payments and obtain discounts, so customers do not necessarily think of these technologies as tools to produce data. Indeed, the primary function of credit cards has long been to facilitate payments; and the function of loyalty cards is, at least on the surface and as their name suggest, to ensure loyalty of costumers to a store or brand, usually by providing special offers. At the same time, such technologies have created vast amounts of data about customers, which provide a wealth of insights to supermarket owners and retail companies. The data can be used to identify purchasing patterns that can indicate which products are most popular, at what time of the day or season, and among which types of customers. These analyses can support marketing strategies as well as helping to manage the supply, distribution and shelving of items.

- ***What kind of data are these?*** *What makes the digital traces left by supermarket transactions into "data"? What characteristics do such data have? What are these data about?*

- ***Can you think of any disadvantages of using these data to inform product supply?*** *Is the use of these data always beneficial? Who benefits?*

- ***Does the scale of data collection make a difference, and how?*** *For instance, does it matter whether we are considering data collected by a large supermarket chain with hundreds of sites around several countries, or whether we are considering data from three local shops whose customers may know each other?*

The combination of data from supermarket transactions with other types of data can provide even better insights. Such combinations become possible when customers sign up to a given loyalty scheme, and typically provide their address, date of birth, and telephone number. Further combination may be the result of doing searches of customers' names and finding out their medical history, personal preferences and lifestyle habits from social media. The combination of these data sources gives rise to an enormous data pool, often called 'Big Data'. From such a data pool, data analysts can extract predictions about what a customer is likely to buy in the future, how their social lives will evolve, or where they are likely to go on holiday.

- ***How do Big customer Data work?*** *When does a given dataset deserve to be called "Big Data"? Why are these data valuable, in which ways and for whom? When different types of data are combined, do they increase the accuracy and reliability of the knowledge being produced?*

An improved understanding of the customer base enables many companies to tailor their services closely to the desires and needs of their clients. It also increases the opportunities to manipulate customers' behaviours via targeted advertising or special offers geared to facilitate addictions to specific types of products.

- ***Does it matter whether customers are aware of who is using their data, and how?*** *Why? Are data derived from consumer services or social media always reliable? Can you imagine cases where the knowledge extracted from such data is not trustworthy? To what extent can repeated suggestions and nudges from companies shape our behaviour? Is it possible to ignore this targeting, and if so, to what extent?*

## Data Story 2: Remote Sensing for Conservation Research

Remote sensing technologies such as drones are widely viewed as the new frontier of data collection, especially when it comes to environmental and biological research and monitoring. Among many other things, they help to map the spread of crop diseases around the world, the extent of deforestation in the Amazon and the degree of damage to coral reefs. In this second data story, we look at the use of UAV (unmanned aerial

vehicles) in conservation projects to detect wildlife and monitor the behaviours of protected species. In particular, we consider how UAVs have been used to produce data documenting the location of chimpanzee nests in Tanzania.

As with other species of great apes, chimpanzees' survival is heavily threatened by environmental changes, deforestation, disease and poaching. To help protect this species, conservationists agree on the need for accurate data on their distribution and density, which need to be gathered at regular intervals. These data can help to identify areas where habitat encroachment, poaching or disease are leading to population loss, thus paving the way for targeted interventions. Given the sheer scale of chimpanzee distribution across western Tanzania (over 20,000 $km^2$), there is urgent need for cost-effective methods of data collection and analysis that can reliably and frequently track chimpanzee numbers and movements across broad spatial scales.

Recently, drones have been used to identify nests from the air. Drones fitted with cameras gather images by flying over large areas. To identify nests, the photographs and videos gathered by the cameras can be analysed automatically using machine learning. This is more effective than using aircraft or satellite data, since those tools do not have enough resolution to be able to detect smaller animals and their traces. Using drones seems to require much less effort and resources than surveys conducted by technicians on the ground.

A key complication is that it is actually difficult to observe chimpanzees themselves through these data. Indirect **evidence** of their presence (for example, dung, calls, or nests) is typically the main parameter used to estimate population size. Much indirect evidence is gathered on the ground by specialised technicians, often using ground surveys. For instance, a "line transect survey" is done by counting all individuals in one specific strip of territory and using that sample to estimate population size over the whole area. This way of acquiring evidence is arguably much more labour - and time - intensive than the use of drones.

- ***What kind of data are acquired via observational techniques like ground surveys?*** *Objects used as evidence are not always numbers resulting from measurements: they may include pictures of footprints, samples of droppings, and handwritten notes about sounds heard in the forest. Are these the same kind of data as the images produced by drones? Are the data obtained from drones comparable to data acquired manually through observations from the ground, and how?*

- ***How can we evaluate the quality and reliability of data in this case?*** *Is quality evaluation possible in the absence of multiple sources of data? For example, would we be able to evaluate the quality and reliability of drone image detection without triangulating with data acquired on the ground? Are drone data "good enough" to warrant stopping ground surveys altogether, thus generating enormous savings for conservation efforts? Or should the drone census be combined with a survey on the ground to confirm the results (which may increase the accuracy and reliability of results, but will also increase the costs)? How could this data be complemented by tracker data, to understand habitat use and range?*

Let us now consider whether the use of drones actually does cost less effort and resources than the use of ground surveys. Drones can be operated by a small team that needs to travel to the various national parks being investigated. To use the drones, at least one member of the team must hold a drone pilot license (which is expensive to obtain) and be knowledgeable about demarcated flying areas, flight height, and privacy laws. Within each location, the team members need to adjust and calibrate the cameras, set them up on the drones, and wait for the right weather conditions to fly the drones according to a pre-determined grid pattern. Depending on weather changes and visibility conditions on the day, the data may be more or less consistent, and the instruments will need to be regularly recalibrated and cleaned up by the data scientists who are in charge of integrating results collected across different sites. Different teams may also use different parameters to calibrate the cameras or set up camp, which again will determine some differences across datasets that will need to be evaluated manually.

- ***Which data are most expensive to produce and analyze?*** *When taking the data work required to make sense of drone data into consideration, does it still make sense to view them as less resource-intensive than survey data? In which ways do these differences matter to the broader effort of designing, implementing and financing a study?*

Now, consider that when conservation officers do a ground survey, they walk through areas that they very familiar with. They know where to look and detect all sorts of signals from the environment, some of which may come to acquire significance later -- for example, in case photographs of local trees reveal symptoms of a newly emerging plant disease. They also talk to visitors and inhabitants of the park. Conservation officers can thus pick up signs of poaching and raise alerts, spot new species moving into the area, or identify changes in behaviours of local animals. By contrast, the drone team, while technically savvy, may not have the same degree of familiarity with the area since they are only occasionally present and tend to cover much broader territory in much less time. The imaging data that they collect can reveal all sorts of unexpected things, but it comes in a highly standardised format and there is little chance to deepen observations or investigate unexpected findings on the spot. At the same time, the use of drones makes data more spatially explicit and has a higher spatial resolution (geo-location of less than 1 m) than human observations (about 15 m accuracy). Aerial surveys can also be done more often, since line transect surveys on the ground are very time consuming.

- ***How do choices about data acquisition affect the type of research – and knowledge - subsequently produced?*** *Given these different ways of working, how do you expect the resulting insights about chimpanzee populations to differ? Are these differences a problem? How might the results of the two teams be evaluated by other conservationists, researchers or policy-makers? How much does precision matter? What are the advantages of a cheaper and faster identification of nests?*

- ***How do choices about data acquisition affect what is valued as relevant expertise in a project?*** *Does reliance on drone data favour the deployment of technical personnel (no matter where they came from) over people*

*with local knowledge and familiarity with the areas in question? Who is best qualified to be doing the identification of animals and estimation of population size and state? Does it matter if knowledge is not produced by people who have ties to the areas, and how?*

The image detection system can identify nests, though researchers feel that this could be improved even more, with better image resolution. When nests are in the middle of the tree crowns, they are more difficult to identify using drones and many nests are missed. To increase detection, researchers want to improve image resolution by using lower altitude drones with more sophisticated multispectral cameras. This use of the drones is likely to have a higher environmental impact, as drones would be visible to animals and their passage may affect the local ecosystem. Also, were these cameras to be flown over inhabited areas, they would capture minute details of the everyday life of humans living there.

- ***What are the broader implications of choices made when creating data?*** *What could be the effects of improving detection in this way? What could be the effect on the data gathered? What could be the effect on the chimpanzees? Who should decide whether the increased resolution and nest detection are worth possible negative effects? Could this be taken into account in the design of the tools?*

Data Story based on Bonnin, Noémie, Alexander C. Van Andel, Jeffrey T. Kerby, Alex K. Piel, Lilian Pintea, and Serge A. Wich. 2018. 'Assessment of Chimpanzee Nest Detectability in Drone-Acquired Images'. *Drones* 2 (2): 17. https://doi.org/10.3390/drones2020017.

# Chapter 2 Big Data in context

## Summary

In this chapter, we focus on how and why Big Data has become so prominent. We review the high expectations associated with Big Data and examine 'Big Data mythology', the often overly optimistic views of what data can and will be able to do. We show how Big Data is part of a longer history, in spite of being hailed as a radical innovation and we highlight the limitations of Big Data. The long history of data across different social circumstances and periods makes clear that what we think of as data, and what we think it is good for, has changed radically over time.

## 2.1 Introduction: The rise of Big Data

The datafication of society is characterised by three main features. First, we see that the creation of data is becoming important and increasingly valued. Second, by using, combining and visualising data in everyday life, data become even more central. And third, we take more and more decisions about current and future actions based on data. If we formulate these three features more conceptually, we can say that (1) data are viewed as valuable commodities that have the power to transform society, and they are no longer viewed as by-products of administrative and research processes; (2) efforts to mobilise, integrate and visualise data are viewed as central contributions to social life, since the more data are pooled together, the higher the chance that they will acquire new significance and meaning; and (3) the consultation of data resources, typically mediated by complex infrastructures and databases, is regarded as the first step in any process of inquiry and plays a key heuristic role in determining future directions for social action.

Together, these features show how the role of data can become increasingly important. The emphasis on the key role of data as starting point for inquiry is rooted on the wish to capitalize on the "data deluge" generated by new technologies through the datafication of human activities. The resulting "**Big Data**" are a resource for research, with ever more sophisticated computational tools being developed to extract knowledge from such data. The data story at the beginning of Section II discussed the case of consumer data garnered from credit card payments and loyalty schemes. Such data can improve current understandings of the nutritional status and needs of a particular population, particularly when combined with data coming from public health and social services, such as blood test results and hospital intakes linked to obesity. Other examples are the use of various different types of data acquired from cancer patients, including genomic sequences, physiological measurements and individual responses to treatment, to improve diagnosis and treatment. Or think of the integration of data on traffic flow, environmental and geographical conditions, and human behaviour to produce safety measures for driverless vehicles. By integrating this data, better approaches can be developed that make it possible to promptly analyse a situation and generate an appropriate response – a child suddenly darting into the street on a very cold day and the driverless car swerving enough to avoid the child while also minimizing the risk of skidding on ice and damaging other vehicles. In each of these cases, the availability of diverse data and related analytic tools is creating novel

opportunities for research and for the development of new forms of inquiry, which are widely perceived as having a transformative effect on society as a whole. In this chapter we will present some of the key characteristics attributed to Big Data, and then we will critique some of those ideas by pointing to the practical obstacles, conceptual problems and social implications involved in the dissemination, aggregation and use of large datasets.

There are multiple ways to define Big Data (Kitchin and McArdle, 2016). Perhaps the most straightforward characterisation is as *large* datasets that are produced in a *digital* form and can be analysed through *computational* tools. The two features most commonly associated with Big Data are volume and velocity. *Volume* refers to the size of the files used to archive and spread data. *Velocity* refers to the pressing speed with which data is generated and processed. As an increasing amount of data is produced and processed, it seems that we need automated analysis.

Precisely those two features, volume and velocity, are however also the most disputed features of big data. What may be perceived as "large volume" or "high velocity" depends on the context. What a large data set is in some fields is perfectly normal in others. Furthermore, technologies used to generate, store, disseminate and visualize the data change rapidly. This is exemplified by the high-throughput production, storage and dissemination of genomic sequencing and gene expression data, where both data volume and velocity have dramatically increased within the last two decades. Similarly, current understandings of big data as "anything that cannot be easily captured in an Excel spreadsheet" are bound to shift rapidly as new analytic software becomes established, and the very idea of using spreadsheets to handle data becomes a thing of the past. Moreover, there are many other ways to qualify data. A focus on data size and speed do not take account of the diversity of data types used. This may include data that are not generated in digital formats or whose format is not computationally tractable. This shows the importance of **data provenance**, that is, the conditions under which data were generated and disseminated. By emphasising velocity and volume, we risk ignoring other important elements that shape the interpretation of data, such as the circumstances of data use, including specific queries, values, skills and research situations.

An alternative is to define Big Data not by reference to their physical attributes, but rather by virtue of what can and cannot be done with them. In this view, Big Data is a heterogeneous ensemble of data collected from a variety of different sources, typically (but not always) in digital formats suitable for algorithmic processing, in order to generate new knowledge. For example (Boyd and Crawford, 2012) identify Big Data with "the capacity to search, aggregate and cross-reference large datasets", while (O'Malley and Soyer, 2012) focus on the ability to interrogate and interrelate diverse types of data, with the aim to be able to consult them as a single body of evidence. The examples of transformative "Big Data research" given above are all easily fitted into this view: it is not the mere fact that lots of data are available that makes a different in those cases, but rather the fact that lots of data can be mobilised from a wide variety of sources (social media, environmental surveys, weather measurements, consumer behaviour).

This account makes sense of other characteristic "v-words" that have been associated with Big Data. These other aspects emphasise functional rather than physical characteristics, and include:

- *Variety* in the formats and purposes of data. Data include objects as different as samples of animal tissue, free-text observations, humidity measurements, GPS coordinates, and the results of blood tests;
- *Veracity,* understood as the extent to which the quality and reliability of big data can be guaranteed. Data with high volume, velocity and variety are at significant risk of containing inaccuracies, errors and unaccounted-for bias. In the absence of appropriate validation and quality checks, this could result in a misleading or outright incorrect evidence base for knowledge claims (Floridi and Illari, 2014; Cai and Zhu, 2015; Leonelli, 2017);
- *Validity*, which indicates the selection of appropriate data with respect to the intended use. The choice of a specific dataset as evidence base requires adequate and explicit justification, including recourse to relevant background knowledge to ground the identification of what counts as data in that context (Bogen, 2010; Mayernik, 2019);
- *Volatility* refers to the extent to which data remains available, accessible and re-interpretable despite changes in archival technologies. This is significant given the tendency of formats and tools used to generate and analyze data to become obsolete, and the efforts required to update data infrastructures to guarantee data access in the long term (Sterner and Franz, 2017; Edwards, 2010; Lagoze, 2014; Borgman, 2015);
- *Valu*e points to how – and to which extent - data come to matter to different sections of society. This is not just in a financial or economic sense. Rather, the attribution of "value" to data encompasses any way in which data could be perceived as significant, whether this is scientific, financial, ethical, reputational or even affective forms of value (Leonelli, 2016a; D'Ignazio and Klein, 2020). Value depend as much on the intended use of the data as on historical, social and geographical circumstances. It is important to note that it is not only individuals or groups who attribute value to data. Institutions, such as the companies and research organisations involved in governing and funding data-intensive science, also have ways of valuing data, which may not always overlap with the priorities of data workers (Tempini, 2017).

This list of features, though not exhaustive, highlights how Big Data is not simply 'a lot of data'. The epistemic power of Big Data lies in their capacity to bridge between different research communities, methodological approaches and theoretical frameworks that are difficult to link due to conceptual fragmentation, social barriers and technical difficulties (Leonelli, 2019). And indeed, appeals to Big Data often emerge from situations of inquiry that are at once technically, conceptually and socially challenging, and where existing methods and resources have proved insufficient or inadequate. Examples range from the attempt to understand biodiversity via integration of highly heterogeneous observations from remote location around the globe (Sterner and Franz, 2017) to the mass aggregation of social data to track consumer demand for specific products or to produce national measures of economic growth.

## 2.2 The Big Data mythology: Data transform society

The emergence of Big Data affects many sectors of society, including scientific research. Promises to enable new and more efficient ways to plan, conduct, institutionalise, disseminate and assess research form a set of expectations that have been embraced by many government agencies, companies and research organisations in the first two decades of the 21st century. They continue to inform the development of analytics and technologies underpinning the use of Big Data. We label these expectations the **Big Data mythology** to emphasise its mythical status and stress the difference between such utopian promise and what Big Data (and related analytics) can actually deliver for society.

The ability to link and cross-reference datasets coming from different sources is expected to increase the accuracy and predictive power of scientific findings, and help researchers – whether they work in universities, industry or policy institutions - to identify future directions of inquiry. The availability of data provides an incentive to build automated procedures and tools to store, organise and analyse the data, in the name of improving the reliability and transparency of knowledge creation. It is widely believed that Big Data are ushering in a whole new way of doing research, which is heavily grounded in data analysis and less dependent on pre-existing theories. This belief is reflected in the renewed attention to data strategies as key component of management for industry. It is also tangible in novel sources of funding and publication venues (such as "data journals") within academia.

Big data are often presented as *comprehensive*. This is the claim that the accumulation of large datasets enables researchers to ground their analysis on several different aspects of the same phenomenon, documented by different people at different times. According to Mayer-Schönberger and Cukier (2013), data can become so big as to encompass *all* the available data on a phenomenon of interest. As a consequence, Big Data can provide an all-encompassing perspective on the characteristics of that phenomenon, without needing to focus on specific details. This is a big promise, and perhaps an understandable one, given the speed and extent of datafication in many areas of life. Yet, it is an extreme promise to hold up to say that 'we have all the data'; that n=everyone, or that all of reality is captured, as invoked by Twitter's slogan, "it's what's happening".

Big Data are also argued to push researchers to embrace the complex and multifaceted nature of the real world, rather than pursuing exactitude and accuracy in measurement obtained under controlled conditions. Indeed, it is impossible to assemble Big Data in ways that are guaranteed to be accurate and homogeneous. Rather, the Big Data mythology encourages analysts to resign themselves to the fact that "Big Data is messy, varies in quality, and is distributed across countless servers around the world" and welcome the advantages of this lack of exactitude: "With Big Data, we'll often be satisfied with a sense of general direction rather than knowing a phenomenon down to the inch, the penny, the atom" (Mayer-Schönberger and Cuckier 2013, 13).

This idea of messiness relates closely to the third key innovation brought about by Big Data, which Mayer-Schönberger and Cukier call the 'triumph of *correlations*'. Correlations can be defined as the statistical relationship between two data values. They are notoriously useful as heuristic devices within the sciences and beyond. Spotting that when one of the data values changes, the other is likely to change too, is the starting point for many discoveries. It is also used to analyse economic activity and many attempts to understand human behaviour: for any big change in the market, the political

situation or the environment (such as an economic crisis, a change of government or an earthquake) changes in citizens' work and spending patterns can be scrutinized to see whether there may be a link. Market research as a whole may be viewed as an exercise in the identification of correlations between a user profile and their preferences for specific products. However, researchers have typically mistrusted correlations as a source of reliable knowledge in and of themselves. This is chiefly because they may be spurious and due to chance – in other words, they may result from serendipity rather than specific mechanisms, or they may be due to factors other than the variables under consideration. For instance, a Netflix user may be using the opening music of a documentary series every night to get his son to sleep, but may never have watched the series and may in fact hate documentaries. In such a case, it would be wrong to interpret Netflix data on his viewing history as being correlated to his viewing preferences.

According to the Big Data mythology, Big Data can override those worries about spurious correlations. In a Big Data world, it is argued, it simply does not matter whether any single correlation is reliable: what matters is the correlations spotted on a very large dataset can help to predict future behaviour with reasonable accuracy. In the example of the Netflix viewer, asking whether he actually liked documentaries becomes irrelevant: what matters is that data analytics can reliably predict that he will be streaming the intro to the documentary again tomorrow. On a much broader scale, Mayer-Schönberger and Cukier give the example of Amazon.com, whose astonishing expansion over the last few years is at least partly due to their clever use of statistical correlations among the myriad of data provided by their consumer base in order to spot users' preferences and successfully suggest new items for consumption (Mayer-Schönberger and Cukier, 2013). In cases such as this, correlations provide powerful predictive knowledge that was not available before, and that can inform society without appearing to be complemented by a causal understanding of *why* a specific effect is predicted. Causal understanding is viewed as simply irrelevant to the useful knowledge yielded from big data. Hence, Big Data encourage a growing respect for correlation and prediction, which comes to be appreciated as more informative and plausible form of knowledge than the more definite, but also more elusive, causal explanation. In Mayer-Schönberger and Cukier's words: "the correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening. And in many situations, this is good enough" (Mayer-Schönberger and Cukier, 2013).

This Big Data mythology is associated with a specific reading of the significance of technology in shaping social life. It is argued that in the past we lived in an analogue world, where computation was extremely limited and mechanical in nature. Now the increased use of information and communication technologies has given rise to a digitisation of experience. In this world, experience that is computable and machine-readable is in turn valued for the possibility of detecting patterns, developing profiles and further predicting behaviour. These practices shape a particular version of human experience – as the sum of our digital traces– as the basis for knowledge. The same holds for expertise. We have seen a shift from expertise as something embodied in a human expert, and developed over time through the active combination and application of knowledge and practice (Daston and Galison, 2007). Increasingly since the 1970s, expertise is embedded in 'expert systems'. These are often considered to be the first forms of artificial intelligence and were developed in the medical field to assist in diagnosis but also in areas as diverse as speech recognition and crisis management.

These systems are usually based on formalising the reasoning of human experts as a set of rules. They promise to fully automate processes and eliminate the need for human intervention. In parallel, a culture of metrics and auditing has spread through institutions, industry and governments around the world, in which data is the main tool used to evaluate outcomes and processes. As a result of these changes, the Big Data mythology highlights how automated use of digital data has become central to *how* we know – with medicine, business, policy, education, environmental concerns all focused on obtaining and analysing data, often for predictive purposes. The availability of digital data is viewed as the best kind of proof or as the best basis for taking action (evidence-based policy). We return to the role of data in decision-making in chapter 10.

## 2.3 A historical perspective: Society transforms data

Some of the claims associated to the Big Data mythology described above look perplexing when evaluated from a historical viewpoint. For one thing, reliance on large datasets is not a novel development, just as data are not necessarily digital objects but include texts (observations about the characteristics of specific territories), drawings (reproductions of the morphology of a newly discovered species of plant) and even specimens (fossils). Data have long been the foundation of empirical inquiry, with long-standing efforts to collect and organise large volumes of data in domains such as astronomy, meteorology and natural history (Daston, 2017; Anorova, et al., 2010; Porter and De Chadarevian, 2018). Similarly, biomedical research – and particularly subfields such as epidemiology, pharmacology and public health – has an extensive tradition of tackling data of high volume, velocity, variety and volatility, and whose validity, veracity and value are regularly negotiated and contested by patients, governments, funders, pharmaceutical companies, insurances and public institutions (Bauer, 2008). The world of research is no stranger to the accumulation of data, and to the quest for ingenious technologies – such as archives, punch cards and statistical techniques - that would facilitate the management and analysis of all that material. In this section, we briefly review key points in the history of research data, to exemplify the ways in which, just as data changed society, social change shaped data.

In the Western world of the 17th and 18th centuries, research data was gathered by visionary individuals, such as gentried naturalists like Charles Darwin and court astronomers like Ticho Brahe. These individuals were backed by wealthy patrons and supported by an extensive network of data collectors, who would roam the globe in search of new biological specimens and gather astronomical and meteorological observations from a variety of different locations. This "age of discovery" was marked by an extractive approach to colonial expansion: rich European countries were focused on identifying and bringing back resources from the colonies that would extend and consolidate their knowledge and power. The large quantities of data thus accumulated were systematised and analysed through models (think about Kepler's laws describing how planets move around the sun, derived from consideration of the astronomical observations collected by Tycho Brahe in the 16th century) or classification systems (such as Linnaeus' taxonomy of different forms of life, which grew out of the study of specimens collected by explorers in the 18th century and underpins how we distinguish between species to this day). These approaches to ordering data were valuable because

they made data more usable and combined principles of organisation with data, so that these systems had both simplicity and explanatory power.

In the 19th century, data shifted from the work of brilliant individuals to a more institutional approach, with national agencies such natural history museums, boards of trade, the census and weather services emerging as a constitutive part of the administrative apparatus of national governments. This means that data were increasingly recognized as social commodities with scientific as well as financial and political value. Data became something to be invested in, regulated and managed – and was clearly marked as a tool for governance and trade. Again, this accompanied a shift of colonial rule: this time it was from extraction to control, with dominating powers increasingly interested in how to manage large populations in the wake of increasingly intense revolts at the turn of the century, and epidemics – and related food shortages – proving increasingly disruptive to urban life and global trade.

Through collaboration between national information infrastructures, more sharing took place, leading to a new informational globalism (Hewson, 1999; Edwards, 2010). With the rise of nation states and the increasing demands of international trade, these initiatives aimed to measure both nature and society in a more systematic, depersonalised manner, and were fostered by an ever-expanding group of data aficionados including researchers as well as administrators, merchants and politicians.

As data became a growing concern, sophisticated techniques of quantification were developed. Statistics became a separate discipline. As more complex techniques and as more experts became available, more complex types of data gathering were developed, such as the census (Oertzen, 2018). International entities such as the League of Nations and the International Monetary Fund had clear aspirations to globalise data collection and analysis for a variety of purposes and across all scientific domains: from drug testing, with the creation of the Permanent Commission on Biological Standardisation to monitor chemical tests and biological assays from 1924, to economic assessments through comprehensive data collection on employment, unemployment, wages, migration by the Economic and Financial Section of the newly instituted International Statistical Commission. Population-level thinking gripped the life sciences through the widespread adoption of the Mendelian theory of inheritance, which was fruitfully combined with attention to new types of data – and specimen – collections focused on genetic mutants of the same model species.

Statistics became the main source of information for emerging insurance practices and public health monitoring systems (Porter, 1995; Desrosières, 2010). While statistics now seems to us an obvious way to think about health, poverty or employment, concepts like rates of unemployment, epidemics or the probability of being victims of a crime are relatively recent developments.

While the two world wars in the 20th century proved severely disruptive to short-term data collection and sharing efforts, the large amount of military investment in intelligence and related information technologies kick-started the post-war drive towards mechanised computing. Investment in information technologies and related infrastructures continued to grow, as did the power of numerical models (such as weather forecast) that enabled number-crunching at a previously unimaginable scale. Research data became well-recognised as political and diplomatic tools. The “one world” movement towards international cooperation eased efforts towards stabling

globalised initiatives of data collection and analysis. Within climate science, the World Meteorological Organisation was founded in 1950 to oversee the international linkage of regional weather services, for instance through the institution of a World Weather Watch and the Global Atmospheric Research Program. In 1957-58, the International Geophysical Year marked both a decisive advance in the commitment of geophysical and oceanographic sciences towards global data exchange, and a diplomatic achievement in fostering good international relations through research communication. This meant, once again, a focus on global infrastructures and related institutions.

From the 1970s onwards digital infrastructures for data sharing were being built in virtually every scientific field (Edwards, 2010; Strasser 2019). There were also efforts to increase global monitoring, which means the tracking of data across many different contexts. The greater availability of computing resources, the growth of expertise and the possibility of sharing data digitally (increasingly over networks) were important factors for this movement towards global data. During this period, the United Nations consolidated its global environmental monitoring system just as the World Health Organisation systematised its efforts to map the spread of infectious diseases. The goal shared across these initiatives was the development of tools, such as numerical models, that could help to manipulate data at a previously unimaginable scale. Recall that this is the age of room-sized computers and that computational power was often a limiting factor. There was also a growing exchange of expertise on statistical and computational approaches to data (UN Division for Statistics). Simulations and future scenarios also increased the use and visibility of global data.

During the 70s, data was increasingly conceptualised as a shareable and re-usable asset, rather than something to be collected to be used only once. Data became an object of exchange and reuse. This approach to data was sparked by the cybernetic movement, with its emphasis on modularity and complexity (Pickering, 2011). It was also accelerated by the rising positioning of science and technology as means towards economic growth, military power and international relations. At the same time, Big Science projects carried out at Los Alamos in the United States and CERN in Geneva became a model for how to do research (Price, 1963). Within these programmes, the production and trade of data were no longer the responsibility of individual researchers, but rather the product of large investment and collective efforts carried out in centralised experimental facilities. Even in fields where such centralisation was unfeasible, such as environmental, biological and climate sciences working with observational rather than experimental data, there was a strong focus on building data-sharing networks so as to feed more information to novel computational tools. For instance, the International Geophysical Year of 1957-58 marked both a decisive advance in the commitment of geophysical and oceanographic sciences towards global data exchange, and a diplomatic achievement in fostering good international relations through research communication (Aronova, Baker and Oreskes, 2010).

The history of data use became ever more tightly intertwined with the history of data infrastructures, and institutions in charge of deciding who should have access to data, what standards and **conventions** should govern what data to collect and how, and how the resulting outputs should be labelled in order to be comparable across time and space (Edwards, 2010; Daston, 2017). This required effective administration and monitoring, a long-term vision of the research domain at hand, and conceptual and technological innovations steeped in specific conceptions of the research objects under

investigation – a repertoire of skills, methods, institutions and tools that took decades to develop and continues to evolve to this day (Ankeny and Leonelli, 2016).

Thus already in the 20th century, and even more in the 21st, data became recognised as objects of public interest, particularly by governments wishing to use data to inform policy and the management of commerce, military assets, diplomatic relations and public health. The rise in the social status of data was accompanied by an increasing recognition that methods and logistics of data access and management play a significant role in channelling analysis and, eventually, interpretation.

Many studies have emphasised how data and datafication coupled to digital networks and computational tools have occasioned a societal shift. One of the early concepts used to describe this shift was the '**information society**' (Bell, 1979). In the information society, information is central to the capitalist system of production, innovation and consumption. The information society is often contrasted to other dominant forms of organization of society, such as industrial activity. A more recent concept is that of the **knowledge society**. In this more utopian view, society generates, processes, shares and makes knowledge that may be used to improve the human condition available to all its members (Castelfranchi, 2007).

These narratives resonate strongly with the Big Data mythology of knowledge emerging from technical developments in data handling. Without taking anything away from the obvious, enormous impact that statistics, computing and related infrastructures have had on knowledge development over the last century, our brief historical overview emphasises even more strongly how society – and more specifically, the social conditions, motivations and governance of data production, exchange and use – has changed, and continues to change, the status, value and uses of data.

## 2.4 Conclusion: Data do not speak for themselves

The historical review in the previous section makes clear that what we think of as data, and what we think it is good for, has changed radically over time. Once regarded as stable objects whose scientific significance was determined by a handful of professional interpreters, data are now recognised as re-usable goods whose significance depend on the extent to which they are mobilised across a variety of contexts and aggregated with other data, thus growing in volume, variety and value – to the point of driving the very process of discovery. We thus see how the mythology of Big Data is strongly tied to the apparatus of institutions, technologies and economic agreements that effectively enabled data to circulate.

Big Data are not contingent products that simply arise out of particular social arrangement. There are powerful forces at work, determining which data are produced, which are circulated and which are used – and how. Data thus do not simply emerge from human encounters with the world. Data are the result of numerous decisions about instruments, the design of data collection, sampling, protocols, statistical tests, categories, scale and granularity. All these decisions are informed by the specific context in which they are made and by the priorities set by the actors. Furthermore, many of these decisions are not necessarily actively made, but are based on tradition, convention, best practices, or what is learned during training. These are not simply

biases that can be eliminated: all data creation involves selection. The many material, social, political, institutional, technological and economic reasons why we create data in specific ways explain why 'data do not speak for themselves'.

This has important implications for the Big Data mythology that we reviewed in 2.2. First of all, it makes clear that technology does not rule everything and does not determine social life and data uses in any straightforward way. Technology has a fundamental enabling role, and keeps facilitating data exchange and use in ways that were difficult to imagine even twenty years ago: we certainly are in the grip of a digital transformation that is touching every aspect of social life all around the globe (Floridi, 2014). By taking into account the social and economic forces that shape technology development, we can better situate and evaluate the characteristics and effects of this digital transformation, to foster its positive impact and avoid its more damaging effects.

Secondly, understanding the history of big data puts in question the extent to which they are truly comprehensive. Data deemed to be useful for trade by powerful countries have certainly received more attention than any other kind of data. Data about low-income countries, documenting the life of people with little access to computing technologies, are certainly lacking; as are data that are not viewed as valuable by those who have the money and power to produce and buy them. By the same token, data deemed to be sensitive for commercial or military purposes has been jealously guarded among close allies, rather than being freely circulated. Given all this, thinking of Big Data as comprehensive can lead us to overestimate what the data reveals and to underestimate how Big Data, like any other data set, is selective and exclusionary.

Third, it is important to note how causation still matters in the Big Data world. While reliance on correlations to predict future events is certainly growing, the appetite for explanations of what causes those predictions to come true is also expanding. Big Data cannot, by themselves, boost causal understanding of the world. The question is therefore what kinds of knowledge, data collection and data analysis could complement the identification of correlations in Big Data, in order to increase human understanding of how the world works, and why.

This brings us to the fourth point, which is that methods – and specifically methods deployed to compare, evaluate and even critique data and related models – continue to be fundamental to Big Data use, despite the 'messiness' often advocated by Big Data advocates. Knowing whether or not a given dataset is representative of a specific population; being able to train algorithm on a well-constructed data sample; using contextual knowledge to assess whether a correlation is valid or not – these are all crucial skills in the big data world, which call for the exercise of human judgement and cannot be fully replaced by automated tools.

The Big Data mythology strongly underestimates the relevance of social context, the theoretical basis of categories, the importance of accountable methods and the human capacity for assessing complex situations to data work, with serious implications for how priorities are allocated when it comes to management of data. In this chapter, we've shown that data is not an autonomous force that shapes society. To understand the potential of data big or otherwise, it is best to consider how, why and by whom data is considered important.

*Additional Reading*

Hey, T., Tansley, S. and Tolle, K., ed., 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Edwards, P.N., 2010. *A vast machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Floridi, L., 2011. *The Philosophy of Information*. Oxford University Press.

Porter, T., 1995. *Trust in Numbers*. Princeton University Press.

# Chapter 3. Characteristics of Data

## Summary

In this chapter, we discuss the characteristics of data and the ways in which they function, focusing particularly on data journeys. This concept is useful to understand the mobility of data: the extent to which their value derives from being passed around and used in a variety of contexts, and what implications such mobility has on how we understand data work. The view of data presented in this chapter builds on the history of data (and particularly Big Data) recounted in the previous chapter, and provides a stepping stone towards the more general understanding of how data fits into the broader knowledge production cycle that we discuss in Chapter 4. The chapter explores how telephone data is shaped by sociological, economic, technological and infrastructural aspects. These interactions all shape the data and make it selective and far from neutral. Several examples in the chapter illustrate the importance of understanding the characteristics of data in order to evaluate it and make sensible use of it.

## 3.1 Introduction: Data do not stay still

One of the important ways in which data can become valuable is through travel and the changes that such travel involves. The value of data as prospective evidence increases the more it travels across sites. This travel makes it possible for people with diverse expertise, interests and skills to probe the data and consider whether they can be useful to answer the questions they are addressing. For instance, genetic data may well be collected as part of a project in molecular biology, and within that context they can be used to study the functions of a specific gene. Once these data are widely disseminated through databases, however, their value multiplies: for instance, clinicians can use the data to investigate the role played by genetics in disease; pharmaceutical companies can use the data to investigate ways to treat patients; and educators can use the data to produce beautiful visualizations of cell biology for use in schools.

To highlight the dynamic and layered creation of data, we now introduce the concept of **data journeys**. Data journeys designate the movement of data from their production site to many other sites where they are processed, mobilised and re-purposed (see Figure 3.1). "Sites" in this definition do not necessarily refer to geographical locations. Sites can encompass diverse times and viewpoints too (Leonelli and Tempini, 2020).

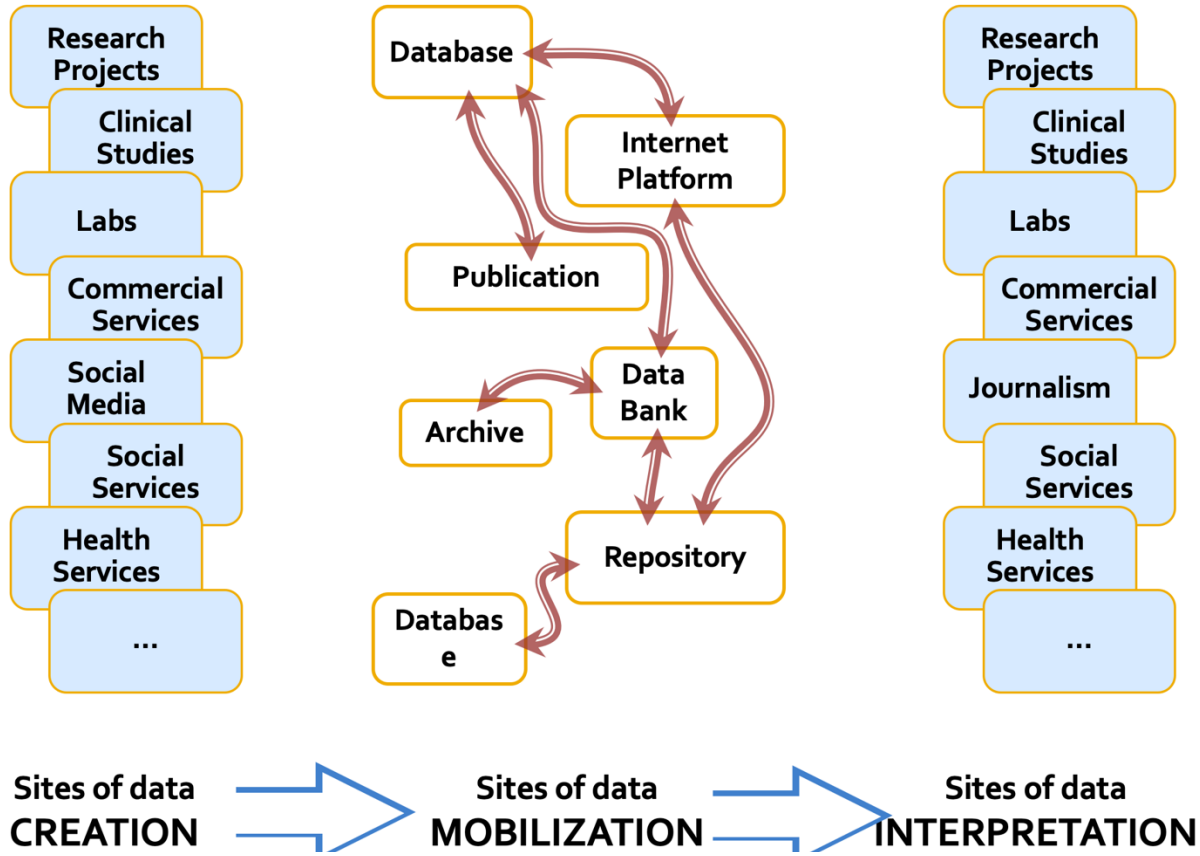



*Figure 2.1 The broad dynamic of a data journey, with data shifting from sites of data creation to sites of mobilization and interpretation.*

The journey of data across these sites involves several stages. At each stage, work with data takes place. It is this work done in each stage that shapes the extent to which data can travel and become usable for analysis and discovery.

The stages include:

- data gathering, which could involve the very production of data through measuring instruments (for example, the generation of location data by satellites) or the collection of pre-existing data (for example, the retrieval of economic data from an archive)
- data processing, which includes the procedures used to make the data useable – such as ensuring that the data are in a machine-readable format
- data cleaning, through which distinctions between "data" and "noise" are drawn, and decisions are made about what features of data to highlight and make machine readable;

- exploratory data analysis to probe what patterns could be extracted from the data, which often uncovers problems, mistakes or gaps in the dataset and thus sends analysts back to the gathering and cleaning stages;
- model design, which could include questions of classification (under which label to categorise the data?), description (what part of the world are data documenting?) and fit to specific problems of interest to the analysts (what are we precisely asking). Again, this stage can require analysts to reconsider their data pool, the ways in which it is processed and the tools chosen to analyse it
- visualisation and interpretation, often conceptualised as the end result of all this work.

Interpretative decisions about what the data may eventually be evidence for, and how to prove it, are made throughout the process. Interpretation is therefore not limited to later stages of data journeys, but is present throughout, even in situations where those busy with cleaning the data do not realise the implications that their choices may have for later analysis. Figure 3.2 illustrates the complex connections and iterations among many of these stages.

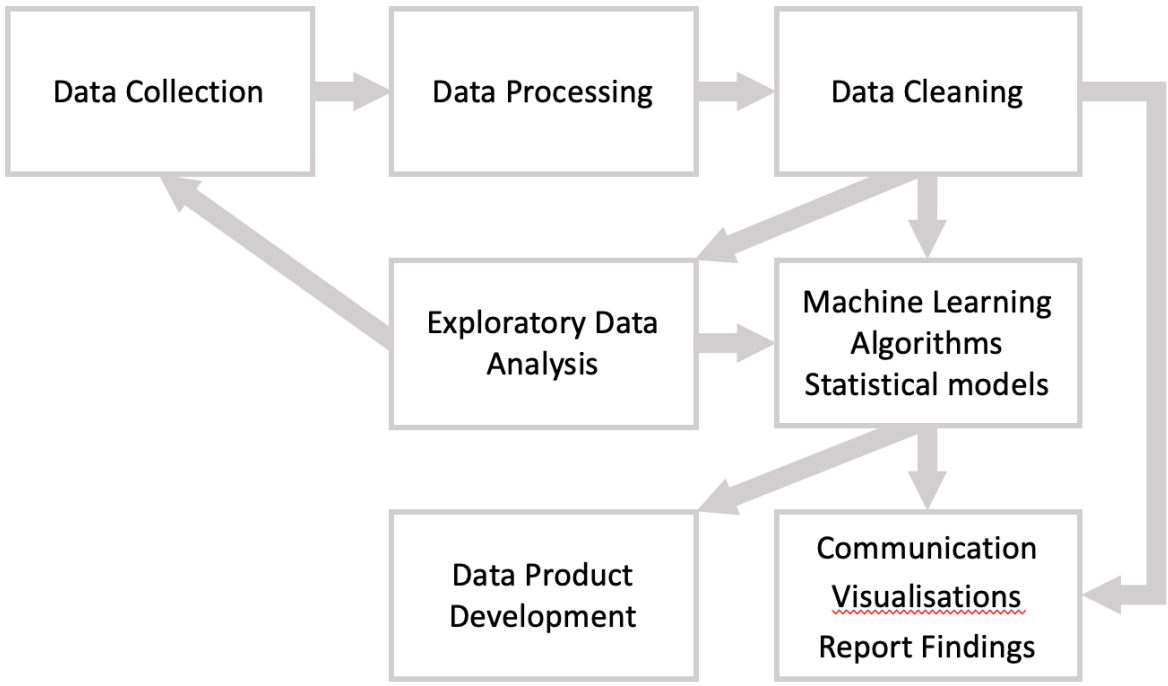



*Figure 3. 2 Stages of a data journey across sites, adapted from Doing Data Science (Shutt and O'Neill, 2013, p.41).[additional arrow needs to be added from Development to Data Collection, since new products frequently seek the support the production and collection of new data—see section on platforms, chapter 5] Note that Shutt and O'Neill define data processing as making data machine readable. We use data processing in a different, more general sense, as making data usable. EU legislation defines data processing in even more general terms, as doing something with data. Such significant variations are a typical challenge of data work.*

Data journeys are anything but linear, and data processing in particular includes many operations that are decisive in determining what comes to count as data, which other data it can be integrated with and how it can be visualised and manipulated as it travels further. Indeed, the metaphor of the "journey" is powerful because, like many human journeys, data journeys are enabled by infrastructures and actions on the part of humans, to various degrees and are not always, or even frequently, smooth. Data may not be able to travel at all, due to proprietary regimes or ethical concerns. It can also happen that strategies developed to make data travel prove to be unfeasible or problematic as unexpected obstacles and disruptions emerge. This could be because a key digital platform is no longer maintained, because of lack of funding for an essential app, or of difficulties in finding analysts with the appropriate skills. Interest in certain data can change swiftly, especially in highly competitive areas such as biomedical research, where interest in data about a given chemical compound can vary dramatically depending on changing perceptions of its potential value as prospective drug.

Data mobility involves risk, so focusing on data journeys helps to identify and evaluate such risks. To understand risks, we have to be aware of the various destinations that data eventually reach and the ways in which they end up being interpreted (Leonelli, 2016a; Bates et al., 2016; Medina-Perea et al., 2019). Prominent examples of such risks include:

- the emergence of errors in the data. For instance, data are copied inaccurately, or irrelevant noise is included in a given dataset by mistake.
- the loss of data due to careless transfers but also to changes in format and storage, which often make it difficult to use current technologies and tools to circulate and re-analyse data collected in the past.
- the misappropriation of data by people who do not have the skills and/or background information to be able to contextualise them appropriately. This can result in problematic interpretations of the data and unreliable/false knowledge being generated. An obvious example is the use of data documenting viewer preferences on streaming services to predict viewers' political affiliation: not only does such a move constitute a potential breach of privacy, but it may also produce false results, since a viewer may choose to watch specific programmes for reasons other than endorsement (e.g. when a right-wing activist watches a left-wing documentary to better understand the opposing side).
- the misalignment between data creation and the uses to which data is put. As we pointed out, data journeys are significant precisely because they bring data into contact with a variety of viewpoints and goals. Where the goals and values of different groups involved in handling data differ considerably, this can give rise to conflict. Consider again a situation where a political party uses data on viewer preferences on streaming services to predict which users may be most susceptible to specific types of political campaigning. There is a potential tension between the interests and goals of the political party in using these data, and the interest and goals of viewers when subscribing to a streaming service. Again, this raises concerns around data privacy which we revisit in Chapters 9 and 10; it also signals the dangers of mobilizing data within a very diverse and unequal social landscape, where some actors have more power to act on the data and use them for their own purposes than others.

The focus on data journeys emphasizes the contextual and dynamic aspects of data creation and use, and particularly the extent to which data themselves may be transformed by their travels (Leonelli and Tempini, 2020). Transformations may well include physical changes, such as a shift in format or medium (from analogue to digital, from a network to a graph-based visualisation). Regardless of whether data change their format or not, what is always transformed as data travel is their significance as evidence.

Think back to the data story on the collection of data on chimpanzees' behaviour in the forest. Data collected on the ground by trackers will be of a variety of types, ranging from numerical counts and samples of chimpanzee droppings to photographs of specimens, weather measurements and observations about the fauna. Once digitised and transferred to a researcher's computer, these data will be used to produce knowledge on animal numbers and movements. But researchers can also choose to post the data on a biodiversity database, where they may be consulted by plant scientists interested in what species of trees are present in the region. Under that new lens, the data will acquire a new significance: rather than functioning as evidence for the understanding and protection of primates, they may be used as evidence for the spread of invasive plant species.

The potential for changes in data use has important repercussions for the data analyst. They need to consider how their techniques and tools are likely to affect the physical properties and evidential value of the data at hand, the speed and ease with which data are circulated and analysed, and the inclusion of certain types of data over others. What types of data best afford which interventions and interpretations? And to which extent do the physical characteristics of data –including their format and the extent to which they fit existing computational tools – constrain possible goals and uses? And what impact do higher or lower speeds of mobilisation have on the reliability of datasets, the amount of uncertainly assigned to them, and the extent to which they are reproducible?

All this is very labour intensive and requires insight into the entire process. Lack of investment and strategy around data travel implicitly supports a naive and unrealistic view of data as "speaking for themselves". This lack of attention to the journey undergone by data can compromise the extent to which data that travel can be reliably interpreted as evidence. Thus, when thinking of a specific dataset, especially one retrieved from a data bank or digital repository, it is very important to gather information about its provenance and the processing it has undergone. In other words, data should be seen in relation to their journey. Some forms of meta-data already exist that help document this (see section on conventions and meta-data in Chapter 5). By paying more attention to data journeys, we are better able to question and critically assess how complete, representative and/or relevant the data are with respect to the questions we wish to ask.

## 3.2 Data are not neutral

Tracking data journeys helps to identify components that are of direct relevance to data use. This helps to take seriously the way data shapes knowledge, rather than dismissing them as lowly building blocks that serve the higher purposes of model building and theory development. This approach is strongly shaped by science and technology

studies that stress the way standards, networks and tools shape data, and connect the nitty-gritty of data production to larger issues of politics and knowledge. Data journeys place the spotlight firmly on the complexity of data and the implications that infrastructures – among many other forces, expectations and material settings – have on their interpretation. Such focus provides a strong counterpoint to many of the hyped expectations and unrealistic promises that have come to surround the use of data, and particularly Big Data. Perhaps the most prominent is the idea that research and discovery can now be fully data-driven, with novel insights arising from the computational analysis of large datasets without the need for preconceived hypotheses or theoretical inputs. This is sometimes referred to as the 'end-of-theory' claim (Anderson, 2008). As discussed in the last chapter, hypotheses, categories and theoretical inputs are already part of how we produce data!

Data journeys help us to discover which conceptual judgments and background knowledge are involved. In the current context of data-intensive research, in contrast to more traditional hypothesis-driven research, such judgments and commitments are made at several different stages of the journeys, and often by different people who may not be aware of (or interested in) what others have decided before them, or why. This means that data journey across highly distributed systems, in which many diverse (or even conflicting) perspectives are embedded. Take for instance the moment in which data are generated, such as when taking satellite pictures of a given location or recording athletic performances on fitness watches. At that moment, what counts as data is affected by the specific sensing technologies at hand and the immediate objectives of data collection (which could be the development of geographical maps in the one case, and monitoring the performance of an individual athlete in the other), which determine the frequency, resolution and scope of the measurements in question. Once data are mobilised beyond that context, however, the objectives and constraints of data handling may change. For instance, if health data are shared with an insurance company, the main objective may become to predict the future probability that the athlete may die of a heart attack, while the main constraint would be data protection legislation which may prohibit specific ways of analysing personal data.

To further explore this issue, we discuss the use of data from telephone communication. The degree of (mobile) phone penetration in many areas of the world is very high by any standard, and it is reasonable to think that data from such practices would tell us a lot. Graphs such as that shown in Figure 3.3 hold the promise of great coverage of data obtained through these technologies. If we compare such rates of penetration (above 100%!) to the response rate of social science surveys for example (usually, the response rate is 12-15%), then it is not surprising that this seems like a wonderfully comprehensive way to obtain data.

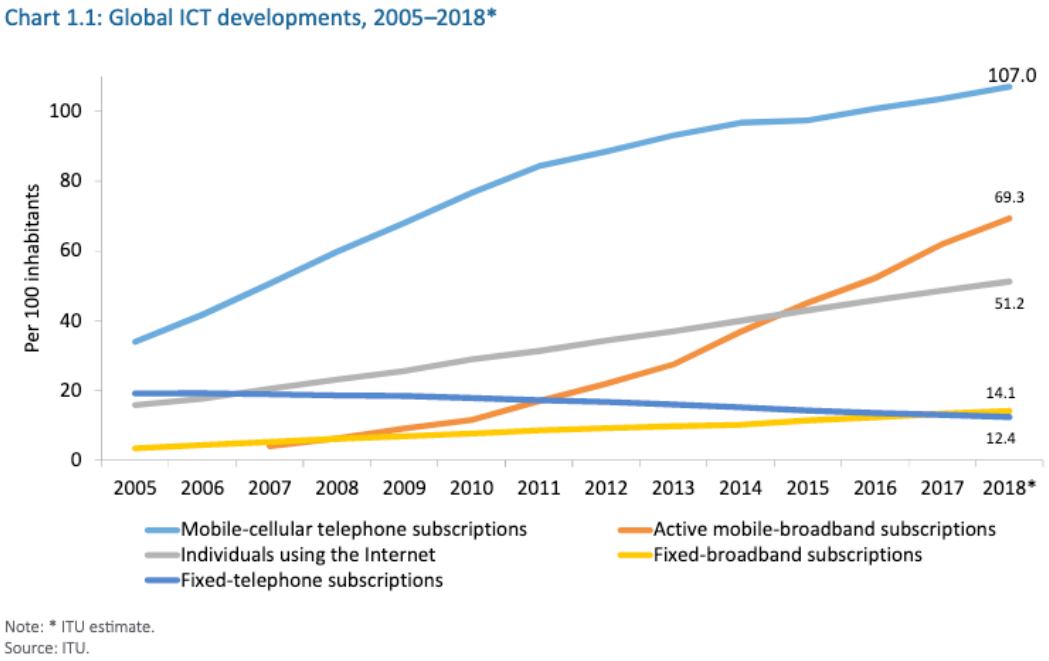



*Figure 3.3 Penetration of different information and communication technologies from ITU, Measuring the Information Society Report 2018 (International Telecommunications Union, 2018).*

A much-used source of data is the call data record (CDR). It is produced from telecommunication transactions, such as a phone call or text message (SMS). Each record has a unique number, and it contains several elements, such as the number called, time of the call, whether it was answered or not, and how long the call lasted. For mobile phone calls, cell-tower IDs (base transceiver station) for both caller and callee are also available, which can provide information on location. CDRs are therefore the basis for plentiful and reasonably well-structured data.

This data is in the hands of service providers who use it for billing purposes, to evaluate the quality of their service, and to analyse the behaviour and needs of their customer base. Beyond this primary context, CDRs are also widely used to generate data for a range of marketing, humanitarian and scientific research. Mobile and landline phone data have been explored as sources of information on mobility following a disaster, on economic welfare (since higher top-ups are taken to mean more income), on social networks, and even on the likelihood of the spread of disease through tracking movement from non-infected to infected areas. Typically, when reporting on the use of such data sets, the main features of the data used in the analysis are described. Exceptionally, authors also report on the collection, cleaning and processing. In the example below, we draw heavily on two texts that respectively describe a data set used for a hackathon activity (Blondel et al., 2012) and a methodological exploration of network analysis using CRDs (Decuyper et al., 2016). It is telling that these detailed accounts of how data is constituted are not typical scientific journal articles. Published journal articles tend to focus on the analytic methods and results, rather than on data work.

Let's go on a data journey and review what happens to CDRs when they are treated as research data. Typical in the study of social networks using CDR data, many researchers start by removing links or nodes that are not active enough (users who don't make enough calls). Researchers also often remove links that are not reciprocal (if one user calls another user, but is never called), or impose a minimum number of calls to take up a link in the analysis (being called at least 6 times) (Decuyper et al., 2016). Users who join or leave a provider during the observation period are also usually removed from the data set. In addition, calls are paired to avoid double counting (an incoming call for a callee is an outgoing call for the caller). The time window chosen (days versus months) can also strongly shape the data set, by excluding less active users. Such decisions can orient the analysis in radically different directions. Decuyper and colleagues (2016) have demonstrated that these decisions can go so far as to lead to different distributions in the sample, and that what has emerged as the accepted form of networks may be based on assumptions made at the data cleaning stage, rather than reflect the actual patterns of social networks.

In the study of social networks, data was traditionally generated using questionnaires in which research participants were asked to list their connections. There was always a concern that participants would either provide incomplete lists, because they would forget some of their connections, or that they would provide socially desirable answers – for example, wanting to seem popular and over-reporting contacts, or else not listing some types of contacts if the participant expected that these might not be socially acceptable. With social media and other telecommunications, there have been claims that radically better data on social networks would become available, since the subjectivity and partiality of the participants' reports would be bypassed. The researcher could capture ALL the relevant data without the introduction of bias from the subject and without missing any of the connections. Clearly, as we discussed above in terms of data cleaning, there is also subjectivity and selection going on with the use of CDR, albeit of a different kind.

When working with data from mobile phone use, data cleaning can involve a number of substantial decisions about network data. There are other considerations in producing location data that also shape the data available. In some situations, the provider may consider that the precise location of the antennas is commercially sensitive information, and so blur the exact location to protect its commercial interests. (The position of antennas affects the quality of service delivered, so positioning them strategically is important). At times, the antenna identifiers are simply not available for technical reasons. Location information may also be removed or blurred in a dataset to protect the privacy of individuals: with enough data on calls, the location of one's home or workplace could be deduced by identifying locations from which calls are regularly made. Such decisions about which users to include in the data set and to which degree of detail location should be included are defensible, and taken for good reasons. Yet, it is important to realise that each of these decisions shapes the dataset, and qualifies the promise that we have all the data about everyone.

We now turn to other elements that shape the data. We indicated above that the data is collected by the network operator. This is important for understanding whose data will be available and what the population contained in the selected sample will be. To what extent are CDRs inclusive? In the context of increasing privatisation of

telecommunications and of the growing use of mobile services, which are nearly always privately operated, the users of a given provider are not necessarily representative of a general population. The data provided by an operator reflect who its customers are, and might be skewed in terms of wealth, gender or culture. If the operator is more popular with certain groups, determined by, for example, age, income, gender, language or occupation, the coverage of the dataset will reflect these tendencies. The data from mobile phone use therefore reflects not only calling behaviour but also market dynamics. Furthermore, call activity is increasingly spread over different operators, so that data from a single operator will give partial information. As Decuyper and colleagues note, such biases are very difficult to remove without access to a dataset with a perfect coverage, which does not exist (Decuyper et al., 2016). Indeed, no data set is ever perfect, and there is no neutral basis from which to generate data. This does not mean we should abandon all hopes of 'good data' or that all data is hopelessly biased. It does mean that we can only properly understand outcomes of data analyses if we also take into account the generation of the datasets that underlie analyses and document data journeys.

Besides the selection by researchers and the biases in data production, the behaviour of users can also shape CDRs. Decuyper et al discuss "flashing", which consists of letting a relative's phone ring a couple of times, then hanging up and waiting for them to call back. Such technique insures that, while either party can take the initiate to have a call, it is always the same person who pays for a communication. But if the filtering technique used to clean the data requires reciprocity (at least one successful phone connection in each direction) then such links will be removed from the dataset (Decuyper et al., 2016). Should such links be retained? It depends on the question. But one may well imagine cases where such relationships, in which one party is willing to foot the financial costs to contact the other party, can be very meaningful social ties and should not be systematically erased from data sets. In addition, it may be the case that phones are shared between individuals, that sim cards get passed around and used in different phones, or that some users have more than one phone (Erikson, 2018) and use different service providers for different kinds of communication (work-related vs personal). All these user behaviours will shape how the data looks and what the data can be assumed to mean.

Finally, technological platforms are also determinant of the generation of data. For CDRs, this is manifest in the fact that they do not exist for all calls! If calls are made via Wire, Skype or Whatsapp (or other types of VoIP phone calls), CDRs are not generated. If particular types of calls are made using VoIP or if specific types of users have strong preferences for these platforms, they will be systematically underrepresented in CDR datasets. Digital platforms also have log files that would enable an analyst to retrieve data about calls made using Skype or Whatsapp. But this data set would, like CDRs, also be shaped by the kinds of uses and users it attracts, and have its own limitations. Digital platforms are furthermore increasingly malleable and generated on the fly—this means that they are less stable as a basis for gathering data. There are always deletions, delays, errors, repetitions, glitches, updates and differences that arise from the many portable technological supports through which platforms are used.

These several considerations about how data sets are shaped even before analyses are performed suggest that we can better see data as a lens, rather than a window. A lens

orients us to a particular way of looking at the world, rather than providing a transparent way of looking at it. The assumption of 'having all the data' tends to stress the view that data is a transparent window. As we saw with the example of CDRs, data only document a fragment of reality, seen through a specific perspective and constrained by specific instruments and formats. Second, there are always non-users of a technology or users who use the technology in a radically unexpected way, so there is never 100% representation in the data either. This is neither a shortcoming of CDRs nor of digital data per se. All data sources have limitations and decisions always need to be made about which data to include in an analysis. The problem arises when this need to evaluate, clean, shape and otherwise select data is erased or considered trivial. Furthermore, loud claims about the total capture possible by Big Data overshadow the partiality of data and can cause analysts to neglect making this partiality visible and to underestimate the need to investigate the sources of that partiality.

Returning to the CDR example: the many qualifications made above do not invalidate the use of this data, but they should make us aware of the need to carefully account for the ways in which data is not neutral. As we saw, to evaluate a CDR data set, we need to be aware of the market share of providers, of the physical characteristics of the transmission system over which calls are made, of privacy regulation, of calling cultures (such as flashing) and of preferences of users for platforms. Sociological, economic, technological and infrastructural aspects all shape the data and make it far from neutral. CDR data can be very valuable, especially when we are able to value it in relation to careful accounts of its non-neutrality. If we take data journeys seriously, then it follows that there is no such thing as neutral data, and that it is not possible to have all the data. The journey of the data has inevitably selected and shaped the data in particular ways. This insight should not be read as a call to remove subjectivity by implementing as much automation as possible or by standardizing as many aspects of data creation as possible. What is more insightful and productive is to ask is how data creation and the rest of the data journey come into being.

## 3.3 Data are context-dependent

It is tempting to think that the scientific significance of data lies in their context-independence, and the extent to which they objectively document the world without any interference from human interests and values. As shown in the preceding sections, this is, however, clearly not the case. In order to do good data science, one must carefully consider how to contextualize the data as well as the processing tools, questions and background knowledge through which the data are analysed. If we consider data in context, this is improves the accuracy and reliability of knowledge being produced, the understanding of the problem and/or situation being studied, and the potential impact of processes of datafication. How data are interpreted often changes depending on the skills, background knowledge, and circumstances of the analysts involved, which is why looking at the same dataset from a variety of viewpoints often yields new knowledge. Maintaining an awareness of how data move across contexts, and with which implications, is therefore crucial to the analyst.

Remarkably, studies of data re-use across contexts also show that the expectations and abilities of those handling and mobilising data determine what is regarded as "data" in

the first place (Leonelli, 2016a; Borgman, 2015). Researchers make choices about which of the objects produced through their interactions with the world – whether they be experimental interventions, observation studies, or measurements – deserve the most attention as potential evidence for claims about phenomena or specific courses of action. Biologists, clinicians and plant breeders differ considerably in the data they will consider most useful towards studying gene-environment interactions; and there are many documented cases in archaeology, astronomy, biomedicine, and physics where objects considered as data at the start of an investigation no longer have that status by the end of it, or vice versa (Leonelli and Tempini, 2020). A set of photographs taken in a forest, for example, could constitute useful data for the study of phenomena as diverse as the growth pattern of a given tree species, the symptoms of an infection, the effect of certain meteorological conditions on photosynthesis, and the presence of parasites in a specific location. Each of these interpretations is affected both by the physical features of the photos (definition, level of detail, focus of attention, colour schemes) and by the manner in which whoever handles these objects accentuates their usability as data (for instance by zooming on a specific detail, adding metadata, and/or changing format to foster interoperability with other botanical data).

*Figure 3.4 Photo as evidence Date: 1940 Location: South Moluccas, Indonesia, Photographer: unknown, Tropenmuseum of the Royal Tropical Institute, KIT. Rights: Photos downloaded from the Essential Lens site are cleared for educational use only.*

A photo such as the one in Figure 3.4 has a data journey spanning decades. It is a photograph that belongs in the collection of the Tropenmuseum in Amsterdam, the

Netherlands. It was probably taken as evidence of the Dutch colonial administrator of the Dutch Moluccas fulfilling his duties. It later became evidence of colonial relations, as it is currently available as part of digital collection entitled Economies and Empire: Colonialism and the Clash of National Visions. It could also be used as evidence of how software can be used to correct horizon… In each of these three instances, different aspects of this photograph count as data, and the photograph has a different relationship to technologies and infrastructure (photographic prints, photo albums, colonial archives, digitization projects, databases of images, image processing software). As you read this, this photo is put forward in a textbook, an educational context, as evidence of the importance of data journeys – yet a different use of this data. Hence, while the features of the objects considered as data certainly shape their use and interpretation, it is often possible to obtain different information from objects depending on how these are managed and interpreted. A particular combination of interests, abilities and accessibility determine what is identified as data in each instance.

Similar considerations apply to data in numerical form. Such data are, on the one hand, eminently transportable: they are easy to copy, aggregate and visualise. On the other hand, what numbers may be taken to represent can vary just as dramatically as in the case of the photograph above. Some numbers may have very different interpretations, such as numbers used to count Covid19 infections: politicians may see them as documenting the success of a public health intervention, while medical professionals could seem them as part of a trend leading to an overwhelmed medical system. Other numbers are taken to indicate different things altogether, depending on who is using them. For instance, the fact that 66% of Greeks had access to a smartphone in 2017 can be seen as indicating a high rate of penetration of digital technology in everyday life in Greece or as indicating that 44% of Greeks have much less access to digital communications. Even numbers that have clear standardised parameters, such as measurements of length, could be interpreted differently depending on which unit of measurement is being considered: it matters a great deal whether a measurement is taken in inches or in meters, for instance.

## 3.4 Conclusion: Characteristics of data

We started out with some of the claims that data is neutral until analysed and that Big Data is comprehensive. Such belief in the power of Big Data to provide neutral access to reality and to describe everything has been called **'Big Data empiricism'**. This indicates a belief that data is the best form of evidence to establish truth, to form opinions about the world, and to make judgments: data can somehow "speak for themselves" and reveal the truth to those who know how to decipher them. Big Data empiricism therefore places Big Data in a privileged position to count as evidence (Rieder and Simon, 2016. This is very significant, especially in combination with the increasing use of algorithms and other automated tools that process such data – something we examine in the next chapter, where we discuss the relationship between data and knowledge.

In this chapter, we gave reasons to mistrust Big Data empiricism. Thinking of data as "speaking the truth" - independently of how they are handled, where, when and by whom - is highly problematic. A much better starting point is to recognise that using

data involves being able to contextualise the data appropriately, which in turn involves understanding as well as possible how data have been generated and mobilised. We highlighted the importance of data journeys, and the implications of emphasising the mobility of data to understand the characteristics associated with data. To summarise our findings, here is a list of the key features of data elicited from studying how they are handled and used in society:

- Data are made, not found: this draws attention to how data are created, which enables us to better understand how to use them as evidence

- Data are partial: this leads us to ask about what was not captured in the data and about who is excluded. It draws attention to what we expect from users, and to the levels of compliance we assume from users. This also helps us understand how strategies of use (some users avoid using systems in specific ways) or design elements (systems can only be used on some devices) systematically exclude the possibility that some data will be gathered.

- Data are limited: this draws our attention to where the data begins and ends. For example, we need to pay attention to how changes in legislation, in technology or in cost structures mark the start or end of types of data. Another consideration is the growing and waning popularity of certain platforms among specific demographic groups.

- Data are shaped by technologies: we can describe how data are generated and retrieved, and evaluate the kind of variation can be expected, due to instability/dynamism of the systems.

- Data are contextual, not neutral: the technologies through which data are generated are themselves products of specific social circumstances, which need to be understood in order to better situate the data.

- Data contain assumptions: about what the data could be used for, which were built into the data at the moment of acquisition, and may have been reinforced or changed as data were processed: Which data were removed in cleaning? Why? Which corrections were systematically applied? Why? What kind of noise is there among these data and what does this noise say about the signal?

- Data use is contingent on goals: the interests and aims of data workers play a central role in determining how data are handled and to which effect. As clearly demonstrated by clever statistical manipulations, it is always possible to interpret data to suit a wide variety of agendas. It is therefore imperative to make processes of data analysis, and the values and interests underpinning data work accountable.

- Data are changeable, not fixed: whenever the format or medium of data travel shifts, for instance when changing the type of file data are encased in to fit a new programme, the ways in which data can be analysed and handled also change – often leading to a shift in what data can come to signify.

We need to keep these points in mind whenever using data. Being mindful of the positionality of data fosters the ability to evaluate data according to their provenance, their merits and disadvantages, and to critically assess their value, quality and limitations. This improves the reliability of the data themselves and of the use we make of them as evidence. And in turn, paying attention to the characteristics of data helps us make the conditions, priorities, interests, and judgments that shape the data explicit and open to scrutiny.

### *Additional Reading*

Borgman, C.L., 2015. *Big Data, Little Data, No Data*. MIT Press.

Leonelli, S. and Tempini N., 2020. *Data Journeys in the Sciences*. Springer International.

Schutt, R., and O'Neil, C., 2013. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media.

Strasser, B., 2019. *Collecting Experiments: The Making of Big Data Biology*. Chicago University Press.

# Chapter 4. Data, Evidence and Knowledge

## Summary

In this chapter, we build on our analysis of the characteristics of data and examine how data can function as evidence. We show that how we think about data matters. It has important consequences for data management and use of data. We begin by introducing two different ways of conceptualising data, the *representational* view and the *relational* view. We analyse the differences between these two views and then illustrate some of the failings of the representational view – and the reasons why understanding data relationally has a positive impact on all aspects of data work. To discuss the contexts in which these two views are used, we use the term "knowledge production". This includes academic research as well as other forms of research grounded on data that produce marketing insights, predictions about future behaviour, or indications for running a business. We conclude the chapter with a reflection on how data fit the broader space of knowledge production and explain how knowledge is extracted from data.

## 4.1 Introduction: The representational and the relational views on data

When we work with data, whether as researchers in a university setting or as analysts in a company, we have a view on data that informs our decisions and actions. By a

"view", we mean some fundamental expectations about how data relates to the world and how we can rely on it. This chapter is about conceptualising how we think about data and what we think it can do. As an illustration, take the tagline for Twitter: "it's what's happening". We might take this at face value—Twitter activity is a representation of what is happening *in the world*, or at least in the lives of the many people who use Twitter. Or we might interpret the tagline differently: this is what is happening *on the Twitter platform*, based on what its users want to communicate about and what is shown to me through my feed as a result of my decisions on who to follow and of the platform's algorithmic selections of which tweets to put in my feed. This very simple example shows the contrast between a representational and a relational view of data.

The representational view of data is perhaps the most intuitive and popular approach to conceptualising data and understanding their role in knowledge production, so we start from there. Within a representational view, data are objects that capture and represent specific aspects of the world. In this view, data constitute the starting point for empirical knowledge. Figure 4.1 provides a graphical explanation of the representational view.

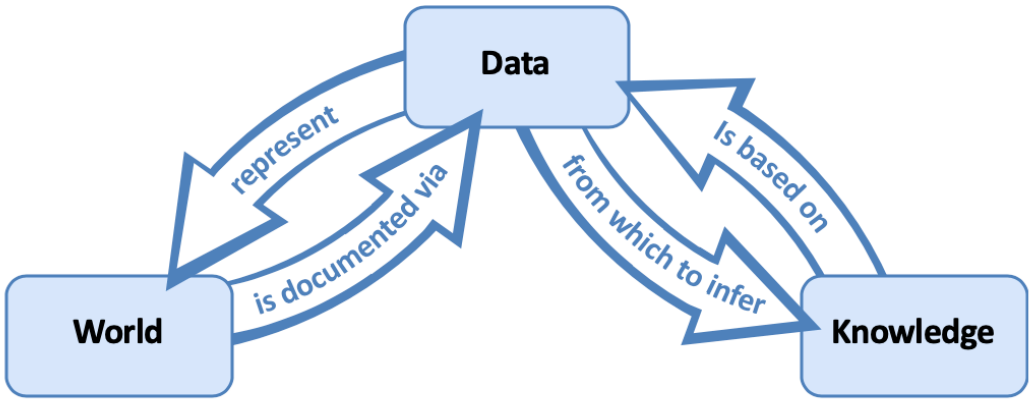



*Figure 4.1 A representational view of data.*

In the representational view, the more data we have, the more information we have about how the world works – from which knowledge can then be extracted. The representational view acknowledges that when scientists gather data, they do so in a structured way, and that certain aspects of the world are prioritised and highlighted whenever a dataset is produced. For instance, scientists will focus on some aspects and downplay others because of the questions they are asking. They select and structure data on the basis of what they already know and what they wish to know. In turn, this is shaped by historical developments and current circumstances. Because of this selection and structuring, data are not simply a mirror of reality, but rather a way to depict it so that it can be analysed and better understood. Of course, scientists are not the only ones gathering data. When data from social media platforms are gathered, there is also a structuring and selection of data, and some data will be of greater interest than others.

In what follows, we use examples ranging from genetics research to the marketing of chocolate.

**Raw data**, in the representational view, are data that have just been generated and have not been further processed. In this sense, they are the closest documentation that we can have of the world "as it is", independently of human ideas and interference. Raw data is thus taken to provide unmediated access to reality. This implies that the information content of data is regarded as fixed, regardless of how data are used. The challenge for research methods is then to uncover what it is that the data indicate about the world, by "cleaning" the data from the noise that may result from imperfect human measuring practices. Controlled conditions of data collection and sophisticated statistical analyses are among the methods that make it possible to evaluate what information a data set contains about the world. The main focus of research methods is to improve the representations and remove as much human (or other) bias as possible from the process of data analysis. Data, in the representational view, are untainted by human interpretation: the main purpose of data is to help test theories independently of the specific biases and fallibility of human perception.

In the relational view of data, by contrast, the main focus is on understanding data based on the relations that data has to many aspects of knowledge production, including existing knowledge, social context and human agency. In this view, data is the product of human interactions with the world. They are not representations per se, which have meaning regardless of context. Instead, what data may or may not represent is the outcome of a process of inquiry pursued by humans in a given context. In the relational view, there thus is no such thing as **'raw data'** as an abstract category. In interacting with the world, we create objects. What information you think these objects provide about the world (and what should instead be viewed as noise) will depend, among other things, on how you want to use them. The objects will only become data once you have decided that they are to be used as evidence for a particular claim. This intended use is what will help you remove the noise and retain the valuable information. What counts as data depends on what you do with them. The question 'What are data?' cannot be answered in the abstract in a relational view. It can only be answered with reference to specific (research) situations, in which investigators make decisions about which data could be used as evidence. This is why it's valuable to define data in terms of their function rather than in terms of intrinsic properties.

When working with a representational view of data, then the focus is on improving the conditions under which the representation is created – thus, the stage of data collection. Once a dataset is created, it is trusted to contain a nugget of truth that needs to be disclosed through appropriate methods. The bulk of research efforts is thus directed at finding good methods (typically statistical tools and/or algorithms) towards analysing the data. By contrast, the relational view defines data as objects that are treated as potential or actual evidence for claims about phenomena in ways that can, at least in principle, be scrutinised and accounted for (Leonelli, 2016a). The relational definition acknowledges that data are powerful but unpredictable objects. Their value as evidence is not fixed and may increase, the more data are shared and scrutinized across multiple contexts. The meaning assigned to data, and thus their value as evidence, is determined on the basis of their provenance, their physical features and what these features are

taken to represent, as well as the motivations and instruments used to visualise them and to defend specific interpretations. Something becomes data when it is used as evidence for a claim. For instance, a poem can be just a piece of literature; when it is used as evidence to claim that Marlowe is actually the author of Shakespeare's sonnets, that poem becomes data. Similarly, a photograph of a tree can be a nice memory of a sunny holiday – but if it is taken as evidence of the health of that tree on that particular day, that photograph becomes data. These everyday examples are no different from the way data work in Big Data contexts or in audit cultures where indicators dominate. A 'like' on Facebook only becomes data when it is recorded, collected, connected to the liked object and/or the profile of the liker, and compared to other 'likes' in order to serve as evidence to make a claim about popularity, social ties, or preferences.

The relational view works with a functional definition of data: data are only such by virtue of their function within a given situation of inquiry, and their relation to the inquirer, the nature of the phenomenon being investigated and other components (such as relevant models and algorithms). This framework acknowledges that any object can be used as a datum, or stop being used as such, depending on the circumstances. This is a well-known consideration to anyone dealing with historical data, often held in forgotten archives and therefore reduced to meaningless objects. The importance of circumstances for determining whether something is data or not also highlights that the mobility of data matters enormously, as discussed in the preceding section. The relational view acknowledges that all aspects of knowledge production are connected: changes in one aspect will affect the other steps as well (see Figure 4.2).

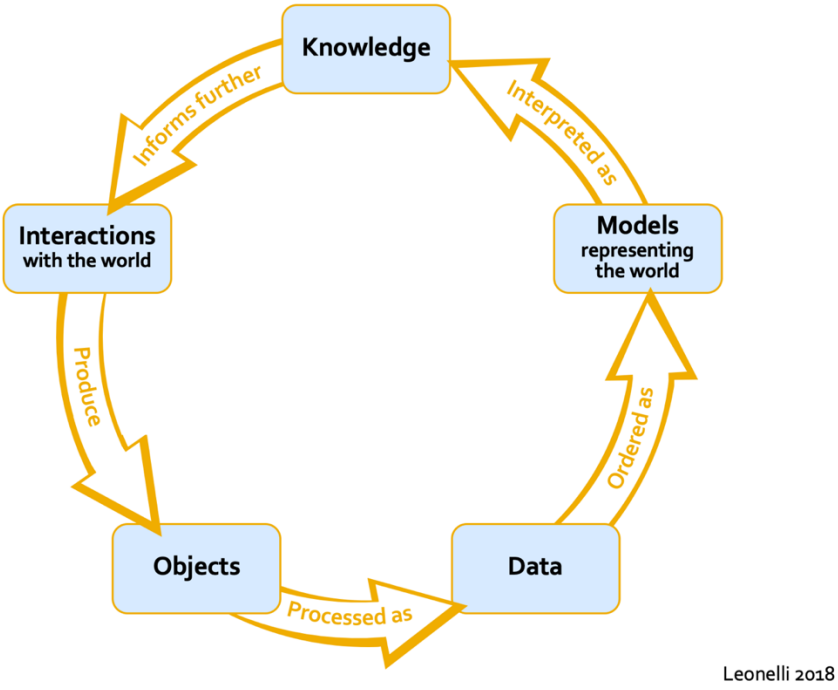



*Figure 4.2 The cycle of knowledge production according to the relational view on data. Adapted from Leonelli 2019 [EJPS].*

## 4.2 What is evidence? The path from data to knowledge

What, then, is evidence? In order to understand this better, we need to take a step back and consider how data is actually transformed into knowledge (see Figure 4.1). Empirical investigation starts from the interaction between humans and the world. These interactions produce various types of objects such as numbers, measurements, symbols, photographs, descriptions and graphs. Some of these objects are then selected and processed to become, at least in part, a source of knowledge. What we call data are objects that we manipulate in this way. What this data are used to stand for is not determined exclusively by the physical features of the data themselves. The preconceptions and context of those evaluating their potential meaning also matter. Interpreting data as a source of knowledge therefore goes through another two stages: (1) the development of ways of ordering data that reveal a specific representative function - often understood as the process of data modelling - and (2) the use of the resulting models as an empirical foundation for the production of knowledge. In this view of knowledge production, therefore, the representative function of data continues to be present, but it is not data in themselves that are taken to represent the world. It is instead the **data model**, selected by whoever is interpreting data and deciding how to organise them, that carries out the representative function. In other words, it is a certain ordering of data – the way in which it is viewed and made relevant to a specific dataset – that represents a particular aspect of the world and makes it accessible to further analysis. It is by ordering data that data become usable as proof of specific facts and as source of new knowledge. Data are not by themselves an objective foundation for knowledge; it is the way we organise and view them that determines its meaning.

It is important to note here that the term "knowledge" can be itself interpreted in at least two ways, and that data are central to how and what we know in both of these senses. The first interpretation of knowledge is as the set of claims that we take to be true, as when "knowing that" something is the case. For instance, if you want to know what role genes play in the cross-generational transmission of cystic fibrosis, you can look this up in a medical textbook which will explain this to you. The second interpretation of knowledge is as the skills and strategies that we need to intervene in the world, as when "knowing how" something can come about. For example, when wishing to know what to do in case of a heart attack, you may sign up for appropriate training by a reanimation specialist. In both of these cases, data are the empirical foundation for the knowledge in question: without data documenting the correlation between the incidence of cystic fibrosis and the possession of certain genetic traits, there would be no evidence for the claim that those genes are a reliable marker for the disease; and without data on the effectiveness of providing certain kinds of help to people suffering of a heart attack, there would be no evidence for the adoption of a certain strategy of intervention rather than another.

Inquiry is thus best depicted as an iterative process consisting of five key steps, represented in 4.2 above: (1) the production of objects of investigation through interaction with the world; (2) the processing of such objects so that they can function as data, which unavoidably involves a restriction in the evidential space within which data can be credibly used; (3) the ordering of data through **data models**, so that they can represent specific phenomena; (4) the use of data models to develop knowledge claims about those phenomena; and (5) the use of knowledge claims to frame further interactions with the world. Theory is involved in each of these steps, but in different

ways. Steps (1) and (2) affect what ends up counting as data, and theoretical commitments are mostly incorporated in the choice of materials and samples, experimental instruments and data sharing procedures – the ways, in other words, in which researchers carve nature's joints and thus limit the conceptual space within which data can be used as evidence. Steps (3) and (4) are where researchers actively question, identify and stabilise conceptual assumptions about the nature of phenomena to be investigated, which shape the content and formulation of the knowledge claims being produced.

## 4.3 Examples of data within the knowledge production cycle

Let us consider the example of botanical data to illustrate this process more concretely. In this instance, an amateur taking pictures in the woods produces objects through their interaction with the world. These objects - their photographs - may then be processed by researchers with the expectation they may function as data (for example when these photos are formatted and loaded into a database). The researchers order and organise data thus obtained in ways that help them represent different phenomena depending on their interests and specialism: morphologists may analyse the shape of leaves of a specific tree species in a certain location, and use the photographs create models that represent different leaf shapes and their relation to the characteristics of different parts of the woods. Or pathologists may look for the visible symptoms of potential tree diseases, and use the photographs to produce models that indicate the incidence of a given disease in the forest. The different models created by these researchers are then tested to verify their reliability and relevance to the phenomena they document - for example, pathologists check that the disease symptom model derived from the photographs found online matches the features of data models coming from other sources and, when possible, they return to the location in question to verify the truthfulness of the model. If the models are found to be adequate, they are used as a source of knowledge on the way that disease manifests in the plants analysed. If they are judged to be inadequate, researchers go back to analysing the data and try to order them in different ways - a process that sometimes requires radically changing the type of objects considered as data and/or the aspect of reality being investigated: perhaps the pathologists have been considering the wrong disease, for instance. This example shows how, in order to give rise to knowledge, data need to be manipulated, cleaned and ordered to fit/inspire/support a particular representation of the world, that is, a model. Modelling is the stage of inquiry where a link is made between what are considered to be data and the aspects of the world that the data are supposed to be documenting. Once data are made to fit a specific model, their value as evidence is established.

Another example is the analysis of telephone call data records ('call data records', or CDRs), a type of data that we discussed in relation to data circulation and cleaning (Section 3.2). CDRs were designed to keep track of telephone service use and to generate reports that would enable companies to claim funds from their users. In this sense, CDRs were designed to act as data for the telephone service provider, since they were used as evidence of phone use for which a customer could be billed. Beyond this original purpose, CDRs have been used to understand a variety of different phenomena, like mobility and migration, traffic patterns, social networks, market development opportunities, etc. A claim like 'there is demand for more capacity of mobile network in this part of the city' is a knowledge claim based on objects (CDRs) that become data once they serve as evidence for this claim. It involves using a model of telephone activity

and of user behaviour in a physical space that explains the real-world activity that leads to the generation of CDRs. This example shows that such a model is not necessarily explicit, and those who produce it and use it may be doing so without being aware of it. The model may indeed be based on everyday experiences or common sense. If you think "when people use their phones, this activity becomes visible as a CDR", this is based on a specific understanding of the relationship between phone use and the creation and recording of a CDR. This understanding can work as a data model when it is used to analyse and interpret CDR data. (Of course, a model can also be more conceptual, expressed as a diagram or in abstract mathematical notation – we return to the issue of data modelling more systematically in the next chapter).

Data are also a factor in other settings, where the traces are produced differently. How about a claim like 'the best time to market our new chocolate bar is when our customers are most likely to buy chocolate-related products, which is between 16.00 and 18.00 local time'? This claim is based on buying patterns that are identified using objects such as timestamps and items on receipts listing chocolate as one of the items purchased. These become data when they are put forth as evidence of purchasing patterns. In their role as data, timestamps and itemized receipts can be compared to check for patterns of purchases, with receipts containing chocolate being clustered depending on time of day. The resulting knowledge about preferred time for buying chocolate will be used to shape future actions, such as prominent placement of ads for chocolate products at given times. In turn, this can be monitored by paying attention to objects that can serve as evidence as to whether the ad campaign is working (for example, trends in sales revenue).

A final example: consider a claim like 'to ensure that our citizens experience the best service from the municipality, we have to ensure that the passport desk has extra capacity 6 weeks before the school holidays begin.' The claim is based on data about the number of times that citizens access the web interface to make an appointment with the passport office in the weeks preceding the school holidays. The evidence to support this claim consists of the traces left by website use, specifically the 'make an appointment' feature. These data that can be combined with the types of data, such as reports produced by employees of the passport office who identify a large number of "last minute" users of their service – thus corroborating the need for higher capacity to ensure that all those who request a passport in the weeks before school holidays can actually get one in time to be able to travel.

Given these examples, we can now come back to the question of what evidence consists of in the relational framework. **Evidence** is the specific arrangement and formatting of data, which is taken to corroborate (provide reasons to believe in) a particular claim. In other words, evidence is "ordered data": data that have been prepared, managed and visualised within a model, to serve as evidence about the world.

## 4.4 Contrasting the representational and relational perspectives

Both the representational and the relational views on data agree on a number of key points:

- data are the result of complex interactions between researchers and the world

- interfaces such as observational techniques, measurement and registration devices play an important role
- rescaling, modification, standardization of objects is needed to make investigation possible, and this holds for many different kinds of objects, including numbers as well as observations.

However, a relational view draws more attention to the whole cycle of knowledge production. A relational view highlights how the whole cycle is important for how objects underlying data are created, the models that go into shaping this, the kind of mobility we give to data, the kind of knowledge we create and the ways this changes how we act in the world. We saw the importance of this when we looked at datafication and made the link between how this process changes society. If we were to take a representational look at datafication, we would be focusing on the creation and statistical treatment of the traces. Any effects we might observe would then be easy to condemn as 'bad use of data' and we would not pay attention to the way particular data orient us to particular uses (and vice-versa). So, if we want to understand how football metrics are changing how we value players, from a representational view, we would ask: are we measuring the right behaviours accurately enough? How can we improve sensors? Is the granularity sufficient? Do we need to measure vertical as well as horizontal acceleration to get good data about speed? From a relational view, by contrast, we would ask: what are the effects of the measurements of players becoming evidence of being a good football player, when we use data about completed passes or ball possession or km run during a match? How do these data make sense as evidence, based on our model of what happens during a football match to make a team successful? We might also look at what this does to how we value players (as good/excellent/exceptional), how we train them and trade them, and how we build stadiums that have sensors, screens in the locker rooms and dedicated spaces for the analysts along the field and how our knowledge of football is changing from an evaluation of teams to an evaluation of players as individuals. We could look at how such data travels to shape FIFA computer games, as well as betting shops and also affects the analytic insights of tv commentators.

Another way in which these views matter is in terms of how we value knowledge in our society. Consider the contrast between the relational view and a pyramidal representation of "data-to-wisdom" often found in knowledge management, systems thinking and management science textbooks (see Figure 4.3)

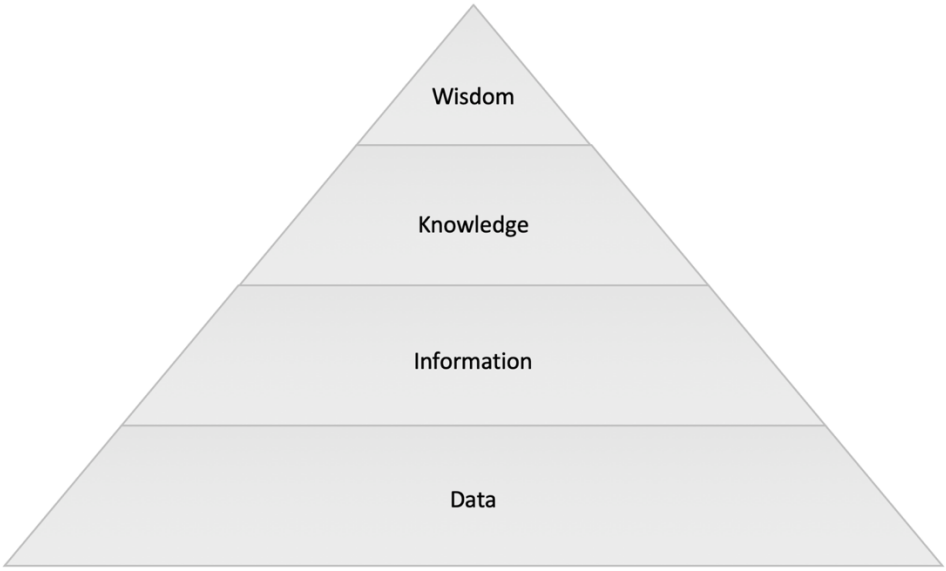



*Figure 4.3 Data-to-wisdom Pyramid Model, from (Ackoff, 1989).*

An often-cited source for this hierarchy is T.S. Eliot's *The Rock*, written in 1934 as a commissioned Christian pageant-play, written as part of a fund-raising effort to build churches in the suburbs of London. These two lines are usually quoted without much attention to their context: "Where is the wisdom that we have lost in knowledge? / Where is the knowledge that we have lost in information?" The pyramid model was formalized by R.L. Ackoff (1989). This model supports a number of ways of thinking about the relationships between its levels, and it is generally understood to go from low to high value, and from low to high meaning. Furthermore, when discussed in the context of automation and digitization, the consensus is that the lower levels can be automated – information might become the realm of AI –but that the upper parts may remain uniquely accessible to humans.

This model assumes that there will always be more data than there are claims or evidence, and that from such masses of data some evidence can be distilled, from which, in turn, a narrower set of claims are extracted or inferred. Furthermore, it assumes that data are the least valuable element, while knowledge constitutes the pinnacle of human achievement and is, accordingly, more challenging and valuable. The mechanisms that lead from one level to the next are not articulated, but the assumption is that data are plentiful and form a stable basis for the pyramid, while wisdom is rare. In this model, knowledge is derived from data through inductive reasoning; and data stand as they are, no matter what subsequent interpretation may be attributed to them.

We have shown how knowledge can be much more dynamic that this pyramid model allows: one person's wisdom may be another person's data. Think of a historian studying letters, diaries or reports. While the diaries contained wisdom for the writers, the diaries have the status of data for the historian. Furthermore, we might ask whether there are other components of knowledge and wisdom than such data alone - for example, experiences and relationships. The pyramid thus assumes a representational view of data, in which it is difficult to recognise the key role played by the history of data as source of knowledge. To those who interpret data as an objective and

unchangeable representation of the world, the conditions in which it is manipulated and ordered mean little: what is important is the revelation of its real meaning. This approach matches the idea that putting a lot of data together equates to an automatic increase in the empirical foundations of knowledge. The accumulation of data means the accumulation of a lot of facts; a treasure chest from which to draw new discoveries, via inductive and statistical techniques. It is easy to see how anyone espousing this view is easy prey to the fast promises of Big Data, such as the idea that it is universally reliable, impartial and usable in any type of analysis.

In the relational view, instead, the derivation of knowledge requires that objects selected to act as data (and therefore their physical features) are positioned in relation to other key interpretative features. Aspects that matter include the objective of research, the conceptual foundations and the type of knowledge – theoretical or practical – that is being sought out. This positioning requires deliberate choices and other selections on many fronts. It is not simply a question of which statistical method to apply. The procedures with which data is processed and ordered are critical to its use as a source of knowledge. The relational view of data therefore acknowledges the huge exertion required to document data journeys and makes it possible to scrutinise these during interpretation.

Data therefore do not define evidence, but the other way around: is it the fact that data can be used as evidence that makes them what they are. Similarly, data do not "contain" information: they are the materials from which meaningful information can be extracted, depending on the circumstances (Floridi, 2011). Data are best conceived as a relational and not an autonomous aspect of knowledge production. If what is taken to be data changes, other aspects of knowledge production also change – and vice-versa. The types of tools and methods we need, the questions we care to investigate, the types of researchers and users of knowledge can and do change.

How we understand data, knowledge and their relation also matters when we discuss what is good, reliable knowledge. An immediate consequence of defining data as relational objects is that there cannot be universal ways of measuring data quality and reliability. There is no underestimating the importance of methods for error detection and countering misinformation in contemporary data science, particularly in the wake of the replicability crisis (Mayo, 2018). Nevertheless, most existing approaches are tied to domain-specific estimations of what counts as quality and reliability – and for what purposes. The estimations cannot be easily transferred across fields, and sometimes even across specific cases of data use (Floridi and Illari, 2014). This is a big obstacle to the development of overarching checks for data quality and begs the question of whether producing such context-independent methods is the most useful way to tackle the problem.

Within the relational framework, the reliability of data depends first and foremost on the credibility and rigour of the processes used to produce and analyse them. The unwillingness to acknowledge the epistemic importance of data handling processes translates into an unwillingness to give these processes attention and document them so as to make them visible and open to constructive criticism. The relational view of data encourages care and attention to the history of data, highlighting their continual evolution and sometimes radical alteration as they travel. It also highlights the impact of such changes on the process of extracting knowledge from data.

## 4.6 Conclusion: Data science in a relational perspective

We can now compare the more abstract cycle of knowledge production to the model of data journeys in data science that we discussed earlier in Chapter 3 (see Figure 4.4).

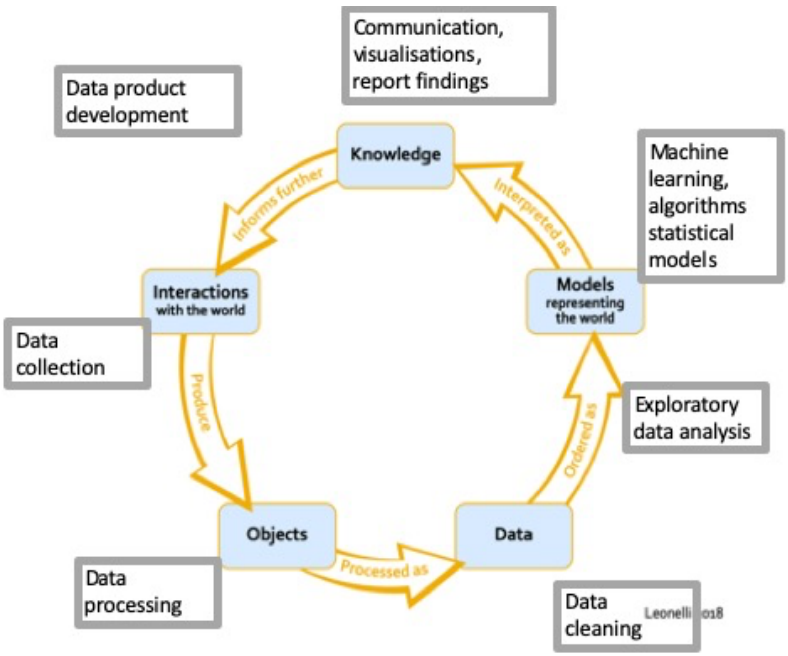



*Figure 4.4 The steps of data journeys in data science (O'Neil and Schutt, 2013) superimposed on the model of the research process. Adapted from Leonelli (2019 [EJPS]).*

This combination of the two models enables us to see how data analysis functions as a type of knowledge production. It also helps illustrate the kinds of work needed in different steps of the process of knowledge production. As we turn to the steps involved in data journeys, we will see how these different steps are structured by infrastructures and conventions (among other elements). These make it possible to put data to work. We will also explore the knowledge and skills needed to ensure the integration of these activities, so that the cycle can be completed.

We see that datafication is increasing, and that data take an increasingly prominent place in all kinds of settings. Someone who holds a representational view of data would address this by saying: we have better and better data about the world, more fine-grained, and are capturing more and more aspects of life. This will enable us to have a more rational, modern society because we can base so many more decisions on representations of the world, on real information about the world, rather than on assumptions. We have discussed the limits of these claims in relation to the Big Data mythology outlined in Chapter 2. Someone who holds a relational view of data would say of datafication: we are seeing the entire knowledge system change, including our models (for example, models of behaviour, with increasingly strong emphasis on individualistic patterns), our interactions with the world (for example creating traces and adding value to traces via platforms), and our knowledge (for example knowledge that is valued contributes to neo-liberal optimization). We are remaking the world

based on our focus on this type of evidence. When we consider the whole knowledge production cycle, we can also see how we are spending a lot of energy on how we package these objects so they can travel, since the focus is not solely on the production of objects.

Datafication happens across different stages of the knowledge production cycle, and what counts as data is affected by infrastructures, conventions, modelling approaches, curation and so on. In the next part of the book, we will explore the broader structures and activities used to make data into accessible and useful sources of evidence.

### *Additional Reading*

D'Ignazio, C. and Klein, L.F., 2020. *Data Feminism*. Cambridge, Massachusetts: The MIT Press.

Leonelli, S., 2020. Big Data and Scientific Research. *Stanford Encyclopaedia for Philosophy.*

Mayo, D.G. and Spanos, A., eds., 2009. *Error and Inference*, Cambridge: Cambridge University Press.

Radder, Hans, 2009, "The Philosophy of Scientific Experimentation: A Review", *Automated Experimentation*, 1(1): 2.

## Section III. Data Circulation

This section builds further on what we have learned about data journeys. It considers what contextual nature of data means for how data can travel and explores how data, though it is not neutral, can be re-used, combined or shared. Chapter 5 explains the practices that enable the flow of data, including infrastructures (networks and platforms), **metadata** and related conventions (standards, annotations), models and visualisation tools, and related expertise (**curation**). These all play a significant role in the multiplication of the uses and users of data. Chapter 6 describes the growing place of data science work in and outside academia, the kinds of skills needed to achieve different types of circulation, sharing or re-use of data and how these skills are interdependent. The interdisciplinary work of data science is illustrated using various examples that show the complexity of data science. Chapter 7 examines regimes and governance of data flows: why we expect that some kinds of data can flow and others not, how this is shaped by values, systems, and regulations. Processes of de-contextualisation and re-contextualisation are also discussed, by revisiting the idea of data journeys put forth in the first section, and considering how looping effects in the data cycle can be implemented. The effects of greater circulation of data and of an increasing diversity of settings in which data is used are examined. The potential of new data flows to support social change or to reinforce of existing inequalities is also set out.

This section will help you to

1. formulate data-problems and understand the different aspects of data projects
2. interact with data science specialists and to translate data issues from their own domain to data science
3. situate their own emerging expertise in relation to the broader data science job market and research landscape
4. assess and formulate the kinds of work needed for data to flow and for data to be re-used, including data handling, data curation, data visualisation and modelling, and data management
5. identify features of different types of data governance and understand their implications for the re/use of data.

## Data Story 4: Geolocation: It's a GIS world

Since 2014, more than half of the world's population lives in cities, according to the United Nations. Increasingly, the way we navigate cities involves digital information. Obvious ways in which we engage with the city digitally are by using Google maps, Uber, and TripAdvisor -- all sources of spatial digital information about cities. Many apps also systematically collect and generate location data. Think of a weather app, where having a precise way to measure location is crucial. Many other apps also collect location data, such as Instagram or Twitter, even though location is not immediately related to the primary or explicit function of these apps. As a result, much digital spatial data is generated by geotagging a huge range of digital transactions or activities. Geo-spatial data can then be combined with data from different sources, such as photos from Instagram, transactions made with a credit card, or searching for a dining recommendation. Geospatial data is linked to everyday actions of individuals, objects and processes, a process that we characterised as datafication.

- ***How can such different data sources be combined in relation to specific places****? How does this data come together in a geographical space? How does this enrich the profiles that companies are able to build and how does it provide valuable data to help predict behaviour?*

Underlying technologies of geo-location make it possible to combine, connect, correlate the data across these many sources. Location data therefore often serve as a reference point to integrate diverse datasets, since it is typically assumed that location measurements are fixed and objective regardless of the instruments that produce them and the context in which they are used.

- ***How do diverse settings and tools come to share infrastructures and standards?*** *What is the motivation to collaborate and trade data? Which actors have to work together to achieve this?*

The Global Positioning System (GPS) uses about 30 satellites and is owned by the government of the United States. It has close ties to the military and is operated by a branch of the American armed forces. The satellites transmit signals to electronic receivers on Earth, so that these receivers can determine their location relative to the satellites' positions. The GPS system connects to a reference system that links position with location, currently the *World Geodetic System* (WGS 84). This requires complex calculations based on data from the different elements in the system, and very precise calibration.

These connected systems are American-funded and US-based endeavours, have global reach and can be used free of charge. But not all countries want to depend on this American suite of standards and technologies, which is seen as imperialistic in many parts of the world as well as a threat to national security, due to the fine-grained accuracy with which it fosters the identification of specific locations. The system is sometimes embraced and sometimes resisted. The Chinese government for instance has developed its own alternative geographical system, called GCJ-02. This system is known to use an algorithm that adds random offsets to the latitude and longitude provided by the WGS-84 system. This means that a location produced within this system is

inherently imprecise, making it impossible to use the system to generate high-definition measurements. Thus, a building in central Beijing may be located in the middle of a lake, even when this is obviously impossible: in reality, the building will be located close to the lake, but the system does not allow to spot its precise location.

- ***What implications is this strategy likely to have?*** *Does this algorithm improve Chinese national security, and how? Should other countries also develop a national system for location measurement? And if many such systems emerge, which standard should be used for international communications and services?*

- ***What does this case tell us about location data?*** *In which sense are measurements of location neutral and context-independent? Is it possible to use location data without considering the differences between geo-referencing systems and their political and cultural significance? Can you imagine situations in which data workers would need to consider these differences in order to produce reliable insights?*

It is important to note that even within a single geo-referencing system, we can expect, given the material discussed in Section II, that geo-located data is not uniform and objective. With regards to geographical data, there are differences due to the economic value attached to certain locations or to the affluence of particular groups who can afford digital devices and connectivity. Some areas may be over-represented because of their importance for the map producers — Google for instance is more interested in mapping commercially-dense areas than deserted landscapes, because these spaces are more relevant to Googles' customers. There are also significant absences, even when data is not produced by corporate actors like Google. Crowd sourcing of geodata in OpenStreetMap reveals that the tagged spaces reflect the gender imbalance of its contributors. For example, the classification of bars and sex clubs is presented with more granularity than types of childcare facilities – the categories that matter to the taggers are included in the maps.

- ***Does it matter that not everyone sees the same map?*** *How do different dynamics of personalization filter and order spatial information we are presented with, and to what extent? How does this digital geo-spatial information shape how we experience space, how we navigate that space and how space is planned and organized?*

Another element to consider with regards to geolocated data is how ubiquitous some types of maps have become. For example, we now expect that we will be able to toggle between satellite and map view. This actually depends on the possibility of data integration and data flow. We also expect to see a route overlaid onto the geographical space, depending on our mode of transportation. Yet, there is nothing obvious or natural about such interfaces – these get overlooked because once we accept them as conventions, they seem transparent.

- ***Why do we take such interfaces with data for granted?*** *Is it possible and advisable to consult more than one type of map, or can we simply trust the map we use every day on our smartphones? Which skills have we learned, to be able to understand and act on these kinds of maps?*

Data story based on: Stephens, 2013; Shaw and Graham, 2017; Kitchin and Dodge, 2011; Brunton and Nissenbaum, 2015.

## Data Story 5: Tracking tuberculosis using phone data

The 20th century has seen major advances in the fight against infectious diseases. So much so that it seems that we might now be able control their spread, since they tend to be highly local. One of the first reactions to the emergence of the Covid-19 pandemic, for instance, was to set up systems for tracking the disease using mobile-phone technology. This approach builds on the processes of datafication that we have discussed earlier. It also relies on interactions between different types of data, experts and users.

Major efforts to track disease have focused on tuberculosis (TB), a complex infectious disease that is still endemic in many countries in the world (a situation made even worse by the COVID-19 outbreak). Many people can be infected with the disease without developing any symptoms, but for those who do develop symptoms, TB is deadly in about 50% of cases. TB can be treated with a 6-month course of antibiotics. The bacteria that cause TB are carried in airborne particles that are generated when a person coughs, sneezes, shouts or sings. Transmission occurs when a person inhales these particles, and the bacteria reach the lungs. Given this infection mechanism, the frequency and duration of exposure to an infected person are two major factors governing the transmission of TB. For this reason, knowing about the contacts of a patient suffering from TB is important, since these contacts may themselves have been infected.

An attempt to map the spread of tuberculosis in India was part of the activities of Big Data for Social Good Initiative, a public-private partnership launched in 2017 aiming to contribute to the Sustainable Development Goals of the United Nations. The government of India was already involved in different public health initiatives as part of Be He@lthy, Be Mobile (BHBM), using SMS in different campaigns. But the BHBM consortium thought that mobile phones could be used in a more innovative way. In collaboration with mobile phone operators, the World Health Organisation, the International Telecommunications Union and Airtel pursued a pilot project in the Indian states of Uttar Pradesh and Gujarat, India being one of the countries most affected, with one quarter of the total number of deaths globally (WHO, 2019).

This project used mobile phone network data in combination with publicly available data about incidence rates of TB for different areas (the incidence rate is the measure of the frequency with which a disease occurs – how many people get sick with TB every year). The incidence rate of TB per region was combined with movement patterns of about 40 million mobile phone users as derived from mobile phone network data. The analysis showed that when an area with few cases of TB had high mobility (for example, through people commuting between home and workplace) to areas with many cases of TB, the low incidence area was at risk of increasing TB levels or could already be under-reporting TB cases. Understanding these patterns could make it possible to act, for

example by implementing vaccination programmes and awareness campaigns or by deploying additional clinics in the affected area.

- ***What do data workers employed in such a project need to know?*** *Is it relevant to know something about TB in order to provide reliable results? What kinds of expertises are required to make some sense of these data?*

In order to understand how mobile phone network data could reveal something about movement of users, the analysts need a good understanding of how mobile phones connect to antennas/transmission towers, of the granularity of this data and its variations in space (in some areas, there are fewer towers for example), and of the kinds of noise that could be expected and would need to be removed from the data set. Hence it could be argued that analysts need to know the characteristics of the areas under study, or work with people who do (such as urban planners or social scientists).

To know which kinds of data analysis would be relevant for understanding the spread of TB, medical expertise is also needed. With TB, repeated exposure is an important factor, which is why regular patterns in movement, like commuting, were taken into account, rather than the absolute number of movements. Known risk factors for TB, such as overcrowded housing, medical malpractice and poor awareness of the symptoms among patients, also need to be taken into account *(Dye, 2014).* Last but not least, public health and knowledge about local healthcare requirements is also necessary. For example, if health care is organized at the municipal level, state-level data will not correspond to a logical level of action for those who make decisions about healthcare. In the TB project, these many capabilities were brought together, in a process of joint learning and iteration.

- ***How can such diverse sources of expertise be integrated in such a project?*** *What do experts need to share, and in which forms and venues? Is it enough to share data? Do experts need to explain to others why particular parameters for analysis are important? Does explaining slow things down, and is this a problem? How can experts appreciate the requirements and skills brought to the table by others? How can they tell the difference between an important requirement and a short-sighted demand from someone who has not understood the problem?*

A final consideration is the kind of regime of which this data was part. The project was a public-private partnership.

- ***Who are the actors involved and what is their stake?*** *Do the relations between the Indian government and Airtel matter for the results of the project? Who is dependent on whom? If Airtel considers that the data set contains commercially sensitive information and does not wish this data to circulate, how can the findings be verified by other scientists? And why shouldn't others benefit, if the WHO and the Indian government, both public organisations, have made major investments in this data set?*

We can also consider what happens in the aftermath of such a project. This seems to have been a one-off intervention. It is conceivable that it might have longer term consequences:

- ***What can we infer from such a project?*** *What if the local health authorities start to organise their work according to the findings based on the algorithms developed in this project, using it to plan where mobile clinics will be set up? Is Airtel obligated to keep providing data? Does state healthcare become dependent on data philanthropy from corporations?*

Such a project might also be pursued at different scales. For example, going from large scale to more targeted identification of TB transmission within small geographical areas.

- ***Who is targeted?*** *What if this project were further developed, and shifted from identifying patterns in the population to identifying individuals? What would happen if rather than using anonymised data at a large scale to find general patterns in the population, mobile phone data were used to track individuals (as was suggested in the early stages of the Covid19 pandemic)?*

Such developments would imply linking phone data to one's health status. Such data is usually considered to be highly personal and private, and access to this information tends to be highly regulated.

- ***Who should be allowed to access this data, and under which circumstances?*** *Would we expect personal data, such as being diagnosed with TB to be available to actors like Airtel, so that this data can be combined with movements as derived from mobile phone data? Would it be acceptable for Airtel to work with such data?*

If these data analyses could lead to monitoring movements at an individual level, the quality of the data would matter a lot.

- ***What could be the consequences of errors in the data?*** *How could the project team ensure that the data is of sufficient quality, so that individuals would not be unduly penalised due to errors in mobile phone data? And could the authorities demand that citizens use an Airtel service, in order to monitor their movements and maintain public health?*

Data Story based on GSMA, 2018; Fleming et al., 2017, Beaulieu, 2021; Dye, 2014.

# Chapter 5 Putting data to work

## Summary
This chapter identifies and discusses technological, infrastructural and practical elements needed to put data to work, and particularly the "messy" data produced by human digital activities. We describe networks and platforms used as infrastructures to store, structure and share data; standards, conventions and metadata; models and visualizations, including infographics and other specific ways of clustering data to extract meaning. And, last but not least, we also explain the curation practices through which data is maintained and cared for, and without which it would not be accessible and usable as evidence.

## 5.1 Introduction: The complexity of putting data to work
Since the second decade of this century, data has been at the centre of promises to revolutionize all kinds of sectors and to provide huge benefits to society. These promises have focused on Big Data and often involve scenarios of merging data flows. The hope is that, starting from globalised data collections, we can combine data, mine them to find patterns that would otherwise go undetected and use these patterns to deliver better services, support development, create new kinds of businesses and activities. The more areas of life become infused with data through datafication, the more potential there seems to be for data to play an important role. The analysis of social data gathered from large internet platforms, social services and more traditional industries is widely expected to inform evidence-based policy, business strategies and education, possibly even replacing traditional data production in social sciences. All these changes also highlight the importance of data science as the field where statistics, maths and AI can be applied to such data riches—a question we will explore in the next chapter.

Clearly, such processes involve not only data, but entire suites of methods and technologies. Most notably since the advent of portable computers and devices, data collection and archives have grown dependent on digital technologies. They are deeply interconnected with the rise of computational modelling and simulations. These technologies include hardware, for example, the use of multiple computers to store data and solve problems in a distributed way. Methods from artificial intelligence are also needed to process the data input, such as machine learning and deep neural networks. Further tools for **data visualisation** are indispensable. There are also technologies that we can qualify as social, in the sense that they function mainly based on practices and conventions rather than on physical properties of machines (Derksen and Beaulieu, 2011). All these technologies function together and constitute potentially powerful assemblages that make big data work towards goals such as personalised medicine and precision agriculture. These also involve the dedicated work of many kinds of experts who have developed new ways of working, repairing and caring for data. Many institutions have adapted their organisation and work flows to ensure that the work needed to make data available and useful can get done. As a society, and particularly after being forced to rely on digital technologies during the coronavirus pandemic, we have also learned to engage with data, becoming adept at decoding data visualisations, infographics or other interfaces for every day or specialised purposes. In this chapter, we identify key elements needed to put data to work – technological, infrastructural, and practical.

## 5.2 The challenge of "messy" data

To understand how data can be put to work, we first consider two different dynamics that shape contemporary digital data. On the one hand, global systems have been painstakingly developed, often based on over a century of analogue data collection and involving complex negotiations among actors around the world. Typical of such global systems are the data collecting and processing practices of the World Health Organisation (WHO, 2019), which monitors specific diseases around the world based on data generated according to specified protocols that yield standardised data, and other similar international agencies. National organisations such as statistics offices are also important actors in creating data and making it widely available for administration or policy purposes. On the other hand, there are more recent developments, where data is the result of very diverse activities –transactions, interactions, communication –rather than the result of deliberate observations or measurements. In this second dynamic, data production and collection are much less regimented. "Messy data" abound as the digital traces of online human activity, and it is their size and the possibility of combining them that makes them valuable. We briefly consider these two dynamics, before moving on the analysis of how data is put to work.

The first way in which data has become increasingly prominent is through the growth of global systems. In many areas, standardising and monitoring systems of measurement and data collection is now a priority, and the power of institutions tasked with these goals increased accordingly. Climate scientists have developed sophisticated ways to use legacy data to reconstruct a history of the atmosphere at the global level, including models to bring together climate and weather data – and this effort in turn fostered further efforts towards the pooling of international data, culminating in the 1992 establishment of the Global Climate Observing System (Edwards, 2010). In biology, the quest to map biodiversity moved to the molecular level with the start of big sequencing projects, first in model organisms such as the worm Caenorhabditis elegans, then through the Human Genome Project (Hilgartner, 2017). Sequencing databases were re-imagined as environments for in silico discovery that would facilitate immediate, low-cost data sharing, visualisation and analysis via the internet, thus helping to transform the massive investment in genomic data production into useful biological and medical knowledge. Many other areas have developed such systems of standardised data collection, for example with regards to biodiversity, pandemics or to support progress on the United Nations' Sustainable Development Goals (Beaulieu, 2021).

The second dynamic seems at first much less ordered and may appear as the result of organic growth – rather than concerted effort. Here, we can think of using Twitter data to assess 'public sentiment' or predicting likelihood that an employee will start looking for a new job through patterns of activity on a site like LinkedIn. Data gathering can therefore seem to be much less organised in these contexts, and to be more valid because it is generated 'spontaneously', rather than through the deliberate design of a scientific experiment or fieldtrip. There are many cases in contemporary society where data that was generated for no explicit purpose can end up being used for seemingly concrete ends, including policy decisions. For example, data from FourSquare (where users check in to a particular location) in combination with Twitter data has been used to quantify the degrees of diversity and homogeneity of neighbourhoods in London. The

data was used to make claims about whether some neighbourhoods had a lot of visitors (high social diversity) or were populated by a fairly constant group of people (low social diversity). These 'social metrics' were correlated with other indicators for wellbeing. The researchers found that signs of gentrification, such as rising housing prices and lower crime rates, were the strongest in deprived areas with high social diversity (Hristova et al., 2016). This innovative use of social media to address classic questions in urban development and policy is exemplary for how messy, seemingly spontaneous data is put to work, to contribute to knowledge and policy.

At the same time, both dynamics – whether they seem scientifically ordered or spontaneous and messy – rely on a set of practices that are essential to putting data to work. These structuring practices involve:

- a high degree of automation in data collection;
- increased data storage, processing and analysis;
- the production of copious amounts of metadata that document the provenance of the data and can themselves serve as useful data depending on future uses;
- and the mobility and interoperability fostered by the development of data infrastructures and related analytic tools.

While many of these developments feel intimately tied to the digital, they also have roots in the histories of bureaucracy, standardization, and engineering. As we zoom in on these various elements, we will note why these roots matter and how datasets need to be constantly maintained and repaired by humans. The digital sphere is both dynamic and multiple and can feel endless in its capabilities and potential. Yet the digital sphere is structured and constrained by series of tools, conventions and practices that support particular kinds of interactions and values. These elements make it possible to coordinate activities, whether scientific, social or commercial, across global regions. These structuring elements are largely invisible, yet they have become central to global capitalism and to cultural forms – effectively defining life in the 'information age' (Castells, 1996).

Why are structuring elements important? To travel by car, we need an infrastructure that takes the shape of a network of roads and highways. We need arrangements for fuelling our vehicles. We also need stop signs and traffic lights, and these function by virtue of a series of conventions (red means stop) that can be enforced (driving above the speed limit can lead to being stopped by a police patrol car). Furthermore, we need to be able to find our way and interact with the road network—think about maps and GPS navigation systems. Roads enable us to go places by car; they also direct us towards particular destinations, since highways make journeys more likely—and discourage us from visiting certain places by car, if no suitable road leads there. Data circulation and use relies on similar, layered suites of technologies, and these suites of technologies orient us to particular practices. If we understand how data infrastructures, conventions, interfaces and curation are organised, then we can better understand why data is used in specific ways and not others. In this chapter, we zoom in on different structuring elements and on the practices that make it possible to put data to work.

## 5.2 Infrastructures

In this section, we consider series of infrastructures that enable us to put data to work. Typically, infrastructures are reliable. They are everywhere and tend to remain invisible until they break down – then we sorely miss them and realise how important they are in structuring our activities. For example, at our place of work, which is the university campus, we are so used to using computers for research, teaching, communicating or studying that when there is a power outage, we are not able to get much done. Or even when the internet is down at work, we find that our access to our documents, sources, data and colleagues also breaks down because cloud storage creates a dependency on this network. Networks and platforms are two key forms of infrastructure for data work.

### *5.2.1 Networks*

In the data story on fighting TB, the very possibility of following mobile phones depends on the presence of a network of towers that provide signal to the phones and therefore track the devices in order to connect them. How did we get to such a networked situation?

**Networks** that can be used to transmit digital content have become ubiquitous, linking billions of devices that are built to produce, store, transmit and handle digital traces that can become data. These networks have their origin in organisations with a public mission (government, military, research) but have become increasingly corporate following the privatisation of the internet in the early 1990s. The early stages of the commercial internet were shaped by the practices and structures of telecommunications, so that telecom providers and system operators often took on the provision of internet services.

Electronic networks make greater circulation possible, which means that the site of creation of digital traces may be far removed from the site of use of data. Traces created in a certain time and place can end up being used in a completely different location and in a different context. Networks should be understood at the elements that enable connectivity: they are not 'simply' connected wires, but also include the technologies that enable wireless connections. Such "networked ICTs" are a combination of hardware and code (Postigo and O'Donnell, 2016). As we enter the second decade of the 21st century, an estimated 3 out of 7 billion people in the world have access to ubiquitous computing (Zuboff, 2019). Circulation and connectivity rely on an infrastructure that makes it possible to sense, record, transmit and process data.

The development of networks is not only pushed by Big Tech. There is also a 'demand' side to the dynamic of growing networks. There are strong association between connectivity and self-realisation, as well as between networks and development. Many activities that contribute to our identity and sense of self involve the use of networks—whether to maintain social connections or to pursue our passions. When understanding our health, staying in touch with our colleagues or finding true love depend on logging on to internet, technology and our daily practices are entwined. This connection between who we are and networks makes them unmissable. Furthermore, many of our more collective aspirations, for example, for sustainable development or decreasing

inequality, have become associated with the roll out of networks: providing access to internet is increasingly considered a pillar of regional economic development or of access to services. Initiatives to combat the 'digital divide' or campaigns that posit internet access as a fundamental right are evidence of the association between networks and the ways of life to which we aspire.

### *5.2.2 Platforms*

Networks are one of the enabling conditions for the rise of **platforms**, the second main aspect of infrastructure we will discuss. Networks and their effect are what enable platforms like Facebook to grow so rapidly (Srnicek, 2016). Platforms are reliant on the connectivity of potential users. This explains why companies like Facebook are eager to provide access to the internet. This can mean supporting technological innovations that would enhance connectivity and circulation, such as balloons carrying internet capacity, or investment strategies, such as Google outbidding other providers for free wifi of Starbucks locations, or corporations pursuing laptop philanthropy.

In Chapter 1, we noted the importance of platforms for datafication. In this section, we focus on the role of platforms in structuring data work. In a first instance, platforms can be seen as enabling interaction and transaction –for example, in the case of FoodDrop or Deliveroo, the platform connects restaurants and customers, enabling one party to advertise its menus, and hungry people to browse through a range of options. The platform facilitates the payment and coordination of orders. As such, it acts like a marketplace. However, this is only part of what platforms do. We will look in more detail about the kinds of marketplaces that platforms shape and at platforms as business models in Chapter 8. For now, we can focus on the other activities that are enabled by platforms and what make platforms a distinctive form of infrastructure.

Platforms connect different actors, provide access to data through application programming interfaces (**APIs**), and foster further development of functionalities. APIs are a kind of interface that makes it simpler to connect applications to existing platforms. APIs enable data to move between different software in a coordinated matter. One way to think of it is as a 'socket' in which to plug in. The connection is made, without having to worry about how the entire house is wired and how electricity is provided. APIs therefore invite connection to the platform, but also shape how this connection can be made (again, think of a socket requiring a specific shape of plug).

Platforms can be defined as a programmable infrastructure upon which other software can be developed and run (Gillespie, 2017). Platforms are distinctive because they are generative and open-ended. This means that they support the development of further possibilities for interactions by providing data, and that they benefit from the increased data generation from increased interactions. Platforms have two important functions: they are the site of data generation and combination (datafication) *and* support the development of applications (programmability). An example of this is the way Facebook can interface with your email account to find 'friends' on Facebook. Facebook has a wealth of data about all kinds of users and builds complex profiles from each account (datafication). The programme makes it possible to search email addresses, parse them and attempt to associate them with profiles and specific accounts on Facebook (programmability). If you use this programme, you are further contributing to the

process of datafication, enabling the platform to make even more connections and enrich profiles, while you re-create your social network on this platform (this is what we mean by generative and open-ended).

Typical of platforms is therefore that they welcome certain types of interactions to further develop their services. They are meant to enable others to innovate and create new services and new connections. At the same time, the further production of digital traces is integral to their set up: an app that only provides a service is not as interesting as an app that also engages users to produce more digital traces. This combination of providing a site that supports innovation and creativity and while tying developers and users to the platform and ensuring that further data is generated ON that platform means that the platform organizes labour and captures data as a form of input for profit for the platform-makers (Plantin et al., 2018; Gillespie, 2010).

Social media platforms are familiar instances. On these platforms, there is significant generation of data, because the platform facilitates the production of content by users (photos, comments, messages), the systematic collection of traces (timing of interactions, downloads, speed of typing, etc), and the extraction of interactions (liking, sharing). Typically, new services are developed on a platform, enabling the delivery of further services or functionalities (for which users may want to pay or that may serve as a support for advertising). These in turn engage users to increase their activities on the platform and to generate increasing amounts of data of different types. Through the combination of different types of datasets, further analyses can be done.

Platforms also organize user participation, and therefore everyday life. Platforms make participation possible along specific types of activity, such as sharing, following, or tagging, and therefore shape user interaction. (Alaimo and Kallinikos, 2017). Participation and interaction generate traces of user activity that can be analysed. These analyses yield profiles or support predictions of user behaviour – or even nudging of user behaviour. Many media outlets have recently featured discussions of 'bubbles' and echo-chambers, which are specific instances of how platforms structure exposure to news. Platforms foreground some information and background other content in ways that are not transparent to users and that may shape public life in unaccountable ways.

Platforms are also open-ended. But this openness is not limitless. Which boundaries affect platforms and how they function? On the one hand, it is difficult to predict how users will precisely engage with the platform and which kinds of new uses will emerge from the efforts of developers using APIs. On the other hand, there is also a kind of lock-in that ties the users, developers and data to the platform they are using, thereby ensuring that the newly generated data flows through the platform and can be harvested by the platform owners.

APIs can be seen as ways to align content developers and platforms: they constrain what actors can do. A typical example is the growing use of authentication via a Facebook or Google profile for other services. By connecting to new services using these profiles, the data you generate in a new setting is tied to your Facebook or Google identity. Each time a Facebook identity is used for a new service or application, these "silently" contribute to the Facebook social graph via the API, extracting data from your shopping habits or information-seeking behaviour and sending it back to Facebook.

Facebook is then able to use these data to personalise advertising and newsfeed or otherwise customize your experience of the platform (Plantin et al., 2018). Similarly, when we speak of 'mashups', we are dealing with the possibilities for data integration provided by an API. Plantin et al., put forth the example of Google Maps, where an API was released very early on after the launch in 2005. The API enabled third parties to add or overlay data onto the Google map. These are effectively 'mash ups' that take Google maps as a base and transform Google Maps into a platform (Plantin et al., 2018).

Platforms often claim to be neutral (Gillespie, 2010), but if we analyse them as infrastructures, we see how they shape digital spaces and their users (participation) as well as who can benefit from them by becoming the owner of data and monetizing their value. Platforms also have requirements about who can join and on what basis. These requirements become visible when there are debates about 'fake accounts' on Twitter, or on the exclusion of 'bots' from social media platforms, a limitation that may be harming research projects.

Finally, platforms are tied to business models and specific ways of making money. Quoting Hal Varian, Google's long-time chief economist, who speaks of "data extraction and analysis" as the core of 'Big Data', Zuboff identifies practices that turn the harvested data into input for the design of modes of prediction that benefit commercial interests. Most of this harvested data is neither personal data (like our address or date of birth) nor explicitly generated by users (actively clicking a 'like' button). Rather, it is data about time of use, length of sessions, hovering of a mouse over particular items and other automatically generated micro-data. On the basis of this seemingly worthless data, new predictive tools are developed that form the basis of surveillance capitalism (Zuboff, 2019) or platform capitalism (Srnicek, 2016). This is what we mean when we say that platforms are generative: more interaction means more data, which further feeds the growth of platforms' profit and influence. Generation of data for combination for added value is the logic around which these platforms are built – an issue we will return to in Chapter 8.

We have singled out networks and platforms for their structuring roles in providing indispensable arrangements of tools and programmes that enable data to be put to work. Networks make the access to platforms possible, and platforms make it possible to access new domains of human behaviour as people increasingly mediate their lives, and to shape the public sphere as well as how we navigate our word. While this section focused on infrastructure because it is often less visible or noticeable, we should not overlook the fact that many other technologies, big and small, are also significant – whether hardware, from giant server farms to tiny wearables like running shoe pods or software like Haddoop's data management system or apps for smart phones.

## 5.3 Conventions and metadata

So far, we have stressed repeatedly how data does not stand on its own but is part of a rich and layered context of production, transformation and use. For data to come into existence, to be transformed and to be analysed in order to support knowledge claims, a lot needs to be in place and much work has to be done. We have also amply illustrated in the data story so far that digital data are transformed across suites of technologies

(Shove et al., 2007), some of which have an infrastructural character. In this section, we zoom in on another aspect of what is needed to put data to work. Part of this work is facilitated by the use of conventions and standards. These help to organise data; to collect and handle metadata deemed necessary to make sense of that data, and they make it easier to combine and compare data. In our discussion of geographical data, we noted that the GPS system connects to a reference system, the World Geodetic System (WGS 84), which links position with location. A convention such as this connection to a reference system makes it possible to layer data. Very concretely, the possibility to switch from map to satellite view in Google maps depends on the conventions about which reference system to use and how to use it. Such conventions are used for all data types, to make scientific research possible with brain scans (Beaulieu, 2002) or for entertainment purposes, like playing Pokemon GO.

Conventions are labels that cover many types of agreements. Conventions can take the form of a protocol (an agreement on process, which steps to follow) or standards (an agreement on the measure of quality, values, or format). They can be about data formats, file systems, metadata, or about using object identifiers, in which case, they are important for automation and for large-scale data collection. Conventions can also vary in their degree of formality. They can be very detailed and internationally implemented, for example, an ISO (International Organization for Standardization) mandated standard. They can also be very informal and local, for example how to annotate data in a spreadsheet used by a few colleagues who collaborate intensively. All these conventions make it easier to link data across different data sets and to give confidence in the combination of data from different sources.

While this may sound very technical and bureaucratic, conventions also have an important role in shaping the value of data and how confident we are in using them. If there is a strong set of conventions around a type of data, and if these are implemented in similar ways across different locations where data are produced, then we are more confident that the data are comparable and can be sensibly aggregated. An example of this are drug testing protocols: we want to be sure that the data are produced and handled in the same way across the different test locations, so that we can aggregate the data and have a sense of whether the drugs have a positive effect. This is especially important in a context where we want to avoid unnecessary risks because human health is involved, and where we want independent confirmation of effectiveness because a lot of money is at stake for the companies that claim to produce effective drugs. Conventions help us navigate such situations. They contribute to ‘quality’ and trust, because they assist in maintaining data integrity, establishing provenance, and preserving privacy. They can also help us discover whether data has been modified, or whether data has been removed from datasets and might lead to a skewed view, for example.

**Metadata** is a specific category of conventions. Metadata are structured data that describe datasets or documents. It helps to make sense of their contents, of how they might be related and of their history. Metadata is therefore a type of information that helps understand how data is structured and that makes it easier to use or manage data. Metadata can be about who is the owner of the data, when and how it was collected and by which means. An everyday example is a barcode. Another example of metadata is a DOI, a digital object identifier, which is linked to a publication. When you use an

application like Zotero to save a reference from a web page, Zotero interacts with the metadata of the webpage to extract the bibliographic reference and import it into the Zotero database. As in the case of data, whether something is metadata depends on the use made of it. A library coding system can be considered metadata, but for a historian writing a history of classification systems, the coding system would be data that is informative about the kinds of classifications – for example, when did 'young adult' literature as a separate category become common. Metadata can also change over time (Li and Sugimoto, 2017).

As data circulates and data-sharing intensifies, it can be critical to keep track of where data originated and how it was produced. The use of metadata to document provenance is a common strategy, not only in biomedical databases (van Hoorn and Toga, 2009), but also in data-sharing platforms across life and social sciences (Dormans and Kok, 2010) and in social and cultural production (Beaulieu et al., 2013; Beaulieu and Rijcke, 2014). Scientific metadata helps to create common ground between different users in different institutions or disciplines (Edwards et al., 2011). There are of course many ways of describing a dataset, and to order these descriptions, 'metadata schemas', 'ontologies' or other types of semantic tools are used. Metadata schemes are often developed for specific types of objects. For digital documents, a widely used schema is the Dublin Core. For websites, the Semantic Web aims to formalize what we know about their contents. For research data, a useful standard has been set by the OBO Foundry, which provides principles around which various computational ontologies pertaining to different scientific domains can be built. As we discussed in chapter 3 with regards to Call Data Records, much metadata about phone calls can be very useful to find out about people, even if the content of the conversation (what you could arguably call the data) is not known.

In Big Data settings, metadata is often referred to as annotations or descriptors. When data from different sources and in different formats is stored together in a repository without a predefined schema (as a way to avoid data silos which occur when you store data in data warehouses), metadata is used as a way to structure data. In such a 'data lake', data remains usable thanks to metadata, since it makes it possible to query data. Again, dealing with the size of metadata is a challenge, since it must remain accessible to users when data is accessed. While we might think of metadata as 'labelling' of data, it is also an essential layer of any information system and that it plays an important role in whether it is even possible to use data. When appropriate and accessible to human or machine processing, metadata enables interaction between objects, such as data, and activities such as discovery, retrieval, provenance tracking or calculation (Greenberg, 2017).

If we think back to the GIS systems discussed earlier, we now see that metadata is needed to make the data from these systems usable. Data about location can only be used reliably in conjunction with metadata about the satellite orbit parameters and metadata about the spatial resolution available. The central role of metadata points to further issues. Because much of spatial data is used 'in real time', it is critical to have rapid retrieval and access of both data and metadata – complex puzzles for the efficient handling of data. When some transformations of data are black-boxed, it is not possible to document those transformations in the form of metadata, which also affects the accountability and therefore trust we have in data. Handling and producing metadata

are therefore important issues that are entwined with data quality, computing challenges and even the use of algorithms.

## 5.4 Models

In Chapter 4, we hinted at the central role played by models in informing the use and interpretation of data. A **data model** is the result of the effort made to structure data, so that they can be analysed (using statistics for example). A data model is often represented in graphical form (see Figure 10). A data model is a necessary step in data analysis: it involves making decisions about how to order and visualize data. This step teaches us about what is being modelled, because the data are selected and ordered in order to represent a specific phenomenon (Leonelli, 2019). In the example Figure 5.1, for instance, the data model shows which categories and features are considered important for human resources processes. There is room in this data model for elements like education and salary, but hobbies or religion are not part of the data model – they are not considered relevant for what we are modelling. The data model has thus identified and restricted the part of reality that data can be used to document.

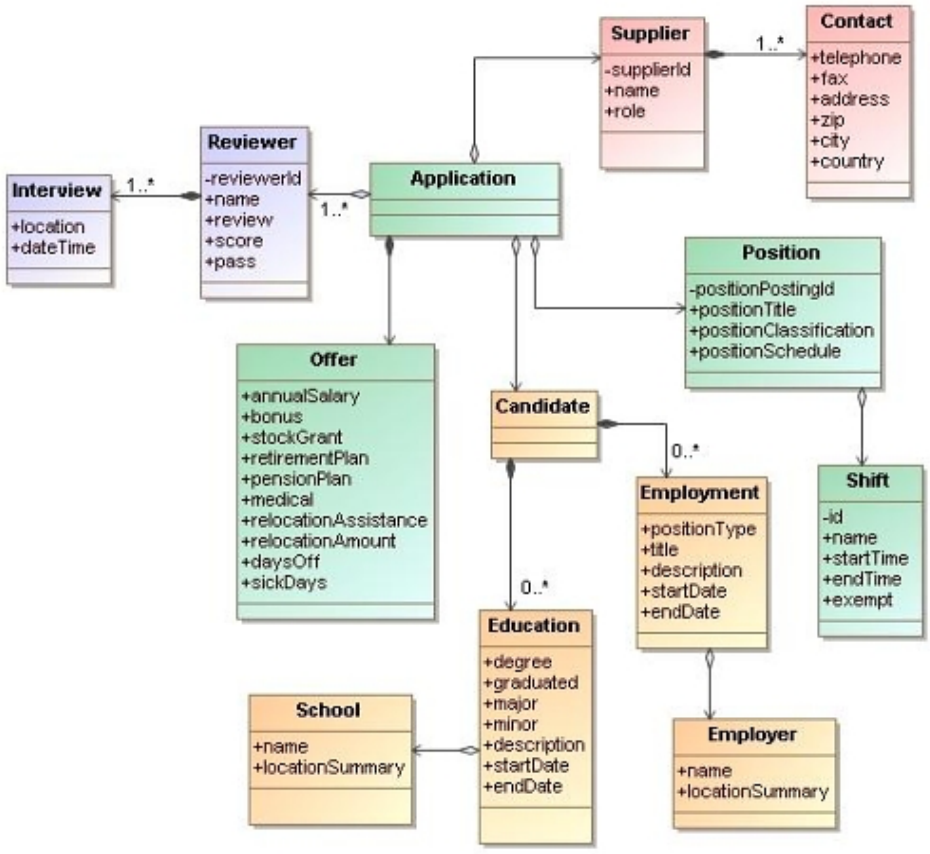



*Figure 5.1 A data model for data on hiring in a private company. https://www.itpedia.nl/2011/01/17/organisatie-rond-modelbeheer/*

This is a fairly simple example, but the same logic applies for different types of data being modelled. The data model makes some aspects of the data count and not others.

In deciding what counts as useable data, researchers define what data can matter and how. The decisions made in setting up the data model shape the range of phenomena that they will be able to consider once they start clustering and ordering data in ways that may help to interpret them as evidence.

Data models help us to capture specific aspects of the world and to pursue analyses. If we do not have hobbies or religion as part of our data model about HR processes, we will not be able to include those aspects in our claims about efficiency of recruitment or about how careers develop. That is the reason why data modelling involves explicit discussions of the value of data as evidence for making representational claims. Data models are where evidential and representational considerations meet. When developing data models, analysts consider – implicitly or explicitly – how ordering the data will make it usable as evidence.

A second kind of model is very important if we want to put data to work: **statistical models**. While data models are usually expressed in graphical form, statistical models are expressed using mathematical notation. Statistical models enable us to describe the data in ways that make it manageable: they are precise and concise descriptions that enable calculations. Statistical models help us make claims because they encompass important aspects of the relationship between variables, and between the data we collected and the world. They guide the methods used to extrapolate patterns from data. They help us evaluate whether or not such patterns are meaningful, and what "meaning" may involve in the first place. Statistical models are therefore central to the credibility of the data-driven approach.

Statistics can help us ask questions like does the data really show an important effect or does data really correspond to a (consistent) phenomenon in the world. In other words, we use statistics to explore the validity and reliability of patterns extracted from data. For instance, statistics is often hailed as a powerful means to detect error within datasets in relation to specific hypotheses (Mayo and Spanos, 2011). In a big data context, there is a set of statistical models and techniques that tend to be used most frequently (see discussion of correlation in section 2.2). Hence, some philosophers and data scholars have argued that "the most important and distinctive characteristic of Big Data [is] its use of statistical methods and computational means of analysis" (Symons and Alvarado, 2016). Examples of this are machine learning tools, deep neural networks and other "intelligent" practices of data handling. There is a lot of emphasis on developing computational methods for data analysis within big data research, and much of the efforts focus on improving inferential tools and methods, in order to extract reliable knowledge from data.

Last but not least, computational models and algorithms are also a central type of model to consider for understanding how data work is structured. Statistical expertise needs to be complemented by computational savvy in the training and application of algorithms associated to artificial intelligence. This includes machine learning and other mathematical procedures for operating upon data (Bringsjord and Govindarajulu, 2018). These are closely linked to the statistical models selected.

Consider for instance the problem of overfitting. This is the mistaken identification of patterns in a dataset, and the result of imposing a model on the data too rigidly.

Overfitting is greatly amplified by the training techniques employed by machine learning algorithms. There is no guarantee that an algorithm trained to successfully extrapolate patterns from a given dataset will be as successful when applied to other data. It is possible to minimize this problem by re-ordering and partitioning both data and training methods. This makes it possible to compare the application of the same algorithms to different subsets of the data. This is called "cross-validation". Another approach is to combine predictions arising from differently trained algorithms ("ensembling"). A third technique is the use of hyperparameters, which are used to constrain the learning process. To do this well requires knowledge of the mathematical operations and of their implementation in code, as well as familiarity with the hardware architecture (Lowrie, 2017). In other words, working with models from statistics and mathematics needs to be complemented by expertise in programming and computer engineering. (The need for this layered knowledge is discussed in the next chapter.)

The point is that structuring of data work in data science is different than data analysis in statistics, as developed over the past century in the social and natural sciences. Whereas regressing or rule-based deduction are used in traditional statistics, machine learning builds programmes that develop their own approach to data description (Lowrie, 2017). Focusing specifically on computational systems, John Symons and Jack Horner (2014) argued that much of big data research consists of software-intensive science rather than data-driven research. These elements structure how we can put data to work: the production of knowledge claims depends on the manipulation of models implemented in database design (data models), in analysis (statistical models) and within software and computation (machine learning and algorithms).

## 5.5 Visualisations: forms, tools and interfaces

### *5.5.1 Data visualisations*

Data visualisations are a way of ordering, encountering and interacting with data. They differ from the graphical data models discussed above (Figure 5.1) insofar as they are aimed at conveying the data rather than the properties of the data. As such, data visualisations are considered to be of much wider interest than the more specialised data models that are mainly used by data workers who deal with databases and computational work. Data visualisations circulate widely and shape what we know and the questions we ask of data. Visualisations are themselves shaped by data practices and technologies. Data visualisations are often presented and perceived as the way of letting data speak for itself, but they are neither transparent nor self-evident. Visualisation are 'acts of interpretation masquerading as presentation' (Drucker, 2014, page 16). In this section, we consider how visualisations have developed over time, the main conventions that shape how data is visualised, and how data visualisation already contain selections and interpretation of data.

We discussed earlier how data are not merely representations of phenomena. We extend this argument to data visualisations: these are not merely representations of data. Data visualisations are the result of many steps, and our appreciation of them as visual renditions depend on conventions that are often so familiar that we don't notice them. What distinguishes a data visualization from other types of expression (text,

numbers) is that they use space in a meaningful way. That is to say, the conventions for displaying data *spatially* are a central component to making data visualisations meaningful. These conventions are often forgotten, but they are both fascinating to understand and important in order to learn from visualisations.

Data visualisations can be organized according to their:

- graphical format (map, table, chart, network diagram)
- purpose or function (navigating, record keeping, calculation)
- type of content (spatial, qualitative, quantitative, temporal, interpretative)
- the way they structure meaning (analogy, connection, comparison, multi-variate, axes)
- disciplinary origins (bar diagrams from statistics, trees from genealogy, flow charts from electrical circuits) (Drucker, 2014)

Across the variations in the types of data visualisations, all have a number of processes in common (Drucker, 2014). The first is the rationalisation of a surface. This is the process of setting a space apart so that it can be meaningful, for example, the separation of the space on a page between the space for the running text and the space for a figure. The second is the distinction between figure and ground. Creating a contrast between an object and the background directs out attention and tells us what to pay attention to. A third process is to have the visual elements of a figure work together, in a relational system. This could be the framing of a visualisation or the use of a legend that enables us to put the visual element in relation to a shared reference. Together these processes create what we consider to be a visualization. Within a visualization, the organized space can then be used to express meaning. Spatial relations like proximity, hierarchy, and juxtaposition indicate how to understand data (Drucker, 2014).

All these visual conventions contribute to make data visualisations work in specific ways. They tend to be treated as transparent, whereas they are already carrying some interpretation of the data. An illustration of this are the reference systems we discussed with regards to GPS data. When we look at maps of the world, we are aware that a three-dimensional space has been brought into a flat, two-dimensional surface. This is done via projections, and different projections take different elements into account (See Figure 5.2), but we pay little attention to this convention because we are so used to it.

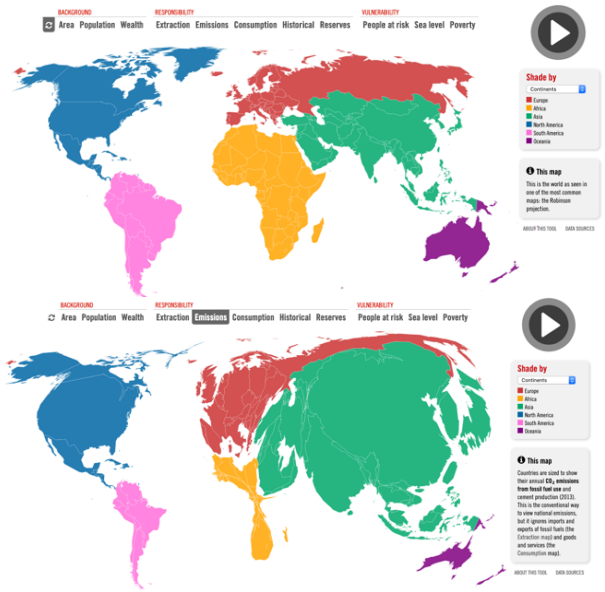



*Figure 5.2 Examples of geo- political vs emissions maps highlight how the projections make different assumptions and therefore show us the world differently.*

Naturalised maps and charts come across as 'what is' – as being the same as the phenomena. When we stop paying attention to which assumptions are taken into account, we start to see data visualizations as transparent. Consider which map in Figure 5.2 looks 'normal'? This applies to all visualisations, even something as simple as a bar chart: we tend to forget all the work that goes into producing these visualisations (sampling, smoothing, colour selection) and how they produce meaning (Drucker, 2014). In Figure 5.3 an association between maleness, genius and scientific ability is expressed in the use of similar shapes (squares). Such associations matter because visualisations shape how we experience the world. According to this chart, females could pass on these traits, but never express them. By visually reinforcing that the symbol for men is related to the symbol for genius, the visualization reinforces a sexist interpretation of the data.

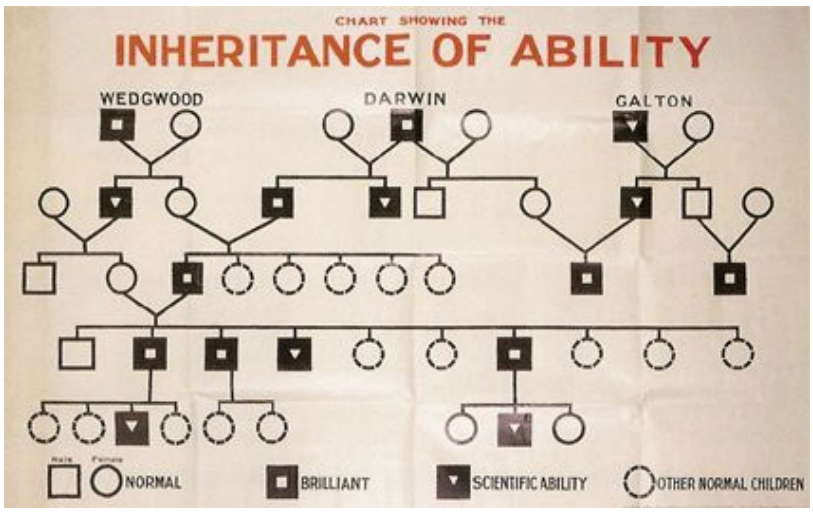


*Figure 5.3 Chart produced by Francis Galton, for the English Eugenics Society. Conventions of visualisation shape the contents: the icon of genius is visually related to the icon of maleness. Chart showing inheritance of ability, Eugenics Education Society. This poster was commissioned by the English Eugenics Education Society and designed and produced by Philip Benson's London advertising agency in 1926*

*Copyright Museum of London.*

Data visualisations are closely associated with Big Data and are a common component of data science training. That visual information constitutes the richest mode of input to human cognitive systems (so-called 'power of the human visual system') is an assumption that has a long history, going back to at least the middle of the 20th century. Each visualization contains assumptions and principles of knowledge, and a particular point of view. Many conventions from older modes of observations persist, even when we are dealing with large-scale information that are beyond human perception. Rather than windows that give us a perfect view onto data, visualisations are better understood as lenses that imperfect, but useful precisely because they are selective (Beaulieu, 2001).

### *5.5.2 Infographics*

In the past decade, the format of the infographic has become a prominent way to communicate visually with data. The term infographic is a contraction of "information graphics". Infographics vary in style, but all have in common that they are visual representations based on statistical data. Their history is sometimes taken to go back to the work of Florence Nightingale in the nineteenth century, but there has been an explosive use since about 2010, related to the availability of web-based tools to create and circulate infographics. Typical of infographics is the combinations of graphical and other elements; assemblage of text and numbers, charts, graphs or maps, and characters to transform data into visually accessible arguments (Featherstone, 2014).

Infographics can be

- statistical graphs (visualizations that rely on quantitative data without adding another layer)
- data maps (combining cartographic form, variables of distance, and visualized quantitative data)
- time series (which place quantitative data within a visual temporal context)
- and relational graphics (which utilize composition to convey differences in size between variables and analogize the properties of data sets with the physical world) (Tufte, 1983)

Across all these categories, it's important to note that digital infographics are sites of intersections between interfaces, people and data. (Amit-Danhi and Shifman, 2018). Furthermore, in a digital context, infographics are not only to be looked at and understood, but like other visualisations, infographics can act as interfaces and sustain interaction (Frosh, 2016). They can also give a more layered sense of engaging with data

(Amit-Danhi and Shifman, 2018). Like other kinds of data visualisation, they are interpretations.

### *5.5.3 Data Visualisations as interfaces*

One of the important features of recent forms of visualization is that they merge the function of engaging with data visually, with other forms of interaction. Digital visualisations have tended to merge with the artefact of the graphical user interface. This means that seeing and doing are brought together. These 'visualisations as interfaces' invite the user to zoom in, select, or otherwise interact with data (Rijcke and Beaulieu, 2011).

As users of interfaces, different actions are open to us. Depending on how the data resource and interface are built, we can view, filter, configure, select, or construct data objects. Some of these interactions are quite shallow, while others allow deeper interaction with data. Infographics tend to be at the shallow end, and the interactions they allow are quire limited. What is important is the way these interactive possibilities 'create a sense of data experience' (Amit-Dahni and Shifman, 2018). By engaging with the data itself, the viewer/user touches the evidence and is put in the position of seeming to have direct access, to be able to explore and draw unbiased conclusions. The viewer/user can make the data speak for itself – but only within the set of infrastructures, conventions and rules of visualisation we have discussed so far.

A final aspect of data visualisation to note is the increased personalisation of data visualisation. Across many apps, it is now very common to have access to data visualisations that can seem like intimate self-portraits. We see ourselves through detailed data visualisations that show how our bodies and activities can me mapped out in time and space. A daily consultation of our mundane practices has become common practice: we are used to seeing graphs of our sleep, our movements, our hoped-for path home… and to respond to cues from such visualisation for our sense of self (you run faster than 86% of users) or to script our lives (move more, drink more, stress less).

Through these visualisations, the very relationship between objects of knowledge and knowing subjects is changing. New interfaces with data are created, such as (dynamic) visualizations. These not only aim to 'show' data, but also emphasize its dynamism through live feeds, interactivity, and the possibilities to layer types of data. Visualizations also function as interfaces to data, to explore and act on it. Visualisation techniques make data seem "transparent and accessible" while the underlying models are opaque (Thayyil, 2018). As we have seen however, visualisations are far from transparent. Critical skills and awareness of the kinds of data assemblages that enable these visualizations are indispensable to using data visualizations responsibly.

## 5.6 Curation

The final element in putting data to work that we will discuss in this chapter is curation. Curation – or, as it is also sometimes called, "data management" - is perhaps the most undervalued aspect of data work. Yet, it is necessary for successful data circulation and use. Data curation is the process of organising and integrating data from different data

sets. It is guided by the desire to maintain the value of the data across the adjustments needed to make it usable. As such, data curation requires a strong understanding of both the data's creation and of its potential further uses.

To illustrate how significant the work of curation can be, think of the experience of grocery shopping in a different country. In the US and Canada, sugar is located in the baking goods section, together with flour and baking powder. In the Netherlands, sugar is located next to the coffee and tea. Sugar is therefore categorised very differently in these two contexts. In the case of a supermarket, it's possible to walk around until you have found the desired item – for a database, the labels assigned to data are the only way to navigate the database. To extend the analogy, how could you ever find the sugar if you look for it with a label of 'baked goods' in a Dutch database? Data curation is therefore fundamental to the usability and quality of a database.

As global data infrastructures, interfaces and related institutions become ever more sophisticated, the resources needed to curate them have also grown exponentially. Within the private sector, the increasing costs of storing and analysing data has heightened their value as commodities, with increasing efforts to license datasets so that they can be incorporated into specific property regimes (more on this in Chapter 8). Companies are also frequently looking to either outsource data maintenance (for instance by using external web and cloud services) or forage for easily accessible data (as in the case of social behaviours expressed in publicly available social media). The extent to which data are circulated, and the restrictions under which such circulation can be placed, can thus vary dramatically.

Within academic research, all the work required to put data to work does not fit contemporary regimes of funding, credit and communication. Monitoring data infrastructures and keeping them up to date requires serious investment, without which the quality and reliability of the Big Data used to fuel artificial intelligence tools cannot be guaranteed. As we will see in Chapter 7, the more data move around and are used for a variety of purposes, the more vulnerable they are to unwarranted and even misleading forms of manipulation and enrichment. And yet, funding agencies focus the vast majority of their resources on supporting novel research and rewarding the publication of high impact scientific papers. Long-running infrastructure do not fit this model of evaluation, since their core business is conservation rather than innovation (though of course they do require constant updates and the uptake of new technologies).

Data curation is done by a variety of experts whose training and titles can vary. They include librarians, information scientists, project managers, consultants or research assistants. Their work can address both upstream and downstream management of data, from the point of data creation to the archiving and sharing via repositories (Palmer et al., 2017). The creation and evaluation of metadata are part of data curation. While we noted that the production of metadata was sometimes automated, there are many aspects of data curation that cannot be formalized and require a high level of familiarity with data creation, competence in overseeing infrastructural aspects and expertise on the needs of users. Even when automatically produced, informal communication about metadata also helps to make data more useful (Edwards et al., 2011). Curation therefore plays an important role in shaping or packaging data and

making them intelligible to users who were not part of their creation. This work often seems technical, outside the more valued activities of doing research. Yet the use of labels to describe data (sometimes called ontologies) is vital for users to be able to retrieve the information they need from a database. A shared understanding of what these labels should be (as in the sugar example above) is very important.

The creation of metadata 'by hand', through annotation and filling in of information by humans for each data point, is highly time consuming. In some fields, there is a high level of professionalisation around metadata, but in many fields this work is considered part of the researchers' tasks. Creating metadata can feel as a burden, on top of a scientist's primary work. As Edwards et al., (2011) explain, "Research scientists' main interest, after all, is in using data, not in describing them for the benefit of invisible, unknown future users, to whom they are not accountable and from whom they receive little if any benefit."

Those who have the expertise to maintain and curate databases are often overlooked and undervalued, since they do not routinely publish in top-ranking journals and may therefore not be recognised nor rewarded as high-level researchers. This affects both the status and the availability of data curation positions at research-performing institutions. It is difficult to make a case for creating jobs in data curation, and even when they do exist, they are often ranked as "service" jobs (on a par with technicians) rather than as "research" jobs (and thus seen to contribute directly to knowledge creation), with serious consequences for the career prospects and salary scales of this type of data workers. There are therefore few incentives to enter this path of work, even though it is central to the large-scale mobilisation and re-use of data that powers contemporary science.

The care for data does not fit the current rhetoric around Big Data. It requires work and time, it is not exciting, and it is expensive. This creates the risk that data are badly managed, unreliable, unfit for repurposing – and that because they have not been curated well (for example, without significant metadata), they cannot be re-contextualised adequately.

## 5.7 Conclusion: Forms of data work

In this chapter, we have gone beyond data journeys to explore how data is put to work. We have looked at what is needed to circulate, share or re-use data. We considered putting data to work from different angles and, for the sake of clarity, discussed in turn the technological, infrastructural, communicative and labour dimensions of putting data to work. Across this discussion, we constantly pointed to how each dimension relies on and affects the others in practice. There is a great variation in the extent to which these elements are formally organised, from the highly regimented databases of the WHO to the looser mashups of start-ups using social media data. In spite of this variation, we saw that to make data work takes work. More data is not enough! The practices that enable the flow of data involve infrastructures (networks and platforms), conventions (standards, annotations), models and visualisation tools, and related expertise (curation). Together, they play a decisive role in the multiplication of the uses and users of data.

### *Additional Reading*

Drucker, J., 2014. *Graphesis: Visual Forms of Knowledge Production*. Cambridge, Massachusetts: Harvard University Press.

Acker, A., 2018. *Data Craft: The Manipulation of Social Media Metadata*. New York: Data & Society Research Institute**.**

Morrison, M., 2015,. *Reconstructing Reality: Models, Mathematics, and Simulations*, Oxford: Oxford University Press.

Starosielski, N., 2015. *The Undersea Network*. Duke University Press.

# Chapter 6 New data skills


## Summary

In the context of data circulation, the level of 'project' is central. In this chapter, we consider different kinds of data work done in projects, from the perspective of people who see themselves and are seen by others as being data workers. One label among many for these people is data scientist. In order to understand what data scientists do, we have to first discuss what is usually understood under data science. It is important to consider what data work involves to really grasp how different aspects of data work influence each other and to value them. We will then zoom in on the skills that are needed for data scientists and data workers and on how collaboration can take place across these skills sets. By discussing data work in detail, we help map out a complex field and give a conceptual basis to understand why collaboration is so important to achieving data projects successfully.


## 6.1 Introduction: Data expertise

Imagine a project that aims to understand traffic patterns in a city. What would be needed, in terms of data work, to pursue such a project? We would need to think about which data is already available and whether additional data needs to be collected. A good plan for data collection would need to be developed, to ensure that data of sufficient quality and scope is collected. Given that the data is generated by citizen's activities in their daily life, we would also need to consider how to engage citizens and how to ensure that this data collection would not be harmful to them as individuals or as groups. We would also need to figure out how to move the data from the points of collection, to storage facilities and to where the data is going to be analysed. Such data might be especially amenable to being mapped out geographically, and we would need to figure out how to best show traffic flows and bottle-necks in the space of the city. And we would need to decide on who will be able to access the data, in which form and for what purpose. A serious set of tasks for any project team!

Clearly, doing data science requires many kinds of work and different types of expertise. This expertise needs to be coordinated. This is true not only across tasks, but also to perform a single task –to decide how much data needs to be collected, you have to understand the implications for both the computational needs of the project and for how you even conceive of traffic density. The requirement for different types of expertise to work closely together is also increasing. For example, as data sets become more diverse, it becomes even more important to link expertise on the domain in which data was created to understand the meaning of the data, with computational expertise to be able to handle the diversity of data.

There are very few individuals who possess skills across all these areas. Furthermore, projects are usually too large for single individuals to take them on. This means that expertise needs to be distributed across individuals within teams, or even across teams in an organization. These experts must not only be competent in their own area, they must also be able to work together if they are to pursue data science successfully. In this chapter, we aim to provide you with greater awareness of the various types of expertise and skills needed. We also describe how experts can work together, by reviewing

different models of collaboration. This material will enable you to better understand your own expertise and skills, to value to expertise of others, and to find effective ways to combine them to work together.

## 6.2 What is data science?

### 6.2.1 A growing field

The increasing attention to data as valuable outputs in and of themselves is occasioning a shift in the division of labour within research and development. The very idea that "**data science**" can be a separate research domain is relatively new, and what data science actually consists of continues to be a matter of debate. This has concrete implications: there are very few people who have been trained as data scientists. Rather than having followed a programme with the title 'data science' in college or university, they have come from different disciplines, such as statistics of computer science, and, over time, have come to occupy a position of data scientist in their organization. This diversity matters because it means that data scientists in different organizations are actually doing very different jobs and might be in very distinct departments within an organization – for example in marketing, business intelligence or R&D. It also means that their expertise might contrast as well, depending on their initial training and the context in which they have worked since studying. This diversity has consequences for careers: in some universities, data workers are considered professional staff, whereas in others they are considered scientific staff. All this variation makes it difficult for the professional status of data workers to be recognized and limits career development or even job security. This is an odd situation, especially given that demand for data scientists far outstrips availability, a trend that has been observed since at least 2008 (Swan and Brown, 2008).

When there is a clearly and widely recognized institutional embedding for a kind of expert, this adds to the legitimacy and recognition of a particular area of expertise. One way to establish the status of a type of expertise is to link it to a scientific discipline. Is this a question of time before data science becomes a coherent, recognized discipline, as a basis for a well-defined profession? A discipline has

- A shared object of study and methods
- An accepted body of knowledge
- A community of scholars who primarily identify with the discipline
- Mechanisms for communication (publishing) and reproduction (teaching)

Whether a discipline is forming can often be traced by looking at mechanisms of communication and reproduction. In other words, whether there is a particular body of knowledge and a discourse in which a community is involved, and whether there are training programmes built around a recognizable set of core elements. With the rise of dedicated journals and the existence of hundreds of accredited training programmes, two key markers of disciplines – mechanisms for communication (publishing) and reproduction (teaching) – seem to be in place. With regards to journals, we can identify

a number of publications that are well-regarded and have data science as their core subject.[3] With regards to educational programmes, data science degrees can be found from college to PhD level, and there are now guidelines and recommendations on what a data science degree should cover (De Veaux et al., 2017; Mikroyannidis et al., 2018). In 2020, over 250 undergraduates programmes could be found at 'bricks and mortar' universities on all continents, and more degrees were offered as online courses.

Obtaining a degree in data science is not the end the story, however. Because of rapid changes in the kinds of data and in computational techniques, data scientists have to be willing to keep learning across their careers. In addition, data scientists need to be aware of the domain in which data is created – data from biology and from marketing need different knowledge to be used properly. This dynamism requires that data scientists be willing to be lifelong learners, and requires that there be support for this learning (for example, Mikroyannidis et al., 2017).

Data science should also be understood in the particular socio-economic and cultural context of the first two decades of this century. In the same period that data science has become increasingly prominent as a field or discipline, there have been important changes in universities. Universities were established around specific disciplines, and from the end of the 19th century in the Global North, faculties and departments became increasingly specialized. This focus on single disciplines has changed over the past decades. For one thing, interdisciplinary education has been increasingly valued. Many programmes in data science are indeed taught via university-wide coalitions between different faculties. These tend to be the most successful, although these initiatives encounter additional administrative overhead to deal with cross-organizational entities (Berman et al., 2018). Teaching of data science is also linked to innovations in universities, in terms of how education is organized:

"Data science is by definition interdisciplinary and requires students to interact widely across academic disciplines and with non-academic partners, since they too are making rapid progress in the field of data science. This requires a new type of education that is future-proof with respect to data science. To achieve this, we also have to adapt the education system, which needs to change from the more classical way of providing education aligned with the traditional academic disciplines (Wijmenga, 2019) "

Such new contours for how universities can function as knowledge institutions have been emerging in the past decades. First, there has been growth in transdisciplinary research involving actors outside the university, often in public-private partnership (for example, businesses, patient groups, NGOs). For data science, this means that students are often involved in projects developed in partnerships with non-academic actors, and that they are learning by engaging with "real world" problems and data sets. Second, activities of 'valorisation' have been increasingly stressed. This means that not only 'scientific discoveries' are stimulated and rewarded in universities, but that innovations, patents, start-ups and other kinds of contributions for which a 'context of application' or societal value seems obvious are rewarded. In the context of datafication of society, we can see how data science would be an area that would benefit from, and contribute to, a

[3] Data Science Journal (CODATA), International Journal of Data Science and Analytics (Springer), PJ Data Science Journal (SpringerOpen), Harvard Data Science Review

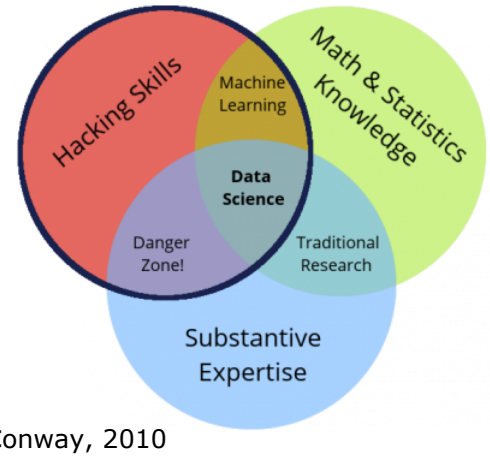
version of the university as a place that produces useful, economically and socially relevant or applicable knowledge.

A final important dynamic is the growing diversity of sites of knowledge production. Research is no longer primarily associated with universities, and knowledge has become central to a great many organisations (Wouters et al., 2013; Gibbons, 1994; Nowotny, Scott, and Gibbons, 2001; Strathern, 2005). The results of all these changes are multiple and layered, and also vary according to funding schemes in place in different national systems of education. Generally though, we can speak of pressures on universities to show that they are 'productive' and useful to society. In the current context, research and teaching priorities are not only shaped by criteria internal to universities, but are also increasingly responsive to societal challenges and corporate interests.

What does this mean for data science? The current trends in universities that prioritize the connection of science to (technological) innovation and economic value, and openness to the world outside academia may work against the internally-focused dynamics necessary for discipline formation. In other words, data science may not solidify into a recognized discipline based in universities, in the way that molecular biology or women's studies have become recognized departments in universities. In addition, there is a strong pull on experts towards industry and away from academic institutions, to the point that it prevents institutions from developing data science programmes (Berman et al., 2018). Higher salaries and short-term benefits are not the only reason for this brain drain. The recent report of the NSF working group on data science noted that an emerging problem with maintaining and developing scientific research in data science also has to do with the fact that "when the best infrastructure environment for cutting-edge research is consistently in the private sector, the opportunity for innovation in the public sector deteriorates (Berman et al., 2018)." This brief discussion of disciplines and of universities as institutions helps understand the context in which data science is developing and why we cannot consider it solely as an academic project: corporate and societal actors are also shaping the contours of data science, and vice-versa.

### 6.2.2 A composite field

If it is difficult to describe data science as an academic discipline, how should we talk about it? Describing data science as field at the intersection of different disciplines or areas of knowledge is a common approach. Venn diagrams are often used to show data science as overlapping areas of knowledge, giving a strong sense of data science as a composite field. One of the early descriptions of data science can be found in *Figure* 6.1.




Drew Conway, 2010

Figure 6.1 Data Science as composite, first presented during a talk in 2010 and later published on a blog in 2015 (Conway, 2015).

This way of representing data science seemed to resonate strongly with the data science community, and this Venn diagram led to countless variations on this theme (see Figure 6.2). The variations are interesting in and of themselves, but the main message is that data science is the result of combinations of different areas of expertise.

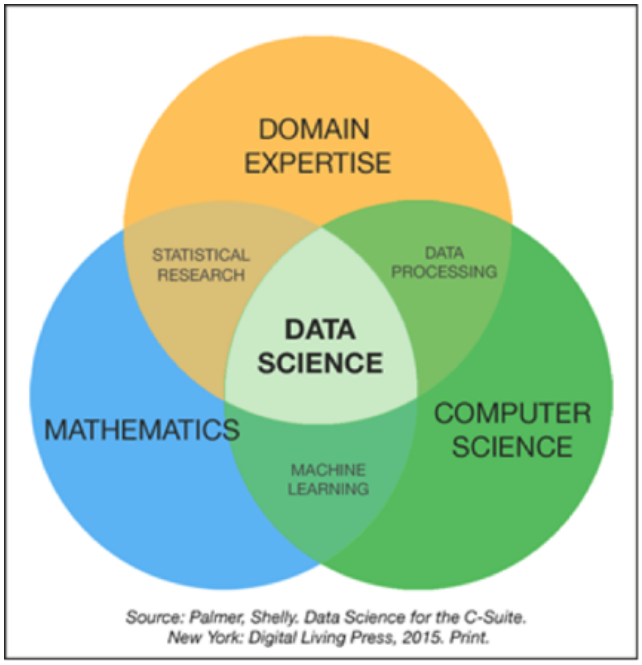



Figure 6.2 Another version of data science as a Venn diagram. This one replaced 'hacking skills'– some found this term objectionable because of its criminal connotations – with computer science and specified the resulting overlapping areas of in more helpful terms. Reproduced from (Palmer, 2015).

The types of expertise that needs to be combined are also debated. Increasingly, there is recognition that efficient and appropriate use of data is not solely the result of combining mathematics with computer science and statistics. As Xiao-Li Meng stated in his inaugural editorial in the Harvard Data Science Review in 2019, data science is "not just machine learning or just statistics" and "not all about prediction." Data science is not even "only about data analysis" (Meng, 2019), given the many stages in the data journeys necessary to make data actually usable. Data science must also address how it gleans knowledge from the world and produces data as evidence to support claims. This means that epistemology is also a core concern that affects both the daily work of data scientists and that shapes the place of data science in society. The very aims of balancing appropriate assumptions with computationally efficient approaches and other trade-offs are epistemic ones. Furthermore, at the heart of technical questions around how to integrate different data formats or how to engage participants in data collection are critical issues about data governance.

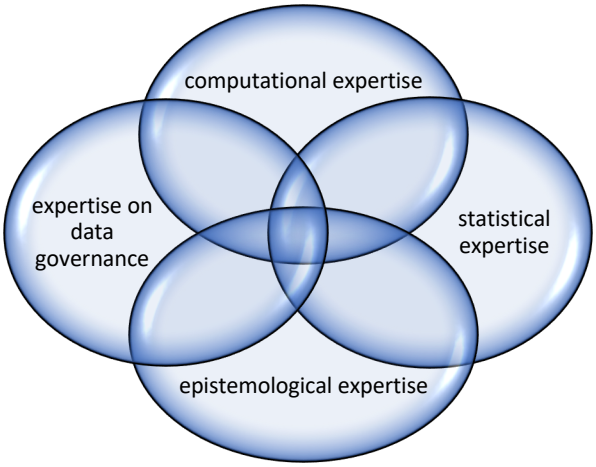



*Figure 6.3 Data work as conceptualized at one of the author's workplaces, Data Research Centre, University of Groningen, the Netherlands.*

We see data science as dealing with four basic types of expertise: computational and statistical, both of which are typically regarded as core "technical" expertise; epistemological expertise, including an understanding of where data fit in the processes of knowledge production and how different stages of a data journey may affect each other; and expertise on data governance, which encompasses the usability, regulation, ethics and curation of data. The relevant areas of expertise may change as data science transforms in response to internal and external factors. It remains that while these Venn diagrams powerfully make the case for the need for a diversity of experts to pursue data science, what really sets data science apart is the need for integration across these different spheres. It is not enough to put these (or other) types of expertise next to each other and expect that successful will ensue. These areas of expertise must truly intersect – rather than co-exist. This is what is distinctive about being a data scientist versus a being a statistician, computer programmer or legal expert. In that sense, the

proliferation of these Venn diagrams may be explained by the fact that they foreground the kinds of expertise needed, while missing out on the interactions that, we argue, are what constitute data science. For this reason, we move from a static view on the disciplinary areas that contribute to data science, to a discussion of the skills that make the dynamic process of data science possible.

## 6.3 Data science skills

Why should we emphasize the dynamism and interrelated aspects of data science? Consider this passage on developing curriculum for data science, and note how it links processes, context and experiential knowledge in learning data science:

"The recursive data cycle of obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses, and communicating the results lies at the core of the data science experience. Undergraduates need understanding of, and practice in performing, all steps of this data cycle in order to engage in substantive research questions. In the words of Google's Diane Lambert, students need the ability to "think with data" (Horton & Hardin 2015, p. 259; see also ASA 2014a and Shron 2014). Data experiences need to play a central role in all courses from the introductory course to the advanced elective/capstone. These experiences should include raw data from a variety of sources and should involve the process of cleaning, transforming, and structuring data for analysis. They should also include the topic of data provenance and how it informs the conclusions one can draw from data. Data science is necessarily highly experiential; it is a practiced art and a developed skill. Students of data science must encounter frequent project-based, real-world applications with real data to complement the foundational algorithms and models." (De Veaux et al., 2017). Similarly, the US National Science Foundation Working Group on the emergence of Data Science stressed the view of data science as a process made up of different practices and skills that could be strengthened by building better connections across the data life cycle to reduce the current gaps (Berman et al., 2018).

Doing data science is therefore about connecting the use of data for prediction, exploration, understanding, and intervention. It requires a thorough understanding of probability, of the relationship between sampling/populations, of the consequences of false positives/negatives, and a firm grasp of correlation and causation – a set of skills that should also be shared widely in society given the growing importance of data science (Garber, 2019). Doing data science also means being able to judge when to simplify or approximate, to make trade-offs to optimize algorithms, and to balance speed and accuracy. Finally, doing data science requires understanding how we shape our world and our knowledge of it by making decisions about samples and populations (Blei and Smyth, 2017; Francois, Monteiro, and Allo, 2020; Bates et al., 2020). Because of the societal significance of data science, it is also important to be able to evaluate the role of data agents who are explicitly or implicitly producing and sharing data, new risks and benefits associated with this, and how configurations of access to data shape power in society. In what follows, we review the types of skills involved in meeting these requirements.

### 6.3.1 Technical Skills

Data science involves technical skills typically associated with quantitative analysis and problem-solving. Mathematics (especially optimization and probability) and statistical inference (including sampling) and modelling remain essential, but they are not sufficient for data analysis. Equally important are computational skills such as coding, machine learning and experimental data mining. A computational approach will focus on interrogating the data in a way that an information processing machine (a computer) can handle. It will also consider whether the analysis and handling of the data will remain feasible as the data set grows. Design and visualisation skills are also highly sought-after: as we saw in Chapter 5, how data are ordered and visualised matters a great deal to their interpretation. Another set of skills has to do with database management and data curation. How to query, retrieve and handle data, as well as cleaning and structuring data are important to be able to meaningfully explore, analyse and visualise data. The combination of these different technical skills is indispensable for data science.

#### *6.3.2 Analytic skills*

Data science requires the skill to formulate productive questions, that is, questions that can be answered with particular datasets. Whereas lots of people have the skills to crunch the data and answer set questions, asking the right questions is an even more valuable skill (Dumit and Nafus, 2018). In dealing with data, a valuable "habit of mind" is to develop a critical stance towards the quality and provenance of the data. This means asking questions like 'How were the data collected?', 'How are the variables defined and the constructs operationalized?' or 'Why, for what purpose, and in whose interest was the data collected in the first place?' (Finzer, 2013).

A question like how the data were collected may seem like a simple descriptive question. However, to use the data well and responsibly, we also need to understand the meaning of interactions around and through data. This is especially important because data used is increasingly created through interactions with networked and mobile data devices used by individuals in daily life. A device like a mobile phone can be an intimate possession, used to quantify the self, and to integrate and manage ongoing intimate relationships. It can also be a common possession that is casually shared between several people. In the first case, the connection of the technology with an individual self will be a valid assumption – and we can equate data from one phone with data from one individual. In the second case, it will not – data from one phone will be data about different individuals. A powerful demonstration of the variation in use and meaning of mobile digital devices is the analysis of mobile phone use in Sierra Leone by Erikson (Erikson, 2018). She shows how a mobile phone can have multiple users and how single users have multiple phones. These practices arise in interaction with markets, with infrastructure and corporate models (multiple phones can be an economical strategy when coverage is highly patchy and using another network than one's own provider is very costly) and with the way the phone is perceived as personal or as a more fluid object (Erikson, 2018).

Paying close attention to how data is collected and how it is framed is also important to maintain a critical stance towards technology use. This makes it possible to keep

questioning assumptions, and to make visible the filters through which Global North actors see the relation to technology of Global South actors. This means that in studies of the Global South users, "their new media practices are predominantly framed as instrumental and utilitarian, partly because development agendas drive this research with a strong historical bias towards socioeconomic impacts (Arora and Rangaswamy, 2013)." For instance, there is more emphasis on farmers checking crop prices online than watching pornography on their mobile devices (Rangaswamy and Arora, 2016). To be able to understand these interactions and to avoid making incorrect or universalizing assumptions about the meaning and import of data are very important skills.

Besides a thorough analysis of data quality and provenance, other habits of mind are also advantageous to learn to think with data across the different steps of data work:

- Acknowledge the need for data to gain insight.
- Look for the data: Ask "Which data could be helpful to reach conclusions, get insight or construct arguments?"
- Graph the data: Construct graphical representations that highlight potentially useful patterns in the data; patterns that are difficult to discern by staring at a table of numbers.
- Become immersed in the data: Use (and invent) measures. Look for and tell the story behind the data (Finzer, 2013).

Because data science is increasingly being used as a decision-making tool and because of the growing role of data in shaping our world, ensuring the quality and reliability of analyses and algorithms is also important. This requires the skill of making the connection between data management and analysis, particularly to enhance the reliability of algorithms and veracity of their outputs, and the accountability of knowledge acquired through data-intensive methods. Doing this well requires not only a solid understanding of statistics but also the ability to understand data journeys. For example, being able to clean a data set requires knowledge about what algorithms are able to do well and knowledge of their ability and limitations in detecting erroneous data. In one of the labs where one of us worked, condoms filled with water were routinely used to calibrate brain scanners. This enabled technicians to test and make technical adjustments to the scanners, without the inconvenience of having to put a person in the scanner. The technicians knew that the data analysis software was unable to systematically remove these scans from data sets and that they were likely to end up being processed as data, along with scans of brains. You can well imagine how this would skew results in very odd ways! Technicians therefore regularly intervened 'manually' in this otherwise highly automated image analysis pipeline, explaining "Computers are not nearly as good at recognizing garbage as humans are." Being able to work in this way requires an openness to really understand the specifics of data producing technologies, algorithms and their interaction with data. This skill can be developed by being exposed to a diversity of data practices and by maintaining a healthy understanding of the limitations of automation.

### *6.3.3 Contextualising skills*

Ultimately, much of data science work requires decisions with ramifications that go well beyond the setting in which a given tool, code or model is being produced. When deciding on which data to use, which algorithm to develop, which problem to tackle and how, data scientists are supporting specific ways of building the digital world that can have immediate and significant effects on every-day, material life. To address the potential implications of their work, data scientists are therefore in need of contextualizing skills.

To illustrate this, consider the very vehement and weighty discussions about sex/gender categories and their construction. Across datasets, gender is overwhelmingly put forth in as a binary. What would it mean to reconsider these categories? There is ongoing research to consider alternative, more inclusive ways of inquiring about sex/gender in population surveys, by using expansive check-all-that-apply gender identity lists or even write-in options (filling in a free text box) that offer maximum flexibility. Another approach is to use a multi-dimensional measure, to capture different dimensions of sex/gender: sex assigned at birth, current gender identity, trans status, lived gender, and hormonal and surgical status (Bauer et al., 2017). This improves the quality and applicability of data collected and can have significant effects. Consider that trans people experience large health and employment disparities. The binary sex/gender data system used in most large population studies makes this group completely invisible and therefore makes it systematically much more difficult to address inequality. Making such changes to the use of categories can improve the data to be collected. But what are the implications for using data we have already collected? How would new categories be mapped on to old ones? Could some of the variation that is currently propped into the m/f binary be retrieved from the data? Or would this be doing violence to the data, imposing an anachronistic understanding of sex/gender? Such questions matter, especially in current data systems that are very much focused on features of individuals and tend to erase social context.

To tackle these challenges, data curators, information management specialists and experts in data studies with contextualizing skills are taking a more prominent place alongside other data experts. While there is still a premium attached to individuals who can deliver an ingenious technical solution or shortcut, recent concerns with the popular image of AI have occasioned more interest in the ability to think about long-term consequences and about the contextual nature of the data structures used. For example, social science research can help understand the categories through which we make sense of our social and material world and how these are embedded in data science systems and in AI. Qualitative research is explicit about data collection and about the interventions that researchers make in the world, by creating conditions for observation or data gathering. These practices can greatly help data science to situate itself in the world and in making better design and implementation decisions (Sloane and Moss, 2019). There is increasing interest in individuals who have the skills to negotiate the intricacies and implications of decisions concerning modes of data access and sharing, and related legal and financial questions around who actually owns the data, whether and under which conditions data sharing constitutes an infringement of privacy and other individual rights, and what credit, if any, the original creators of a dataset should get when others successful re-use their data. In the corporate world, data is no longer simply input for management or logistics, but can form the core of an enterprise's business model. In this context, businesses will also want data scientists to

have affinity with business models and money-making. Data scientists will be expected to find novel ways to capitalize on data that is available (we delve into the value and re-use of data in Chapter 8).

### *6.3.4 Communication skills*

A final set of skills are those related to communication. Data scientists must be able to communicate with other team members as well as with stakeholders who may be closer or more distant to projects. In addition, communication with different publics is also important, via infographics or other visualisations. It is therefore very valuable to be able to communicate across a range of forms (oral, visual, textual) and with different groups.

As we've noted across different discussions in this book, data is heavily shaped by its context of production. Sharing data, even between academic research groups, is a complex undertaking. To get different research cultures to share information in ways that are intelligible and accommodate different views on the world and different values is a challenging task (Hilgartner, 1995; Hine, 2001; Leonelli, 2016a; Beaulieu, 2001). For example, the areas of biodiversity and ecology research would seem to share a deep concern for gathering and sharing data on where species of animals can be found. However, ecologists are very much oriented to documenting the kind of location (the particular ecosystem), whereas biodiversity researchers are concerned with the absolute location (geo-location). These differences translate 'down' to how the data fields of databases are ordered, as well as 'up' to the way the data is used as evidence for decreasing biodiversity or disappearing ecosystems. The communication skills needed to address this are very important: in such a situation, the data worker needs to make clear where the differences between understandings of a concept like location are coming from, the possibilities for implementing these differences in both the architectures and data flows of the digital systems, and to communicate the consequences of such decisions on implementation. The same holds for communicating the meaning of categories in surveys or the respective goals of different users. Furthermore, it is only in conversation with users that the data infrastructure can be designed or adapted to suit their multiple needs – again highlighting the importance of communication skills. The communication challenge only increases as users of data diversity and different kinds of industry, government agencies, citizen groups and academic institutions become involved with data. Of course, when successful, such communication across actors can be especially powerful and stable knowledge production systems can emerge across disciplinary, organizational and national borders (Edwards, 2019).

## 6.4 Bringing Skills Together

Let us now reconsider the thought experiment from the opening of this chapter, where we imagined what would be needed to put together a data science project on traffic patterns. Read through table 6.1 below (inspired by Finzer, 2013) and consider the tasks listed, as well as who has the required expertise to pursue these tasks and which skills would be most essential. Note that the column primary expertise refers to the types of expertise in Figure 6.3.

*Table 6.1: Components of a data science project on traffic patterns, divided by tasks, primary expertise and main skills required*.

| Illustration | Task | Primary expertise | Main skills required |
|---|---|---|---|
| What data do we have about traffic in the city? Which approaches might we use to collect more data? | Review existing data sets. | Governance<br>Epistemological (reflexive domain expertise) | Analytic<br>Contextualising<br>Technical |
| What else do we want to know and which evidence do we need to collect? | Decide what data should be gathered. | Epistemological (reflexive domain expertise) | Analytic<br>Communication |
| How can we engage enough citizens living and travelling in the area and how to ensure that the data collection will not harm of disadvantage individuals or groups? | Data collection should be designed to ensure privacy and avoid discrimination for individuals and groups; data subjects should be involved and informed | Governance<br>Statistical | Contextualising<br>Communication |
| How are we going to gather and store data from sensors in the city? | Set up a server as a repository of data streaming in real time from a large array of geographically distributed sensors. | Computational | Technical<br>Contextualising |
| How will we ensure that the data is flowing smoothly? | Develop data pipelines/process to annotate, filter, clean the data. | Governance | Technical<br>Communication |
| How will we assess the quality of the data and remove 'impossible' data, for ex. a sensor indicating that a truck that is 300m long rode through the historic city centre? | Explain the origin of outliers in a particular data set. | Epistemological (reflexive domain expertise) | Analytic<br>Technical |
| How will we assess whether the patterns in traffic are specific to the area studied? | Decide to what extent the conclusions drawn from analysis can be generalized. | Statistical<br>Epistemological (reflexive domain expertise) | Analytic<br>Technical |
| How will we present the data to other researchers, | Design a data visualization suitable | Computational | Communication<br>Technical |

| | | | |
|---|---|---|---|
| to the city's policy-makers and citizens? | for publication in an article | | |
| Should we use health records to examine traffic and air pollution effects? | Decide whether certain disparate data sets can be meaningfully merged. | Epistemological (reflexive domain expertise)<br>Statistical<br>Governance | Analytic<br>Contextualising |
| Which factors are most important to analyse the data? | Reduce the number of variables that need to be considered for a particular analysis. | Statistical | Technical<br>Analytic |
| How can we keep collecting data securely over a long period? | Set up a version management system for data that will be gathered over a number of years. | Computational | Technical<br>Contextualising |
| How can this project benefit as many researchers as possible? | Ensure proper data curation, deposit in a repository and licensing | Governance | Communication<br>Technical |

From this table, we see that multiple combinations of expertise and skills are needed to undertake such a project. It would be quite extraordinary to see that a single person can possess all the expertise and skills needed, and, on top of that, be able to deploy them in different ways in different projects. Instead, data scientists tend to have a particular combination of skills with which they have most affinity and which they further develop. They work in teams where their strengths can be complemented by those of others with different backgrounds. It is very common within large organisations to see data science groups include people trained as statistician, a computer programmer, an information management specialist and a social scientist, for example. Larger teams can be composed of several specialists in each of these areas. In some contexts, data curators and visualisation specialists will also be involved. Simply setting people with different expertise to work on a joint project is not enough to form an effective team. What makes a successful data science team lies in the ability for these different areas of expertise to work together: the key is collaboration.

### *6.4.1 Importance of collaboration in data science*

Why is collaboration difficult? While interdisciplinarity has been growing in universities, disciplinary training is still very strong in most educational programmes. Academic training socializes students to become members of disciplinary communities of practice (Lave and Wenger, 1991). They are taught by mentors and teachers, they develop similar experiences by doing lab work or practicals, and come to share a set of skills, ways of doing things and of pursuing work. This means that a discipline orients its members to a particular domain of enquiry, certain ways of defining problems, and

agreement on what counts as evidence. When we speak of turning a business question into a data science question or of reconciling the different ways of defining location as in the example of ecology versus biodiversity, we are talking about overcoming disciplinary differences. This work is sometimes described as translation or brokering. It involved language and terminology, but also ways of seeing data. When working in multi-disciplinary teams, members have to learn to interact across these practices and orientations, which means being aware of one's own approach and being able to understand how it differs from that of others. Disciplinary differences can run very deep and affect not only the direction of a project but also what counts as making progress in the project. In one study of bio-medical scientists who worked with data scientists, the bio-medical researchers used intermediate results of the project to revise their initial research question. To the biomedical scientists, this was a positive outcome since they were now asking a better question. For the data scientists whose goal was to transfer the initial research question into a well-defined data science question and to resolve it by using machine learning and optimising performance, this felt like they had wasted their time working on the initial question (Mao et al., 2019). Collaboration between experts is challenging and the extent to which is shapes a project should not be underestimated.

Collaboration also requires that all data scientists, no matter their expertise, understand the need for different profiles in a team and the role that these other skills can play in data analysis. This is one of the reasons why data science teaching and training increasingly seeks to provide students with skills for multi-disciplinary teamwork. Two common approaches to helping students develop these skills are to set up interdisciplinary teaching teams, and to expose students to 'real world' rather than textbook problems. The idea is that these experiences will help students understand how the different aspects are entwined and the variety of expertise needed.

### *6.4.2 Types of collaboration*

As we noted above, collaboration across disciplines has been praised for being more open to messy, societal challenges, and for supporting a diversity of sites of knowledge production. This means that collaborative science might be more likely to create responsive types of knowledge that lead to innovation and socially relevant research. On a day-to-day level in data science projects, the need to collaborate is often formulated much more pragmatically, as the need to get a job done. Perfect mutual understanding is rare in data science teams, as in any other area of life. Collaboration between experts and professionals has been studied by scholars from many fields, from computer-supported-collaborative-work, organisational psychology or science and technology studies and anthropology. All this work has not led to a perfect recipe for collaboration. A number of patterns have been observed, however, and awareness of these patterns can be a useful to tool to help set up teams and to gain insight into the kinds of interactions that are likely to occur.

To end this chapter, we propose a short overview of different models of collaboration, as a way to become aware of how relationships in teams and between teams can develop. This awareness will help in recognizing patterns of interaction. Following Barry and Born (2013), we propose two paradigms of interaction. In the first paradigm, collaboration is done across disciplinary boundaries, while the boundaries of disciplines

are maintained. This collaboration within this paradigm is often labelled multi-disciplinary or cross-disciplinary. Within this paradigm, one model of collaboration is the integration of the outcomes of different approaches to provide a synthesis of results. You can think of this as the 'happy family' model: where families might grow through marriage, and where the family culture is enriched by new additions – individuals interact harmoniously while maintaining their differences. In a data project, imagine team members pursuing their work according to their expectations and training, and exchanging the results of their efforts with other team members. In this model, there is sharing across disciplines, but members maintain their way of working and their assumptions are not questioned.

A second type of collaboration in this paradigm (where disciplinary boundaries are maintained) is more hierarchical. The relations between disciplines are organized so that one is subordinated to the other. You can think of this model as an 'upstairs-downstairs' situation, where one discipline provides a service to the another, which is more powerful (some data scientists are more equal than others). Typical of this kind of collaboration are projects in which engineers develop new technological approaches and social scientists (often social psychologists) are brought in at a late stage of development to ensure fit with social factors or to organize 'public acceptance'. The example of the collaboration between biomedical scientists and data scientists observed by Mao et al., also had features of this kind of collaboration: the biomedical scientists changed the research question partway through the project in a way that surprised the data scientists, and the latter had to take this new question on board and develop ways to address it. The data scientists were in service to the biomedical scientists, each group pursuing their work according to their disciplinary assumptions, but with the goals of some members taking precedence over those of others.

Sometimes this relationship is clear from the outset and made explicit in the roles, where one part of a team is labelled as 'research' or 'scientific staff' and the other as 'support' or 'professional' staff. A long-standing hierarchy in the sciences – that places natural science and engineers at the top – is an important determinant of whose agenda takes priority. Some disciplines are more powerful within the university or within corporations, often wielding more cultural capital and more resources (think engineers versus social psychologists or the sales department versus marketing). But sometimes the differences are subtler, at least at the beginning of a project. In such situations, there is often an imbalance in the perception of collaboration, where some members consider that they collaborate with others, but this perception is not reciprocated (A reports that they collaborate with B, but B does not report collaboration with A) (Zhang, Muller, and Wang, 2020). Underlying this skewed perception is the way different contributions are valued as being substantive to the core objectives of the project or whether they are seen as non-essential, 'nice-to-have' elements.

The second paradigm in collaboration seeks to make the whole greater than the sum of the parts, and to question and go beyond the limits of established disciplines. In this paradigm, the disciplinary boundaries are not maintained. Models of collaboration in this paradigm seek to synthesize perspectives – not just results. Such collaborations have been heralded as promising of a new kind of knowledge production (Weingart and Padberg, 2014), that may also be more open to lay knowledge in problem solving. In practice, this kind of collaboration involves the creation of common understanding of

both processes and contents. This is sometimes described as the sharing of a 'trading zone' or 'third space', "where people can compare, negotiate, and integrate goals, perspectives and vocabularies, as well as discuss shared meanings and protocols." (Mao et al., 2020) Ideally, this leads to shared criteria of quality and success. This is a challenging route (Mauthner and Doucet, 2008) that may feel risky to those involved. Such a mode of collaboration takes time and can feel uncomfortable or even unproductive. Whether it is possible to go for this kind of collaboration depends on the institutional setting, the kind of funding available and whether one is in a corporate or academic regime (or a mix). Yet, insofar as teams and team members can shape how they collaborate, it is possible to foster conditions that lead to better collaboration. Such learning and discomfort should be embraced as part of the job of pursuing fruitful collaboration. They can even be seen as the key to success. Teams should pay attention to creating the conditions for learning and to nurture learning in participants (Freeth and Caniglia, 2020).

Freeth and Caniglia offer the following suggestions. First, the creation of a common ground is important: it can mean sharing a concept or a method that is relevant to all involved. Creating a sense of safety and trust is also necessary. This can be done by making it acceptable to discuss failures in a team and to explore the diversity that may exist within a team. Spaces for interaction that do not reinforce hierarchies are also helpful (not feeling like some team members are 'guests' on the home territory of the rest of the team). Finally, teams should ensure that there is time to discuss both procedural issues (how to work together) as well as outcomes of the project (Freeth and Caniglia, 2020). Other factors such as personal affinity between team members and track record of participants also matter in developing effective collaboration.

Finally, from a more technical angle, there is a growing set of tools to support collaboration. Some tools are specifically aimed at data science collaboration, to support the documentation of the provenance of data and of code. Such tools, if well implemented and supported by a local culture, can contribute to keeping data work transparent and accountable. Other tools are directed to documenting data processing and analysis, such as GitHub, Slack, or Jupyter Notebook. Of course, many aspects of collaboration are supported by more generic tools like email, document sharing and co-writing tools like Google docs and by file sharing services, as well as meetings, in person and online.

## 6.5 Conclusion: Becoming a data scientist today

While being a data scientist has been labelled 'the sexiest job' for almost a decade, what is associated with this role has changed. Data scientists were meant to help organisations capitalise on Big Data, later, to help personalise products or unleash 'intelligence'. Recently, data scientists have been seen as the key to achieving the goals of implementing machine learning and AI. Across these different kinds of hype, it has generally been recognised that being a data scientist means dealing with expectations from other parts of the organisation. Dealing with pressures and politics of organisations is yet another set of skills, and while they may be important for every job, when high hopes are pinned on establishing a new unit or project team, these

expectations can be especially determinant. Besides the growing demand for data scientists, another significant development has been the 'mainstreaming' of data science. Many jobs in sectors not primarily associated with data science, such as education or policing, are increasingly requiring some awareness of data work and data science skills. There is also a growing value attached to skills and activities relating to the stewardship of data – this is not as yet implemented in most universities, but is strongly championed by funders and policy-makers. For example, the job of 'data manager' is not yet well-defined and those doing this job tend to be hidden under other administrative labels that do not do justice to their centrality in research. Overall, there seems to be increasing awareness that coding is neither the sole nor primary skill involved in doing data science.

People seeking to enter the data science job market will also find that it is strongly shaped by portfolio development and by participation in hackathons and other data competitions (Kaggle is probably the best-known brand). This means investing in demonstrating that you have particular abilities by sharing code or taking part in events – and providing free labour. This is quite different than relying on formal credentials, such as a degree from a recognised institution. It also means that networks and networking skills are especially important. This approach to recruitment tends to favour homogeny – people know people like themselves. So if job opportunities depend on who you know, there is little opportunity to diversify the work force. This means that there should be increased attention to diversity of backgrounds of data workers, and particularly data scientists and data curators. Having only white, middle-class, technically-trained men working as data scientists – as is still overwhelmingly the case in corporate data analytics firms – is far from ideal in terms of bringing a variety of experiences and viewpoints to the table. Going back to the questions we posed about how to handle binary sex classifications, it is clear that having data scientists with some knowledge of gender studies and an understanding of intersectionality would contribute to ensuring that the categories used for data classification do not discriminate or otherwise adversely affect relevant individuals and groups. Making positive efforts to address the current lack of diversity in terms of gender, age, ethnicity and class is a constructive step towards a better data science (we return to this topic in the concluding chapter of this book).

Finally, there are new civic and corporate roles for data scientists. The security and sensitivity of data, consequences and privacy concerns of data analysis, and the professionalism of transparency and reproducibility are all increasingly important areas of expertise in contexts beyond data science units in companies or research groups in universities. For example, data scientists have the expertise needed to work with or as journalists, to help make governments and businesses accountable because they understand how data is being generated and used. A recent investigation by the British independent daily newspaper *The Guardian* established that over a quarter of British councils (local government authorities in the UK) have invested in software contracts with large firms to support the administration of benefits to citizens (Marsh, 2019). The software systems are used in the activities of local councils. These include providing housing benefits (a subsidy for rental expenses), detecting signals of child abuse and allocating places to pupils in schools. Across these different applications, *The Guardian* reported concerns about privacy and data security, the ability of council officials to understand how some of the systems work, and the difficulty for citizens to

challenge automated decisions. In particular, the performance of some systems on 'predictive analytics' had been problematic: in detecting cases of potential fraud, low risk claims for benefits had been wrongly labelled as high risk. This case of too many false positives had harsh consequences, delaying the payment of benefits and causing undue hardship on vulnerable groups (Marsh, 2019). This is an example of how social and political issues require an understanding of data science in order to be tacked. Many NGOs and citizen movements are therefore drawing on the expertise of data scientists to navigate and evaluate the use of large-scale data systems, whether to create alternative data sets (citizen sensing projects) or to audit and demand more responsible systems.

As machine learning and data-driven policy spread across different levels of government and areas of life, the need for such experts will also increase – data scientists can be heroes of social justice, as well as the champions of new business models and drivers novel scientific insights.

### *Additional Reading*

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C. and Borgman, C.L., 2011. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science* 41(5), pp.667–90.

Meng, Xiao-Li. 2019. Data Science: An Artificial Ecosystem. Harvard Data Science Review 1(1). https://doi.org/10.1162/99608f92.ba20f892

Wouters, P., Beaulieu, A., Scharnhorst, A. and Wyatt, S. eds., 2013. *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*. MIT Press.

Jemielniak, D. and A. Przegalinska. 2020. *Collaborative Society*. MIT Press.

## Glossary

**AI/Artifical Intelligence** is an umbrella term used to encompass a broad range of automated forms of data analysis, typically combining statistics, modelling, programming, and computing.

**Algorithm** in the context of machine learning refers to the operations and calculations performed on data.

**APIs** (application programming interfaces) are an interface built by platforms that makes it possible to connect applications and share data.

**Big Data** is a loose term often used to refer to the large and diverse body of data generated by digital technologies through the datafication of human activities.

**Big Data empiricism** is the belief that data are the best form of evidence to establish truth, to form opinions about the world, and to make judgments.

**Big Data mythology** is a set of inflated expectations around how Big Data enables new, cheap and efficient ways to plan, conduct, institutionalise, disseminate and assess research.

**Conventions** are standards agreed upon by the relevant stakeholders, which make it possible to retrieve and link data across different platforms.

**Correlation** is a statistical relationship between two data values. When two values are correlated, we know that when one changes, the other will change as well.

**Curation** in the context of data work is the process of organising, formatting and annotating data so that the data can be retrieved, shared and re-used.

**Datafication** is the process of turning of objects and activities into data.

**Data ethics** is the study of what it means for data work to be socially responsible and beneficial to life on earth.

**Data fairness** involves considering how data work can help to treat people in ways that are right and reasonable.

**Data governance** refers to the ensemble of regulations, norms and socio-technical systems that enables and directs data work and particularly how, where and why data can travel.

**Data journeys** designate the movement of data from their site of production to many other sites where they are processed, mobilised and re-purposed. Sites can encompass diverse times, disciplines or viewpoints.

**Data justice** concerns the specific circumstances of data work, and how those circumstances may affect whether such work is socially damaging or socially beneficial (and to whom).

**Data mobility** is a label for the extent to which data move across space and time.

**Data models** are the result of the efforts made to structure and visualize data, so that the data can be used as representations of a specific aspect of the world.

**Data provenance** refers to the conditions under which data were generated.

**Data science** is a newly emerged research domain which includes several types of expertise relevant to data analysis, such as computational and statistical skills; epistemological expertise, including an understanding of where data fit in the processes of knowledge production; and expertise on data governance and ethics.

**Data subject** is a label for a person who can be identified directly or indirectly by an identifier such as name, location data, online identifier or by facets of one's identity, be they physical, physiological, genetic, mental, economic, cultural or social.

**Data visualisation** is the process of ordering and interacting with data in visual form. Most data visualisations aim to facilitate the discovery of patterns.

**Data workers** are individuals who are in a position to take decisions concerning what data should be gathered and used, for which purposes, and in which ways.

**Ethics consists of** philosophical reflection on what it means to be a good person.

**Evidence** is the use of data to provide reasons to believe in a particular claim.

**FAIR** is an acronym that stands for four principles introduced in 2016 as guidance for data management and sharing: Findability, Accessibility, Interoperabiilty and Reuse.

**GDPR** (General Data Protection Regulation) is a piece of European legislation introduced in 2018 to protect individuals from abuse of their personal data and encourage the development of sophisticated and responsible ways of collecting, archiving, mobilising and reusing personal data.

**Information society** denotes a society where information is central to the capitalist system of production, innovation and consumption.

**Knowledge society** refers to a society that generates, processes, shares and makes knowledge that may be used to improve the human condition available to all its members

**Knowledge commons** is a label used to describe knowledge as a public good that contributes key insights on human life and that should be accessible without restrictions.

**Machine learning** is a branch of artificial intelligence (AI) where a data set is used by a computer to build and/or further refine a computational approach to solving a specific problem, such as image recognition or classifying information.

**Metadata** are structured data used to describe the characteristics of a given dataset, such as its provenance or significance.

**Metrics** are measures used for assessment.

**Morality** is the systems of norms and rules that tell us what is right and what is wrong, i.e. how we should behave.

**Networks** are systems of interconnected things, processes or individuals. Digital networks linking computing devices are central to the transmission of digital content.

**Open Science** is a movement committed to promoting collaborative research practices and the widespread distribution and reuse of data, results and methods.

**Platforms** are programmable infrastructures upon which other software can be developed and run.

**Raw data** are data that have just been generated and have not been further processed.

**Space of agency** (for data workers) is the space within which data workers can make decisions and take responsibility for the implications of those decisions.

**Statistical models** are precise and concise mathematical descriptions of datasets that enable calculations.

## Works Cited

### A


Abrieu, R., Rapetti, M., Aneja, U. and Chetty, K., 2019. How to Promote Worker Wellbeing in the Platform Economy in the Global South. *G20 Insights,* Japan.

Ackoff, R., 1989. From Data to Wisdom. *Journal of Applied Systems Analysis* 16, pp. 3–9.

Alaimo, C., and Kallinikos J., 2017. Computing the Everyday: Social Media as Data Platforms. *The Information Society* 33(4), pp.175–91.

Amano, T. and Sutherland, W.J., 2013. Four Barriers to the Global Understanding of Biodiversity Conservation: Wealth, Language, Geographical Location and Security. *Proceedings. Biological sciences*, 280(1756), p.20122649.

Amit-Danhi, E.R., and Shifman L., 2018. Digital Political Infographics: A Rhetorical Palette of an Emergent Genre. *New Media & Society* 20(10), pp.3540–59.

Amoore, L., 2009. Lines of sight: On the visualization of unknown futures. *Citizenship Studies*, 13(1), pp.17-30.

Amoore, L., 2011. Data derivatives: On the emergence of a security risk calculus for our times. Theory, *Culture & Society* 28(6), pp.24-43.

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, 23 June 2008.

Ankeny, R. A., and Leonelli, S., 2016. Repertoires: A Post-Kuhnian Perspective on Collaborative Research. *Studies in History and Philosophy of Science Part A*, *60*, pp.18-28.

Aouragh, M., Gürses, S., Pritchard, H. and Snelting, F., 2020. The Extractive Infrastructures of Contact Tracing Apps. *Journal of Environmental Media*, *1*(2), pp.9-1.

Aronova, E., Baker, K.S. and Oreskes, N., 2010. Big Science and Big Data in Biology: from the International Geophysical Year Through the International Biological Program to the Long-Term Ecological Research (LTER) Network, 1957 - Present. *Historical Studies in the Natural Sciences*, *40*(2), pp.183-224.

Arora, P. and Rangaswamy, N., 2013. Digital leisure for development: reframing new media practice in the global South. *Media, Culture & Society*, *35*(7), pp.898-905.


### B


Barry, A., and Born, G., eds. 2013. *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*. 1 edition. London; New York, NY: Routledge.

Bates, J., Cameron, D., Checco, A., Clough, P., Hopfgartner, F., Mazumdar, S., Sbaffi, L., Stordy, P. and de la Vega de León, A., 2020, January. Integrating FATE/Critical Data Studies into Data Science Curricula: Where Are We Going and How Do We Get There?. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp.425-435, New York, NY, USA: Association for Computing Machinery.

Bates, J., Lin, Y.W., and Goodale, P., 2016. Data journeys: Capturing the socio-material constitution of data objects and flows. Big Data Soc. 3, 2053951716654502.

Brown, Lydia X.Z., Michelle Richardson, Ridhi Shetty, Andrew Crawford, and Timothy Hoagland. 2020. Report: Challenging the Use of Algorithm-Driven Decision-Making in Benefits Determinations Affecting People with Disabilities. Georgetown, DC, USA: Center for Democracy and Society. https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/.

Bauer, S., 2008. Mining Data, Gathering Variables and Recombining Information: The Flexible Architecture of Epidemiological Studies, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 39(4), pp.415–428.

Bauer, G.R., Braimoh, J., Scheim, A.I. and Dharma, C., 2017. Transgender-Inclusive Measures of Sex/Gender for Population Surveys: Mixed-Methods Evaluation and Recommendations. *PLoS ONE 12*(5).

Bell, D., 1979. The Social Framework of the Information Society. In: *The Computer Age: A Twenty-Year View*, (eds.) M Dertoozos and Joel Moses, pp.500–549. Cambrdge, MA.: MIT Press.

Beaulieu, A. 2001. Voxels in the Brain: Neuroscience, Informatics and Changing Notions of Objectivity. *Social Studies of Science* 31(5), pp 635–80.

Beaulieu, A., 2002. Images are not the (only) truth: Brain mapping, visual knowledge, and iconoclasm. *Science, Technology & Human Values*, 27(1), pp.53-86.

Beaulieu, A., 2004. From Brainbank to Database: The Informational Turn in the Study of the Brain. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 35(2), pp.367-390.

Beaulieu, A., 2021. Organising Knowledge for Sustainable Futures. In: *Handbook for the Anthropology of Technology*, (eds.) Britt Ross Winthereik and Klaus Hoeyer. Palgrave.

Beaulieu, A., and Estalella, A., 2012. Rethinking Research Ethics for Mediated Settings. *Information, Communication & Society* 15(1), pp. 23–42.

Beaulieu, A., de Rijcke, S., van Heur, B., Wouters, P., Scharnhorst, A., Wyatt, S., 2013. Groningen Energy & Sustainability Programme, and Energy and Sustainability Research Institute Groningen. Authority and Expertise in New Sites of Knowledge Production. In: *Virtual Knowledge*, 25–56, Cambridge, Massachusetts: MIT Press.

Beer, D., 2015. Productive Measures: Culture and Measurement in the Context of Everyday Neoliberalism. *Big Data & Society*, *2*(1), p.2053951715578951.

Beer, D., 2016. *Metric power*. London: Palgrave Macmillan.

Beer, D., 2018. *The data gaze: Capitalism, power and perception*. Sage.

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V. and Szalay, A.S., 2018. Realizing the Potential of Data Science. *Communications of the ACM*, *61*(4), pp.67-72.

Bezuidenhout, L., Leonelli, S., Kelly, A. and Rappert, B., 2017. Beyond the Digital Divide: Towards a Situated Approach to Open Data. *Science and Public Policy*, *44*(4), pp.464-475.

Bezuidenhout, L. and Ratti, E., 2020. What does it mean to embed ethics in data science? An integrative approach based on the microethics and virtues. *AI & Society*, pp.1-15.

Bigo, D., Isin, E. and Ruppert, E., 2019. *Data Politics: Worlds, Subjects, Rights*, (eds.) Didier Bigo, Engin Isin, and Evelyn Ruppert, London and New York: Routledge.

Birch, K. and Muniesa, F. eds., 2020. *Assetization: Turning Things into Assets in Technoscientific Capitalism*. Cambridge, Massachusetts: MIT Press.

Birhane, A. and Cummins, F., 2019. Algorithmic injustices: Towards a relational ethics. Pre-print, Available at: https://arxiv.org/abs/1912.07376 [Accessed: 21.02.2021].

Blasimme, A., and Vayena, E., 2020. What's Next for COVID-19 Apps? Governance and Oversight. *Science 370* (6518), pp.760–62.

Blei, D.M. and Smyth, P., 2017. Science and Data Science. Proceedings of the National Academy of Sciences, 114(33), pp.8689-8692.

Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z. and Ziemlicki, C., 2012. Data for Development: The D4D Challenge on Mobile Phone Data. *Computers and Society [cs.CY] ArXiv:1210.0137*.

Bogen, J., 2010. Noise in the World. *Philosophy of Science*, 77(5), pp. 778–791.

Bonnin, N., van Andel, A. C., Kerby, J.T., Piel, A.K., Pintea, L. and Wich, S.A., 2018. Assessment of Chimpanzee Nest Detectability in Drone-Acquired Images. *Drones* 2(2), pp.17.

Borgman, C.L., 2015. *Big Data, Little Data, No Data*, Cambridge, Massachusetts: MIT Press

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O., Rawlins, M., Thornton, J., Vallance, P. and Walport, M., 2012. Science as an open enterprise. *The Royal Society*.

Boyd, D. and Crawford, K., 2012. Critical Questions for Big Data. *Information, Communication & Society* 15(5), pp.662–79.

Brayne, S., 2020. *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press, USA.

Bringsjord, S. and Govindarajulu, N.S., 2018. The Stanford Encyclopedia of Philosophy: Artificial Intelligence.

British Academy & Royal Society, 2017. *Data Management and Use: Governance in the 21st Century. A Joint Report of the Royal Society and the British Academy*, Available at: https://www.thebritishacademy.ac.uk/publications/data-ai-management-use-governance-21st-century/ [Accessed: 15.02.2021].

Brown, L.X.Z., Richardson, M., Shetty, R., Crawford, A. and Hoagland, T., 2020. Report: Challenging the Use of Algorithm-Driven Decision-Making in Benefits Determinations Affecting People with Disabilities. *Centre for Democracy and Technology*. Georgetown, DC, USA, Available at: https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/ [Accessed: 24.02.2021].

Bruns, A. and Burgess, J., 2016. Methodological innovation in precarious spaces: the case of Twitter. In: *Digital methods for social science: An interdisciplinary guide to research innovation*, 17-33, (eds.) Helene Snee, Christine Hine, Yvette Morey, Steven Roberts and Hayley Watson, Palgrave Macmillan.

Brunton, F. and Nissenbaum, H., 2015. *Obfuscation: A User's Guide for Privacy and Protest*. Cambridge, Massachusetts: MIT Press.

Buolamwini, J. and Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in Commercial Gender Classification. In: Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

Burgelman, J.C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K. and Schouppe, M., 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data 2*(43), pp.1-6.

Burton, P.R., Murtagh, M.J., Boyd, A., Williams, J.B., Dove, E.S., Wallace, S.E., Tasse, A.M., Little, J., Chisholm, R.L., Gaye, A. and Hveem, K., 2015. Data Safe Havens in Health Research and Healthcare. *Bioinformatics*, *31*(20), pp.3241-3248.

## C

Cai, L., and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14(0), pp.2.

Castelfranchi, C., 2007. Six Critical Remarks on Science and the Construction of the Knowledge Society. *Journal of Science Communication* 6(4), pp.C03.

Castells, M., 1996. *The Rise of the Network Society*. Cambridge, Massachusetts: Blackwell Publishers.

Chun, W.H.K., 2016. *Updating to Remain the Same: Habitual New Media*. Cambridge, Massachusetts: MIT Press.

Cinnamon, J., 2019. Data Inequalities and Why They Matter for Development. *Information Technology for Development*, *26*(2), pp.214-233.

Coeckelbergh, M., 2021. AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges. *AI and Ethics* 1(1), pp.67–72.

Conway, D., 2015. The Data Science Venn Diagram. Drew Conway. 2010. Available at: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram. [Accessed: 15.02.2021].

Couldry, N. and Mejias, U.A., 2019. *The Costs of Connection: How Data Are Colonizing Human Life and Appropriating It for Capitalism.* 1st edition. Stanford, California: Stanford University Press.

## D

D'Ignazio, C., and Klein, L.F., 2020. *Data Feminism*. Cambridge, Massachusetts: MIT Press.

Daston, L., 2016. History of science without structure. *Kuhn's Structure of Scientific Revolutions at Fifty: Reflections on a Science Classic*, pp.115-132.

Daston, L., 2017, *Science in the Archives: Pasts, Presents, Futures*, Chicago, IL: University of Chicago Press.

Daston, L., and Galison, P., 2007. Objectivity. New York: Zone Books.

De Chadarevian, S., 2018. Things and Data in Recent Biology. *Historical Studies in the Natural Sciences 48*, 5(2018), pp.648-658.

De Rijcke, S., Wouters, P.F., Rushforth, A.D., Franssen, T.P. and Hammarfelt, B., 2016. Evaluation practices and effects of indicator use – a literature review. *Research Evaluation*, *25*(2).

De Rijcke, S. and Beaulieu, A., 2011. Image as Interface: Consequences for Users of Museum Knowledge. Library Trends, 59(4), pp.663-685.

De Rijcke, S. and Beaulieu, A., 2014. Networked Neuroscience: Brain Scans and Visual Knowing at the Intersection of Atlases and Databases. In: *Representation in Scientific Practice Revisited*, (eds.) Catelijne Coopmans, Steve Woolgar, Janet Vertesi, and Michael Lynch. Cambridge, Massachusetts: MIT Press.

De Veaux, R.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R. and Kim, A.Y., 2017. Curriculum Guidelines for Undergraduate Programs in Data Science. Annual Review of Statistics and Its Application, 4, pp.15-30.

Decuyper, A., Browet, A., Traag, V., Blondel, V.D. and Delvenne, J.C., 2016. Clean up or Mess up: The Effect of Sampling Biases on Measurements of Degree Distributions in Mobile Phone Datasets. *ArXiv:1609.09413 [Physics]*, September.

Derksen, M., and Beaulieu, A., 2011. Social Technology. In: *The Handbook of Philosophy of Social Science*, (eds.) Ian C. Jarvie and Jesus Zamora-Bonilla, pp.703–19. SAGE Publications.

Desrosières, A., 2010. *La Politique des Grands Nombres*. La Découverte.

Dormans, S., and Kok, J., 2010. An Alternative Approach to Large Historical Databases; Exploring Best Practices with Collaboratories. *Historical Methods* 43(3), pp.97–107.

Drahokoupil, J. and Jepsen, M., 2017. The Digital Economy and Its Implications for Labour: 1. The Platform Economy. *Transfer: European Review of Labour and Research*, *23*(2), pp.103-119.

Drucker, J., 2014. *Graphesis: Visual Forms of Knowledge Production*. Cambridge, Massachusetts: Harvard University Press.

Dumit, J., and Nafus. D., 2018. The Other Ninety per Cent: Thinking with Data Science, Creating Data Studies – Joseph Dumit Interviewed by Dawn Nafus. In: *Ethnography for a Data-Saturated World*, (eds.) Hannah Knox and Dawn Nafus, pp.252–74. Manchester, UK: Manchester University Press.

Dye, C., 2014. After 2015: Infectious Diseases in a New Era of Health and Development. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1645), pp. 20130426.

## E

Ebeling, M., 2016. Healthcare and Big Data: Digital Spectres and Phantom Objects. London, New York: Palgrave Macmillan.

Edelman, B., Luca, M. and Svirsky, D., 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, 9(2), pp.1-22.

Edwards, P.N., 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, Massachusetts London, England: MIT Press.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C. and Borgman, C.L., 2011. Science Friction: Data, Metadata, and Collaboration. Social Studies of Science 41(5), pp.667–90.

Edwards, P.N., 2019. Knowledge Infrastructures under Siege: Climate Data as Memory, Truce, and Target. In: *Data Politics*: *Worlds, Subjects, Rights*, (eds.) Didier Bigo, Engin Isin, and Evelyn Ruppert, 21–42, Routledge.

Erikson, S. L., 2018. Cell Phones ≠ Self and Other Problems with Big Data Detection and Containment during Epidemics. *Medical Anthropology Quarterly* 32(3), pp.315–39.

Eubanks, V., 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Illustrated edition. New York, NY: St Martin's Press.

European Commission, 2016. *Open Innovation, Open Science, Open to the World. A vision for Europe*. Luxembourg, Publications Office of the European Union.

Eyert, F., Irgmaier, F. and Ulbricht, L., 2020. Extending the framework of algorithmic regulation. The Uber case. *Regulation & Governance*, Available at: https://doi.org/10.1111/rego.12371 [Accessed: 21.02.2021].

## F

Featherstone, R., 2014. Visual Research Data: An Infographics Primer. *The Journal of the Canadian Health Libraries Association. Journal de l'Association des Bibliothèques de la Santé du Canada*, *35*(3), pp.147-150.

Fecher, B. and Friesike, S., 2014. Open Science: One Term, Five Schools of Thought. *Opening Science*, p.17.

Finzer, W., 2013. The Data Science Education Dilemma. *Technology Innovations in Statistics Education,* 7(2).

Fleming, L., Tempini, N., Gordon-Brown, H., Nichols, G.L., Sarran, C., Vineis, P., Leonardi,G., *et al.,* 2017. Big Data in Environment and Human Health. Oxford Research Encyclopedia of Environmental Science.

Floridi, L., 2011. *The Philosophy of Information*. Oxford University Press.

Floridi, L., 2013. Distributed Morality in an Information Society. *Science and Engineering Ethics*, 19(3), pp.727-743.

Floridi, L., 2014. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford: Oxford University Press.

Floridi, L., 2015. *The Ethics of Information*. Oxford: Oxford University Press.

Floridi, L., 2017. Robots, Jobs, Taxes, and Responsibilities. *Philosophy & Technology*, *1*(30), pp.1-4.

Floridi, L., and Illari, P., eds. 2014. *The Philosophy of Information Quality*. Synthese Library. Springer International Publishing.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People – an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), pp.689-707.

Francois, K., Monteiro, C. and Allo, P., 2020. Big-Data Literacy as a New Vocation for Statistical Literacy. *Statistics Education Research Journal*, *19*(1).

Freeth, R. and Caniglia, G., 2020. Learning to Collaborate While Collaborating: Advancing Interdisciplinary Sustainability Research. *Sustainability Science*, *15*(1), pp.247-261.

Fricker, M., 2009. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press.

Frosh, S., 2016. Relationality in a Time of Surveillance: Narcissism, Melancholia, Paranoia. *Subjectivity*, *9*(1), pp.1-16.

Eubanks, V., 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Illustrated edition, New York: St. Martin's Press.

## G

Gabriel, L. and Casemore, R. eds., 2009. *Relational ethics in practice: Narratives from counselling and psychotherapy*. Routledge.

Galbraith, J.K., 2017. *Economics in perspective: A critical history*. Princeton University Press.

Garber, A.M., 2019. Data Science: What the Educated Citizen Needs to Know. *Harvard Data Science Review 1*(1).

Gibbons, M. 1994. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Gillespie, T., 2010. The Politics of “Platforms” - Tarleton Gillespie, 2010. *New Media & Society*, February.

Gillespie, T., 2017. The Platform Metaphor, Revisited. *Culture Digitally* (blog). Available at: http://culturedigitally.org/2017/08/platform-metaphor/ [Accessed: 15.02.2021].

Given, L.M., 2008. The Sage Encyclopedia of Qualitative Research Methods: Vols. 1-2, Thousand Oaks, California: SAGE.

Garcia, P., Sutherland, T., Cifor, M., Chan, A.S., Klein, L., D'Ignazio, C. and Salehi, N., 2020. No: Critical Refusal as Feminist Data Practice. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, pp. 199-202.

Green, S. and Vogt, H., 2016. Personalizing Medicine: Disease Prevention in Silico and in Socio. *To appear in Humana. Mente Journal of Philosophical Studies*, *30*.

Greene, D., Hoffmann, A.L. and Stark, L., 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: *Proceedings of the 52nd Hawaii international conference on system sciences*.

Greenberg, J., 2017. Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata. *Journal of Data and Information Science* 2(3), pp.19–36.

GSMA, 2018. Helping End Tuberculosis in India by 2025. Available at: https://www.gsma.com/betterfuture/wp-content/uploads/2018/12/Helping_end_Tuberculosis_in_India_by_2025.pdf.

Gstrein, O.J., Bunnik, A. and Zwitter, A., 2019. Ethical, Legal and Social Challenges of Predictive Policing. *Católica Law Review, Direito Penal*, *3*(3), pp.77-98.

**H**

Harris, A., Kelly, S. and Wyatt, S., 2016. *CyberGenetics: Health genetics and new media*. London and New York, Routledge.

Hazard, L., Cerf, M., Lamine, C., Magda, D. and Steyaert, P., 2020. A Tool for Reflecting on Research Stances to Support Sustainability Transitions. *Nature Sustainability*, 3(2), pp.89-95.

Hess, C. and Ostrom, E., 2007. Understanding Knowledge as Commons. Cambridge, Massachusetts: MIT Press.

Hewson, M., 1999. Did Global Governane Create Informational Globalism? *Approaches to Global Governance Theory*, pp.97-113.

Hey, T., Tansley, S. and Tolle, K., ed., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.

Hilgartner, S., 1995. Biomolecular Databases: New Communication Regimes for Biology? *Science Communication*, *17*(1), pp.240-263.

Hilgartner, S., 2017. *Reordering Life: Knowledge and Control in the Genomics Revolution*. Cambridge, Massachusetts: MIT Press.

Hine, C., 2001. Web Pages, Authors and Audiences: The Meaning of a Mouse Click. *Information, Communication & Society*, *4*(2), pp.182-198.

Hoang, L., Blank, G. and Quan-Haase, A., 2020. The Winners and the Losers of the Platform Economy: Who Participates?. *Information, Communication and Society*, *23*(5), pp.681-700.

Hogle, L.F., 2016. The ethics and politics of infrastructures: Creating the conditions of possibility for big data in medicine. In: The Ethics of Biomedical Big Data, pp. 397-427. Springer, Cham.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. and Wallach, H., 2019, May. Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. CHI '19. New York, NY: Association for Computing Machinery.

Hood, L. and Friend, S.H., 2011. Predictive, Personalized, Preventive, Participatory (P4) Cancer Medicine. *Nature reviews. Clinical oncology*, *8*(3), pp.184-187.

Horton, N.J. and Hardin, J.S., 2015. Teaching the Next Generation of Statistics Students to "Think with Data": Special Issue on Statistics and the Undergraduate Curriculum. *The American Statistician*, *69*(4), pp.259-265.

Hristova, D., Williams, M.J., Musolesi, M., Panzarasa, P. and Mascolo, C., 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In: *Proceedings of the 25th International Conference on World Wide Web*, 21–30. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee.

## I

International Telecommunications Union, 2018. Measuring the Information Society Report, 1(1) Geneva, Switzerland, p.204.

Irwin, A., 2018. No PhDs Needed: How Citizen Science Is Transforming Research. *Nature*, *562*(7728), pp.480-482.

## J

Jasanoff, S., 2003. Technologies of Humility: Citizen Participation in Governing Science. *Minerva*, 41(3), pp.223-244.

Jones, K.M., Ankeny, R.A. and Cook-Deegan, R., 2018. The Bermuda Triangle: The pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology*, *51*(4), pp.693-805.

## K

Kaldestad, Ø.H., 2016. 250,000 Words of App Terms and Conditions. Available at: https://www. forbrukerradet. no/side/250000-words-of-app-terms-and-conditions/ [Accessed: 21.02.2021]

Kallinikos, J. and Tempini, N., 2014. Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation. *Information Systems Research*, *25*(4), pp.817-833.

Kaye, J., Whitley, E.A., Lund, D., Morrison, M., Teare, H. and Melham, K., 2015. Dynamic consent: a patient interface for twenty-first century research networks. *European journal of human genetics*, 23(2), pp.141-146.

Kennedy, H., Hill, R.L., Aiello, G. and Allen, W., 2016. The Work That Visualisation Conventions Do. *Information, Communication & Society,* 19(6), pp.715–35.

Kitchin, R. and Dodge, M., 2007. Rethinking Maps. *Progress in Human Geography* 31(3), pp.331–44.

Kitchin, R. and Dodge, M., 2011. *Code/Space: Software and Everyday Life*. Cambridge, Massachusetts: MIT Press.

Kitchin, R., 2014. *The data revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Kitchin, R., and McArdle, G., 2016. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society* 3 (1), pp.2053951716631130.

Kuner, C., 2017. Reality and Illusion in EU Data Transfer Regulation Post Schrems. *German Law Journal*, *18*(4), pp.881-918.

## L

Lagoze, C., 2014. Big Data, Data Integrity, and the Fracturing of the Control Zone. *Big Data &* Society, 1(2), pp.2053951714558281.

Lane, J., 2020. Democratising our Data: A Manifesto. Cambridge, Massachusetts: MIT Press.

Lave, J., and Wenger, E., 1991. *Situated Learning: Legitimate Peripheral Participation. Learning in Doing: Social, Cognitive and Computational Perspectives.* Cambridge: Cambridge University Press.

Letouze, E., 2015. Big Data & Development: An Overview. Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative.

Leonelli, S., 2012. Introduction: Making Sense of Data-Driven Research in the Biological and Biomedical Sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), pp.1-3.

Leonelli, S., 2013. Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. *Bulletin of Science, Technology & Society*, *33*(1), pp.6-11.

Leonelli, S., 2016a. Data-Centric Biology: A Philosophical Study. University of Chicago Press.

Leonelli, S., 2016b. Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production. *Philosophical Transactions of the Royal Society: Part A*. 374: 20160122.

Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal*, *16*.

Leonelli, S., 2018. Rethinking Reproducibility as a Criterion for Research Quality. In *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise*. Emerald Publishing Limited.

Leonelli, S., 2019. *Data Governance is Key to Interpretation: Reconceptualizing Data in Data* Science. Harvard Data Science Review.

Leonelli, S., 2020. *Big Data and Scientific Research.* Stanford Encyclopaedia for Philosophy.

Leonelli, S. and Tempini, N., 2018. Where Health and Environment Meet: The Use of Invariant Parameters in Big Data Analysis, *Synthase*, pp.1-20.

Leonelli, S. and Tempini, N., 2020. *Data Journeys in the Sciences*, Cham: Springer International Publishing.

Leonelli, S., Lovell, B., Flemingm, L., Wheeler, B. and Williams, H., (forthcoming)*. From FAIR data to fair data use: Methodological data fairness in health-related social media research. Big Data and Society.*

Leslie, D., 2019. Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. The Alan Turing

Institute, e-Prints, Available at: https://ui.adsabs.harvard.edu/abs/2019arXiv190605684L [Accessed: 21.02.2021].

Li, C. and Sugimoto, S., 2017. Provenance Description of Metadata Vocabularies for the Long-Term Maintenance of Metadata. *Journal of Data and Information Science* 2(2), pp.41–55.

Lindstrom, M., 2016. *Small Data: The Tiny Clues that Uncover New Worlds.* St Martin's Press.

Lowrie, I., 2017. Algorithmic Rationality: Epistemology and Efficiency in the Data Sciences, *Big Data & Society*, 4(1), pp.1–13.

Lucivero, F., 2020. Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. *Science and engineering ethics*, *26*(2), pp.1009-1030.

Lucivero, F. and Prainsack, B., 2015. The Life Stylisation of Healthcare? 'Consumer Genomics' and Mobile Health as Technologies for Healthy Lifestyle. *Applied & translational genomics*, *4*, pp.44-49.

Lupton, D. & Michael, M. 2017. "For Me, the Biggest Benefit Is Being Ahead of the Game": The Use of Social Media in Health Work. *Social Media + Society*. 3(2).

**M**

Mackenzie, A., 2017. *Machine learners: Archaeology of a Data Practice*. Cambridge, Massachusetts: MIT Press, p.46.

Magalhães, J.C. and Couldry, N., 2020. Tech Giants are using this crisis to colonize the Welfare System. *Jacobin*, Available at: https://jacobinmag.com/2020/04/tech-giants-coronavirus-pandemic-welfare-surveillance [Accessed: 21.02.2021].

Marsh, S., 2019. One in Three Councils Using Algorithms to Make Welfare Decisions. *The Guardian*, 15 October 2019, sec. Society. Available at: http://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits [Accessed: 15.02.2021].

Mao, Y., Wang, D., Muller, M., Varshney, K.R., Baldini, I., Dugan, C. and Mojsilović, A., 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?. *Proceedings of the ACM on Human-Computer Interaction*, *3*(1), pp.1-23

Mauthner, N.S. and Doucet, A., 2008. 'Knowledge Once Divided Can Be Hard to Put Together Again': An Epistemological Critique of Collaborative and Team-Based Research Practices. *Sociology*, *42*(5), pp.971-985.

Mayo, D., 2018. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press.

Mayo, D. G. and Spanos, A., (eds.), 2009. Error and Inference, Cambridge: Cambridge University Press.

Mayo, D. G., and Spanos, A., 2011. Error Statistics. In: *Philosophy of Statistics*, (eds.) Prasanta S. Bandyopadhyay and Malcolm R. Forster, Handbook of the Philosophy of Science, Amsterdam, Holland, 7, pp.153–98.

Mayer-Schönberger, V., and Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

Mayernik, M. S., 2019. Metadata Accounts: Achieving Data and Evidence in Scientific Research. *Social Studies of Science* 49(5), pp.732–57.

Medina-Perea, I.A., Bates, J. and Cox, A., 2019. Using Data Journeys to Inform Research Design: Socio-Cultural Dynamics of Patient Data Flows in the UK Healthcare Sector. *iConference 2019 Proceedings*.

Meng, Xiao-Li. 2019. Data Science: An Artificial Ecosystem. Harvard Data Science Review 1(1).

Merelli, I., Pérez-Sánchez, H., Gesing, S. and D'Agostino, D., 2014. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed research international*, *2014*.

Mikroyannidis, A., Domingue, J., Phethean, C., Beeston, G. and Simperl, E., 2017. The European Data Science Academy: Bridging the Data Science Skills Gap with Open Courseware. In: Cape Town, South Africa.

Mikroyannidis, A., Domingue, J., Phethean, C., Beeston, G. and Simperl, E., 2018. Designing and Delivering a Curriculum for Data Science Education Across Europe. In: *Teaching and Learning in a Digital World*, (eds.) Michael E. Auer, David Guralnick, and Istvan Simonics, 540–50. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing.

Mirowski, P., 2018. The Future(s) of Open Science. *Social Studies of Science* 48(2), pp.171–203.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. and Lum, K., 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1).

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L., 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), pp.2053951716679679.

Mittelstadt, B., 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1(11), pp.501–7.

Morrison, M., 2014. *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press.

Moss, E., and Metcalf, J., 2020. *Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies*. New York, NY, USA: Data and Society Research Institute.

**N**

Noble, S.U., 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. Illustrated edition, New York: NYU Press.

Nowotny, H, Scott, P. and Gibbons, M., 2001. *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. Cambridge (UK): Polity Press.

Nuffield Council on Bioethics, 2015. The Collection, Linking and Use of Data in Biomedical Research and Health Care. *Ethical Issues.* Chapter 3: Public Interest, pp53-55, Available at: https://www.nuffieldbioethics.org/publications/biological-and-health-data [Accessed: 15.02.2021].

**O**

*OECD, 2014.* OECD Science, Technology and Industry Outlook 2014*.* p11*.* Available at: https://doi.org/10.1787/19991428 [Accessed: 21.02.2021].

OECD, 2017. Recommendations of the Council on Health Data Governance. OECD/LEGAL/0433, Available at: http://www.oecd.org/els/health-systems/health-data-governance.htm [Accessed: 15.02.2021].

O'Malley, M.A. and Soyer, O.S., 2012. The Roles of Integration in Molecular Systems Biology. Studies in History and Philosophy of Science Part C: Studies in *History and Philosophy of Biological and Biomedical Sciences*, *43*(1), pp.58-68.

O'Neil, C., 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown.

O'Neil, C. and Schutt, R., 2013. *Doing Data Science: Straight Talk from the Frontline*. 1st edition, Beijing, Sebastopol, CA: O'Reilly, pp.1-50.

**P**

Palmer, S., 2015. *Data Science for the C-Suite*. Digital Living Press, New York. Print.

Palmer, C.L., Thomer, A.K., Baker, K.S., Wickett, K.M., Hendrix, C.L., Rodman, A., Sigler, S. and Fouke, B.W., 2017. Site-Based Data Curation Based on Hot Spring Geobiology. *PloS one*, *12*(3), p.e0172090.

Pasquale, F., 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pickering, A., 2011. The Cybernetic Brain: Sketches of Another Future. Chicago, Ill: University of Chicago Press.

Pitcan, M., Marwick, A. E. and Boyd, D., 2018. Performing a Vanilla Self: Respectability Politics, Social Class, and the Digital World. *Journal of Computer-Mediated Communication* 23(3), pp.163–79.

Plantin, J.C., Lagoze, C., Edwards, P.N. and Sandvig, C., 2018. Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook. *New Media & Society* 20(1), pp.293–310.

Poell, T., Nieborg, D. and van Dijck, J., 2019. Platformisation. *Internet Policy Review*, *8*(4), pp.1-13.

Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton Univ. Press.

Porter, T.M. and de Chadarevian, S., 2018. Introduction: Scrutinizing the Data World, *Historical Studies in the Natural Sciences*, 48(5), pp.549–556.

Postigo, H., and O'Donnell, C., 2016. Chapter 20: The Sociotechnical Architecture of Information Networks. In: *The Handbook of Science and Technology Studies*, Cambridge, Massachusetts: MIT Press, pp.583–608.

Prainsack, B. and Buyx, A., 2017. *Solidarity in Biomedicine and Beyond*, Cambridge, UK: Cambridge University Press.

Prainsack, B. and Vayena, E., 2013. Beyond the Clinic: 'Direct-To-Consumer' Genomic Profiling Services and Pharmacogenomics. *Pharmacogenomics*, *14*(4), pp.403-412.

Prey, R., 2020. *The Performance Complex: Competition and Competitions in Social Life.* Stark, D. (ed.). Oxford University Press, pp. 241-258.

Price, D.J., 1963. *Little Science, Big Science*. New York: Columbia University Press.

**R**

Rangaswamy, N. and Arora, P., 2016. The Mobile Internet in The Wild and Every Day: Digital Leisure in The Slums of Urban India. *International Journal of Cultural Studies*, *19*(6), pp.611-626.

Rappert, B. and Selgelid, M.J., 2013. *On the Dual Uses of Science and Ethics Principles, Practices, and Prospects*, p. 390. ANU Press.

Rieder, G., and Simon, J., 2016. Datatrust: Or the Political Quest for Numerical Evidence and the Epistemologies of Big Data. *Big Data & Society* 3(1), pp.2053951716649398.

Rieder, G. and Simon, J., 2017. Big data: A New Empiricism and Its Epistemic and Socio-Political Consequences. In: *Berechenbarkeit der Welt?* Springer VS, Wiesbaden, pp.85-105.

Roussi, A., 2020. Resisting the Rise of Facial Recognition. *Nature 587,* pp.350-353, 18 November, Available at: https://www.nature.com/articles/d41586-020-03188-2 [Accessed 21.02.2021].

**S**

Sample, I., 2019. Maths and tech specialists need Hippocratic oath, says academic. *The Guardian*, 16 August, Available at: https://www.theguardian.com/science/2019/aug/16/ mathematicians-need-doctor-style-hippocratic-oath-says-academic-hannah-fry [Accessed: 21.02.2021].

Schutt, R., and O'Neil, C., 2013. Doing Data Science: Straight Talk from the Frontline, O'Reilly Media, Incorporated, Sebastopol.

Shapin, S., 1995. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago: University of Chicago Press.

Sharon, T., and Zandbergen, D., 2017. From Data Fetishism to Quantifying Selves: Self-Tracking Practices and the Other Values of Data. *New Media & Society* 19(11), pp.1695–1709.

Shaver, L., 2010. The Right to Science and Culture. *Wisconsin Law Review*, (1), pp.121-184.

Shaw, J., and Graham, M., 2017. An Informational Right to the City? Code, Content, Control, and the Urbanization of Information. *Antipode* 49(4), pp.907–27.

Shove, E., 2007. *The Design of Everyday Life*. Berg Publishers.

Shove, E., Watson, M., Hand, M. and Ingram, J., 2007. Reproducing Digital Photography. *The Design of Everyday Life*. New York: Berg.

Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L.H. and Merchant, R.M., 2017. Twitter as a Tool for Health Research: A Systematic Review. *American Journal of Public Health*, 107(1), pp.1-5.

Sloane, M. and Moss, E., 2019. AI's Social Sciences Deficit. *Nature Machine Intelligence*, *1*(8), pp.330-331.

Srnicek, N., 2016. *Platform Capitalism*. Cambridge, UK; Malden, MA: Polity Press.

Stark, L. and Hoffmann, A.L., 2019. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics* 1(1), pp.11052.

Starosielski, N., 2015. *The Undersea Network*. Duke University Press.

Stephens, M., 2013. Gender and the GeoWeb: Divisions in the Production of User-Generated Cartographic Information. *GeoJournal,* 78(6), pp.981–96.

Sterner, B., and Franz, N.M., 2017. Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data. *Biological Theory,* 12(2), pp.99–111.

Strasser, B., 2019. *Collecting Experiments: Making Big Data Biology.* Chicago: University of Chicago Press.

Strasser, B. J., and Edwards, P., 2015. *Open Access: Publishing, Commerce, and the Scientific Ethos*. Bern: Swiss Science and Innovation Council, SSIC Report 9(2015), Available at: https://citizensciences.net/wp-content/plugins/zotpress/lib/request/request.dl.php?api_user_id=424601&dlkey=BDEJXNRZ&content_type=application/pdf [Accessed: 15.02.2021].

Strathern, M., 2005. Anthropology and Interdisciplinarity. *Arts and Humanities in Higher Education* 4(2), pp.125–35.

Swan, A., and Brown, S., 2008. *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*. Key Perspectives, Consultants in scholarly information, Report to the JISC, Truro.

Symons, J. and Alvarado, R., 2016. Can We Trust Big Data? Applying Philosophy of Science to Software. *Big Data & Society*, 3(2), p.2053951716664747.

Symons, J. and Horner, J., 2014. Software Intensive Science, *Philosophy & Technology, 27*(3), pp. 461–477.

**T**

Taylor L., 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society,* 4(2), pp.205395171773633.

Tempini, N., 2015. Governing PatientsLikeMe: Information Production and Research Through an Open, Distributed, and Data-Based Social Media Network. *The Information Society*, *31*(2), pp.193-211.

Tempini, N., 2017. Till Data Do Us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures. *Information and Organization, 27*(4), pp.191–210.

Tenner, E., 2018. Who's Afraid of the Frightful Five? Monopoly and Culture in the Digital Age. *The Hedgehog Review* 20(1), pp.68–78.

Thayyil, N., 2018. Constructing Global Data: Automated Techniques in Ecological Monitoring, Precaution and Reification of Risk. *Big Data & Society, 5*(1), pp.2053951718779407.

The Economist, 2020. Winners form the Pandemic: Big Tech's Covid-19 Opportunity. *The Economist*, Leaders, 4 April 2020, Available at: https://www.economist.com/leaders/2020/04/04/big-techs-covid-19-opportunity [Accessed: 21.02.2021].

Thrift, N., 2005. *Knowing Capitalism*. London: Sage.

Ticona, J., 2016. Phones, but No Papers. Medium, Points, 30 November, Available at: https://points.datasociety.net/phones-but-no-papers-e580e824ed6 [Accessed: 21.02.2021].

Tufte, E., 1983. *The Visual Display of Quantitative Information*. Cheshire, Connecticut. Graphic Press.

Turnhout, E., Dewulf, A. and Hulme, M., 2016. What Does Policy-Relevant Global Environmental Knowledge Do? The Cases of Climate and Biodiversity. *Current Opinion in Environmental Sustainability*, 18, pp.65-72.

Tutton, R., 2016. Personal Genomics and its Sociotechnical Transformations. In: *Genomics and Society*, Academic Press, pp. 1-20.

**U**

UN News, 2016. UN Agency and Google Collaborate on Satellite Data Tools to Manage Natural Resources. Available at: https://news.un.org/en/story/2016/04/526802-un-agency-and-google-collaborate-satellite-data-tools-manage-natural-resources [Accessed: 15.02.2021].

United Nations, 2018. Sustainable Development Goals Report 2018. New York, NY, USA.

## V

Vallor, S., 2018. Technology and the Virtues: *A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

Van Dijck, J., Poell, T. and De Waal, M., 2018. *The Platform Society: Public Values in a Connective World*. Oxford University Press.

Van Dijck, J., 2020. Governing Digital Societies: Private Platforms, Public Values. *Computer Law & Security Review*, *36*, p.105377.

Van Doorn, N. and Badger, A., 2020. Platform Capitalism's Hidden Abode: Producing Data Assets in the Gig Economy. *Antipode*, *52*(5), pp.1475-1495.

Van Horn, J.D. and Toga, A.W., 2009. Is It Time to Re-Prioritize Neuroimaging Databases and Digital Repositories?. *NeuroImage*, 47(4), pp.1720–34.

Vayena, E. and Prainsack, B., 2013. Regulating Genomics: Time for a Broader Vision. *Science Translational Medicine*, 5(198), p.198ed12.

Vayena, E. and Tasioulas, J., 2015. " We the Scientists": A Human Right to Citizen Science. *Philosophy & Technology*, 28(3), p.479.

Veale, M. and Binns, R., 2017. Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data. *Big Data & Society*, 4(2).

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M. and Fuso, N., 2020. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications* 11(1), pp.233.

Von der Leyen, U., 2019. Political Guidelines for the Next European Commission 2019-2024. *European Commission, Brussels. PE*, *658*, pp.18.

Van der Vlist, F.N., 2016. Accounting for the Social: Investigating Commensuration and Big Data Practices at Facebook. *Big Data & Society*, *3*(1), pp.2053951716631365.

Von Oertzen, C., 2018. Datafication and Spatial Visualization in Nineteenth-Century Census Statistics. *Historical Studies in the Natural Sciences*, *48*(5), pp.568-580.

## W

Weingart, P., and Padberg. B., 2014. University Experiments in Interdisciplinarity: Obstacles and Opportunities. *University Experiments in Interdisciplinarity*. Bielefeld: Transcript.

Wessels, B., Finn, R.L., Wadhwa, K. and Sveinsdottir, T., 2017. *Open data and the Knowledge Society*. Amsterdam University Press.

Wijmenga, C., 2019. Our Opportunities Lie in Data. Presented at the Opening of Academic year 2019-20, University of Groningen, The Netherlands, September 2, Available at: https://www.rug.nl/about-ug/latest-news/news/archief2019/nieuwsberichten/0823-speech-rector-magnificus-september-2019.pdf [Accessed: 15.02.2021].

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, *3*(1), pp.1-9.

Wilsdon, J., Bar-ilan, J., Frodeman, R., Lex, E., Peters, I. and Wouters, P.F., 2017. Next-Generation Metrics: Reponsible Metrics and Evaluation for Open Science. Report of the European Commission Expert Group on Altmetrics.

WHO, 2019. Global Tuberculosis Report 2019. World Health Organisation.

Wolff, J. and Atallah, N., 2020. *Early GDPR Penalties: Analysis of Implementation and Fines Through May 2020*. SSRN Paper ID 3748837. Rochester, NY: Social Science Research Network.

Wouters, P., Beaulieu, A., Scharnhorst, A., and Wyatt, S., 2013. *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*. Cambridge, Massachusetts: MIT Press.

## Y

Yeung K., 2018. Algorithmic Regulation: A Critical Interrogation. Regulation & Governance 12(4), pp.505–523.

## Z

Zhang, A.X., Muller, M. and Wang, D., 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *ArXiv:2001.06684 [Cs, Stat]*, April.

Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A. and Nelson, A., 2017. Ten Simple Rules for Responsible Big Data Research. *PLoS Computational Biology*, 13(3), pp.e1005399-e1005399.

Zuboff, S., 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st ed. New York: PublicAffairs.

Zwitter, A. 2014. Big Data Ethics. *Big Data & Society* 1 (2): 2053951714559253. https://doi.org/10.1177/2053951714559253.

Zwitter, A. J., O. J. Gstrein, and E. Yap. 2020. Digital Identity and the Blockchain: Universal Identity Management and the Concept of the "Self-Sovereign" Individual. *Frontiers in Blockchain* 3. https://doi.org/10.3389/fbloc.2020.00026.