# The Sciences of Data – Moving Towards a Comprehensive Systems Perspective

Victor Garcia, Claus Horn and Thomas Ott

**Abstract** Data science's rapid development in a dynamically growing data environment endows it with unique characteristics among scientific disciplines, juxtaposing challenges typically encountered in theoretical as well as empirical sciences. This raises questions as to the identification of the most pressing problems for data science, as well as to what constitutes its theoretical foundations. In this contribution, we first describe data science from the perspective of philosophy of science. We argue that the current mode of development of data science is adequately described by what we term the *differentiational-expansionist* mode. This leads us to conclude that data science concerns the acquisition of scientific theories relating to the application of methods, workflows and algorithms that generate value for users – which we term the *integrative view*. This definition emphasizes the interdependent nature of human and algorithmic elements in complex data workflows. We then offer four challenges for the future of the field. We conclude that since full control of entire data

Victor Garcia · Claus Horn · Thomas Ott
Zurich University of Applied Sciences ZHAW, Institute of Computational Life Sciences, Im Schloss 1, 8820 Wädenswil
✉ gara@zhaw.ch
✉ horc@zhaw.ch
✉ ottt@zhaw.ch

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 8, No. 2, 2022

DOI: 10.5445/IR/1000150241
ISSN 2363-9881

workflows is unfeasible, attention should be redirected towards the creation of an infrastructure by which data workflows will self-organize in a useful manner.

# 1 Introduction

In the past decades, profound advances in technology have brought about shifts in how and at what scale information is processed, thereby giving rise to data science. Cheaper and more accurate sensors have facilitated the gathering of vast amounts of data often termed a "data deluge" (Cao, 2018; Manyika et al., 2011; Bell et al., 2009). Stark reductions in the cost of computation have allowed for those data's exploration and analysis and has spawned new avenues for knowledge acquisition from data (De Roure, 2010).

In the wake of these developments, data science has emerged as a unique juxtaposition of elements from various disciplines (Braschler et al., 2019b). Dhar defines data science as *"the science of identifying generalizable knowledge from data"* (Dhar, 2013), in other words, as the science (Heilbron, 2003) of learning from data. In the following, we will use this definition of data science as the *standard definition of data science.* According to Dhar, data science draws from heterologous and unstructured data and therefore utilizes methods that are adequate for the data sizes to allow for knowledge discovery (Dhar, 2013).

In contrast to the hypothesis-testing machinery at the heart of statistics, data science is geared towards discovery of patterns in large amounts of data and intends to find *interesting* and *robust* patterns that satisfy the data (Leek, 2013; Dhar, 2013). Here, with "interesting data", Dhar means data that are *actionable* (Davenport and Patil, 2012) and unexpected (Leek, 2013; Dhar, 2013). Actionable data hold predictive power, where the outcome expected from an action can be relied upon to appear with high confidence (Popper, 2014). Robust patterns are patterns that are expected to reoccur in the future (Perlich et al., 2003; Dhar, 2011; Meinshausen, 2007). Crucially, data science employs methods that allow the researcher to analyse data without strong *apriori* assumptions about the relationships between the variables (Dhar, 2013).

Some of these features are in line with a prescient analysis by De Roure (De Roure, 2010), which in 2010 heralded the advent of a new mode of conducting science. The digitization of almost all data measured by researchers has brought about an automation of scientific practice that he termed e-Science. He surmised that this shift should change the relationship between hypotheses

and data. When data were scarce, hypothesis formation was the dominant partner in this relationship: Hypotheses directed researchers' search for data. The abundance and ubiquity of data shifted dominance towards data: Data now guide research towards the most plausible hypotheses.

Thus, data science is not just the statistics of large data sets (Dhar, 2013; Braschler et al., 2019a). Two main aspects that separate it from statistics are, first, that its central focus lies in pattern discovery, that is, in the discovery of hypotheses that are compatible with the data and perhaps satisfy additional conditions (Dhar, 2013). Second, that especially those parts of machine learning that are concerned with exploratory data analysis through unsupervised learning deviate from classical computational statistics (Bishop, 2006; Friedman, 1998).

As data science has emerged and established itself as a scientific discipline in its own right, it has embarked on a fruitful process of maturation, but particularly of differentiation and expansion. By maturation we mean a process by which a body of theory is assembled and continually refined (Rheinberger, 2008). Concepts are sharpened, theory made more rigorous and consistent, and the understanding of problems is deepened by novel methods of analysis. This mode of scientific development is described by Kuhn as a period of "normalcy" under an existing paradigm (Kuhn, 1962). The differentiation and expansion of a scientific discipline denotes two complementary processes that refer to one another (Rheinberger, 2008). The differentiational aspect refers to the emergence of novel subbranches of the discipline around practical problems of particular interest and depth. As this differentiation occurs, the scientific discipline integrates these objects of study into its body of knowledge, which is understood as an expansion into novel territory by the discipline. This second differentiational-expansional mode of scientific development has been described in detail by Fleck (Fleck, 2019), and more generally by Bachelard (Bachelard and Lepenies, 2016; Bachelard, 1988, 2015), and has been termed the "process character" of science (Rheinberger, 2008). Importantly, it is driven by the encounter with practical problems arising from the application of commonplace instruments and methods within a discipline to some initially fringe case.

Data science possesses the distinct peculiarity that it is the differentiational-expansional mode that dominates its development today, and with it, the problems that arise from data science's application in real-world contexts. For instance, the rapid adoption of data science methods in many business domains has unearthed many new challenges that have fed back into its theoretical

development (Cao, 2018), such as self-driving cars (Saha and De, 2022), video games (Silver et al., 2017) and protein folding (Jumper et al., 2021). In this –very strict– sense, we observe a transition towards practicality within data science and relative to other sciences.

The shift towards practicality that follows from the particular expansion-dominated course of data science's development results in a tension. This is a tension between data science's unifying thread at its inception (Dhar, 2013) with today's reality of everyday data science. To use a geographical metaphor, the process of differentiation has spread the activity in data science across such a vast territory that a reexamination of what unites these colonies is warranted.

This current state of affairs has considerable disadvantages. For instance, it can cause confusion about what actually constitutes a data science project, which entails a risk of resource misallocation. In national research funding bodies, as well as enterprises, resources may be directed away from the resolution of pressing problems because these are not recognized as pertaining to the field of data science. Furthermore, individuals or organizations engaged in data science may not be recognized as doing so. Most importantly, a discipline's self-image affects the vision of its future development, and where the greatest effort should collectively be directed. To circumvent these issues, it is worthwhile to ask whether the current state may not be better encompassed by an alternative conception of data science.

To address these issues and to accommodate for these novel developments within it, here we propose to extend the definition of data science as follows: Data science is the study of interdependent data workflows embedded in real-world processes. We argue that this outline of data science allows us to resolve some of the tension we diagnose as arising from its current understanding. It shifts the emphasis of data science into the domain in which most practical activity is located. It permits to subsume a wide range of activities under a unifying paradigm. It redirects our view towards the next necessary steps for data science's development. And lastly, it helps to clear the view for the main challenges ahead and assists us in the preparation of students to future work environments.

In the following, we will give a more detailed and founded account of our proposition for the refocusing of data science. We will additionally present a set of challenges to achieve more beneficial and useful digital products via the utilization of data science.

## 2 Data Science's Object of Study: The Interdependent Data Workflow Embedded in Real-World Processes

In this section, we recapitulate in greater detail why we believe that data science's scope ought to be expanded and embed these arguments into the broader context of the history and philosophy of science.

We first begin by describing some of the pertinent characteristics of data science's rapid development as we perceive them. Subsequently, we place data science in the space that is spanned between two established descriptions of the process of scientific evolution. We proceed to draw conclusions with regards to a useful definition of data science on the basis of this placement.

The main source of phenomena that may act as midwives to new branches of data science today stem from practical problems from industry and related contexts (Cao, 2018). This is a development that has also been observed in other scientific disciplines. In physics, for instance, the development of statistical mechanics was influenced by analysis of movements of suspended particles in a liquid – termed Brownian motion (Rigden, 2005). Similarly, the development of special relativity was heavily influenced by patent applications on the synchronization of clocks by means of electrical signals (Galison et al., 2004).

This cross-pollination from practical problems has been studied in depth in the past. In the philosophy of science of both Gaston Bachelard and Ludwig Fleck, practical problems act as crystallization points, as seeds, for new branches of science (Rheinberger, 2008, 2006a; Bachelard, 2015). For Bachelard, there does not exist a theoretical system of philosophy that may encompass the breadth of phenomena studied under science (Bachelard, 1988). This process is due to two distinct factors acting in unison. The first factor, he argues, is the direct result of the inherent intertwinedness of experience and thought, reality and reason: These categories refer to each other and thus coexist in a semantic circularity (Bachelard, 1988). This intertwinedness precludes the clean separation of these aspects without contaminating one partner with remnants of the other (Rheinberger, 2008).

The systems analyzed in science share the same property, which Bachelard calls *complementarity*. Complementary indicates that a system cannot be reduced into individual qualities or properties without suffering a loss to its full description. This vantage point leads Bachelard to conclude that the appropriate object of study of philosophy of science is "phenomenotechnology"

(Bachelard, 1988), a self-recursive system of experimental instruments and methods that generates knowledge that in turn feeds back into the preconditions of its further development.

Within the framework of phenomenotechnology, scientific insights are not discovered in a Platonic or Cartesian sense. Instead, scientific facts and insights are created, or "realized" within a phenomenotechnological system (Rheinberger, 2008). They are forced into existence by the coupling of the technical assortment of instruments with the thought process.

The second factor generating differentiation within scientific disciplines, Bachelard asserts, is that by necessity scientific insights must carry a social component and will therefore be contingent on their past developmental trajectory (Bachelard, 1988). This social constitution of science is also mirrored in science's tendency to fragment into separate fields of work. This differentiation is the necessary consequence of the character of experimentation, since it revolves around concrete and particular issues and concerns. This gives rise to an "epistemology of detail" (Bachelard, 2015). It is this process of realization that manifests novel scientific facts, and lies at the root of new disciplines.

Fleckian sociology of science mirrors these observations by Bachelard. Epistemology cannot be built from the individuum up, but is a holistic, social and historically determined process in which single individuals participate (Fleck, 2019). The experience of the scientists as a *community* is indispensable to obtain an accurate picture of the reality of scientific activity. This experience relates to the everyday practice of science making, and not to its written and digital traces in the literature (Rheinberger, 2008). It is thus a profoundly cultural phenomenon, and driven by contact with the object of study (Fleck, 2019; Rheinberger, 2008).

Thus, according to Fleck, science proceeds "like the river that forms its riverbed" (Fleck, 2019) – in a manner similar to the process emphasized by Bachelard. The process is contingent on the previously taken trajectory, and thus deeply historical (Rheinberger, 2008). Further developments depend on the order of previously realized steps, and may thus lead to bifurcations in the research interests of the scientific community, differentiating it further. Revisiting the river metaphore, the stream is in a process of perpetual division into substreams. This "inner historicity" (Rheinberger, 2008) endows the process with a unique quality that is independent of the individual, and emerges as an entity onto its own.

In contrast to Bachelard and Fleck, Pierre Duhem has emphasized processes that enhance the internal conceptual order and consistency of bodies of theory (Duhem, 1892), which we term the maturation mode of scientific development and which we review in the following.

According to Duhem, the mind comes into contact with the world, that is physics, all the time. The resulting observations form a body or a set of experimental facts. These facts cannot stand alone for themselves. Conceptions attached to each of them are going to be set into some sort of relation to each other by the mind. Therefore, through induction, the mind constructs experimental laws out of these facts. To Duhem, the purpose of this inductive process is plain:

> *" Theoretical science has as its aim to relieve the memory and to assist it in retaining more easily the multitude of experimental laws."* Duhem (1996, p.2)

The mathematical theory that ultimately describes the processes of interest is therefore a *form of summarization and classifications* of experimental laws. This solely is the aim for the construction of a mathematically formulated theory.

Thus, according to Duhem, a theory does not need to relate directly with the world as it is. It is a relative entity, a convention on how to represent the world. That is why physicists should look for a "systematic coordination" of laws in a theory, and not for explanations: Physicists "wish to clarify laws, not to reveal causes." Duhem (1996, p.17)

Hence, Duhem sees theory-building as a kind of housekeeping, a consolidation and tidying of the preexisting concepts associated with experimental facts. It is in this sense that Duhem's mode of scientific evolution can also be interpreted to give rise to a *maturation of science*: By enhancing a body of theory's internal coherency and facilitating its applicability to the user.

Similarly to this notion of a maturation process, the philosopher of science Thomas Kuhn describes a mode of development of science that predominates in between phases that he terms *paradigm shifts* (Kuhn, 1962). According to Kuhn, science does not develop in the form of a gradual accumulation of knowledge alone. Its development is discontinuous, characterized by interspersed, profound disruptions and revolutions. Revolutions establish new scientific paradigms around which scientific communities organize. These paradigms must be open enough to point to a range of problems that can be expected to be addressed or solved by means of the tools contained in the paradigm. However, once

established, the scientific community will begin to apply the newly created framework to open problems, cementing the paradigm and strengthening internal coherency – that is, maturating the paradigm.

These observations lead us to conclude that there exist at least two modes of scientific development that we refer to as the *differentiation-expansion mode* and the *maturation mode* of the evolution of science. The differentiation-expansion process can be understood as being shaped by a cultural process and by real-world phenomena. Duhem and Kuhn, on the other hand, identify efforts within scientific developments that are aimed at generating an internal consistency within the body of ideas that represent science, thereby maturing them.

These prerequisites allow us to place data science within a framework of established descriptions of the nature of scientific enquiry. Although the differentiational-expansional and the maturation modes are no logical opposites, they nevertheless emphasize distinct, and to a degree mutually exclusive aspects of development: One is oriented towards integration of novel phenomena into the existing body of knowledge, and the other is oriented towards the integration of existing knowledge into a conceptual superstructure. In the following, we will use these as reference points for our reasoning about the present challenges of data science.

Here, we argue that current data science is engaged in an intense process of differentiation-expansion. Data science today is awash with problems from practice that may serve as crystallization points for new branches (Cao, 2018). This is evidenced by the cross-pollination between academic data science and applied data science in industry, in particular where it concerns automation of the scientific workflow (Foster and Kesselmann, 1999; De Roure, 2010).

Industrial and business applications are providing a steady stream of problems and challenges that require ingenious approaches for their resolution (Cao, 2018; De Roure, 2010). For instance, the startup GainForest based in Zurich, Switzerland, is utilizing data science approaches to monitor sustainable land use – for example by validating reforestation efforts – by means of satellite, drone and field imagery (Dao, 2022; Masterson, 2021). This business application poses new, very specific problems on how to evaluate images for sustainable land usage that are blurred or contorted by high cloud coverage and sparse imagery. Their solutions feed back into the existing knowledge.

That a novel scientific discipline should expand is inherent in its natural developmental course (Niiniluoto, 2019; Schickore, 2018). However, data

science's largely differentiational-expansionally driven mode of development constitutes an idiosyncratic evolutionary path. The requirements imposed on practitioners of data science by real-world problems overshadow the need for internal theoretical coherency. Indeed, the collections of challenges and problems in the literature that are readily subsumed under the umbrella term of data science is growing steadily (Davenport and Patil, 2012; Braschler et al., 2019b; Cao, 2018).

Summarizing, our review of some schools of thought about the mode of evolution of scientific disciplines leads us to believe that data science is engaged in an intense differentiation-expansionist mode of evolution. This mode gives rise to the emergence of novel branches of data science around particular problem stems from a vast array of applications –the crystallization points–, ranging from industrial applications to academic projects, that cross-pollinate. This raises the question of whether these disparate areas are still united by a common theme, a common element contained within all these distinct efforts: What is it that makes these activities precisely data science?

The Bachelardian and Fleckian views of science suggest a fundamental reflexivity in data science that may serve as starting point for the identification of a unifying theme: The knowledge required to devise and operate an analytical instrument appropriately is precisely the knowledge that these instruments are meant to uncover (Wind, 2000). This circularity forces the epistemic science into a mode of perpetual self-adjustment, and is the driving force that governs its evolution (Rheinberger, 2008, 2006b). In data science, this translates to a perpetual feedback-loop, where the knowledge contained in the data is needed to adequately design the methods utilized to acquire that knowledge in the first place.

For instance, in physics, the prevailing understanding of the physical laws that govern microscopic environments enter the design of the instruments for studying these same environments, such as for example quantum mechanics is necessary to conceive of the scanning tunneling microscope. These devices then sense phenomena within their respective environments, generating data that may spark a reinterpretation of the prevailing understanding, and thus give rise to novel devices that accommodate for such new understanding. Analogously, in data science, "sensing" devices are deployed to extract knowledge from data that are based on core assumptions. As vast swaths of data are explored by means of these

devices, the limitations of their underlying assumptions are gradually revealed, giving rise to novel, better adapted knowledge extraction devices.

Here, we argue that it is this circular self-referentiality that lies at the heart of data science. We argue that this is not merely the consequence of disparate conventions, but that it points to novel territory of thought and intellectual engagement with a fertile landscape of new challenges. Thus, we propose that an appropriate object of study in data science is the question of how real-world data workflows arise from interdependent combinations of algorithmic computation and thought. We term this the *integrative view of data science*.

This perspective has two major ramifications on our understanding of data science. First, it continues to reflect the idea that data science is an applied science discipline; that is, it examines urgent or awkward problems (for instance, "genuine problems in real industrial contexts for which someone will pay for the development of a solution", see Brodie, 2019a) that arise from everyday life (Braschler et al., 2019b). It thus retains key aspects of the standard definition of data science.

Second, and crucially, it enables us to expand the standard definition of data science in a coherent fashion, integrating further notions and objects of study. Studies in the wild of data and applications require an understanding of that wild environment, i.e., the context in which the data were created, processed, or used. Thus, the object of the science is not only the data themselves, but how we come into contact with them, how they are created, manipulated for knowledge generation and how they affect our perspective towards them. Thus, our perspective naturally expands the view of data science as a science that studies how all components in the production process of the digital product interact to derive value from data.

Such a perspective does considerably increase the complexity of the object of study. While the machine learning component is currently often viewed as the heart of a workflow in data science projects, this augmented complexity requires a shifting of the focus to a more comprehensive understanding of data science.

Based on these assumptions, in the following, we will formulate four challenges regarding the development of data science from this new vantage point.

# 3 Towards a Comprehensive Complex Systems Perspective: Four Challenges on Current Data Science

## 3.1 Challenge I: Perform Data Workflow and Pipeline Engineering at a Formal Level

Data workflows are connected to and embedded in many real world applications, and practicability requires that they become more and more automated (see e.g. Masterson, 2021; De Roure, 2010). To diminish the level of complexity experienced by humans, data and data workflows should be represented in abstract ways. For instance, automated workflows consist of algorithmic building blocks such as automated application programming interfaces (APIs). These elements conceal an ever larger amount of technicality and detail to allow for an abstract treatment of the element and thus to assure the element's function.

In the future, data science practice will thus be less about manually preparing data and training algorithms, and more about workflow design tasks at a high level of abstraction. Data science will become the art of adapting and optimizing schemes of data workflows for specific applications.

According to what we call the *integrative view* of the data science workflow (illustrated in Figure 1) at a very high-level of abstraction the workflow consists of a set of elements or protagonists whose omission greatly diminishes the workflows's ability to fulfill its purpose. Furthermore, also societal and environmental boundary conditions and requirements such as regulations ought to be accounted for.

To achieve these modeling goals, researchers may draw from a tradition of well-established modeling concepts and approaches (Zeigler et al., 2019; Zeigler, 1970, 1971) and enrich them with own solutions that account for the particularities of the challenge.

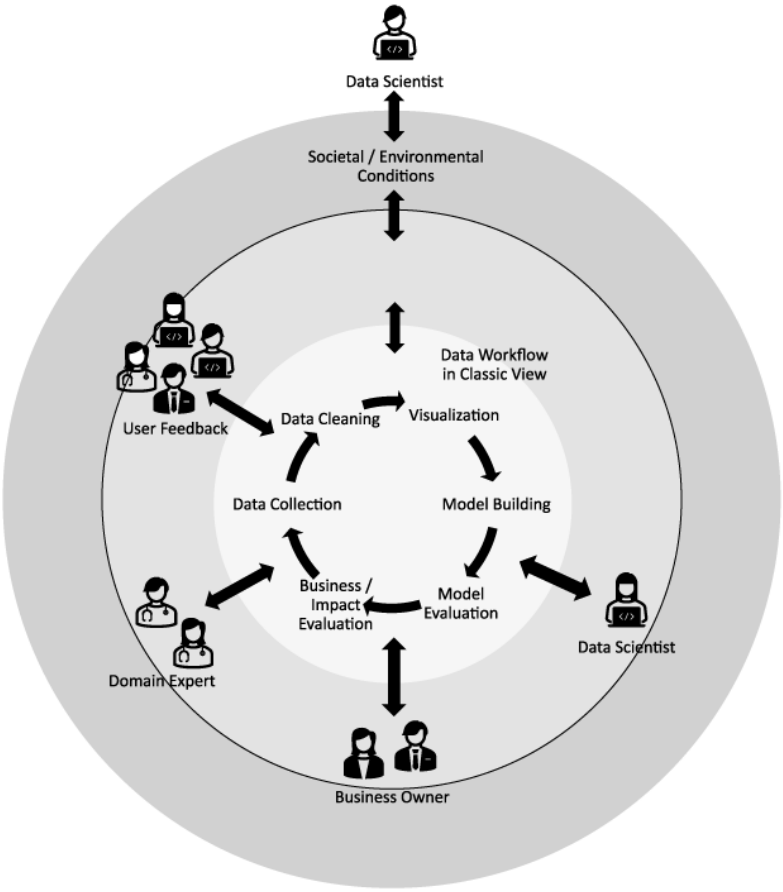


**Figure 1:** Integrative view of the data science workflow. Users, domain experts, data scientists and business owners influence the outcome of the data science process with their knowledge, motivations, and biases. Their roles should be explicitly addressed in a view that goes beyond the narrow view of the "inner" (classic) workflow. In a fully developed data science perspective, the (data) scientist's task is thus not only to develop data workflows, but also to reflect on the surrounding spheres of influence and underlying conditions.

## 3.2 Challenge II: Integrate the "Human in the Loop" Perspective into the Workflow Design

A recently published survey gives a series of instances where humans-in-the-loop are leveraged to improve utility (Wu et al., 2022). For instance, within object detection, humans are still employed to correct annotations proposed by a machine learning detector in images in an approach termed *interactive object detection*. In this approach, human expertise is utilized to annotate images with the maximum predicted annotation cost to achieve the greatest gain for the overall object detection pipeline (Yao et al., 2012).

The survey therefore suggests that humans remain essential agents and participants in many data science workflows. As the data workflow is perfected, they actively learn and develop their skills in unison with machine learning algorithms. This "human in the loop" perspective changes the way workflows need to be framed, designed and analysed. In this framework, a data workflow should no longer be considered as a sequence that is processed in a strictly linear temporal order. Rather, the workflow may be more appropriately described as a complex, self-organizing network of value adding processes that may be operating simultaneously or in sequence. Singular process nodes may receive data from or export data to other singular processes, sets of processes or the whole network itself.

The challenge for data science will lie in adopting the notion of data workflows as self-organizing, adaptive complex systems and how to account for the role of humans within them. In fact, within such a network, the human data scientist will have to fulfil different roles at different levels. On the one hand, data scientists must be seen as part of the loop interacting with the inner workflow (see Figure 1). On the other hand, they must assume the role of a meta-level supervisor who critically reflects on and defines their own function in the workflow.

For instance, during the Covid-19 pandemic, scientists from ETH Zurich created a web-application and dashboard to visualize the current and predict the future epidemiological status in Switzerland (ETH Zurich, 2020). Parts of this application were later taken up by Swiss health authorities and their results made publicly available in governments websites. In these applications, epidemiological models were fitted to current Covid-19 case data and utilized to estimate the effective reproductive rates of the virus, amongst other relevant information. In this workflow network, humans occupied different roles. On the

one hand, they perform non-automatizable tasks necessary for the pipeline to work, such as taking the test swabs from patients, sending the paperwork to health authorities, setting up a model and fitting it to data to obtain a prediction. On the other hand, they also take on roles as workflow soundness reviewers –meta-level supervisors–, verifying that the chain of information transmission is working, that measurement instruments measure what they intend to, and that reporting delays do not unduly bias model estimates. They run plausibility checks on the estimates, and constantly monitor whether some fundamental aspect of transmissibility has changed, such as for instance immune evasiveness (Salathé et al., 2020). If any of that is the case, model parameters or architecture are adapted to the new reality, as well as other components of the workflow. In summary, in such a workflow network, humans occupy different crucial roles in data processing at different levels of abstraction. Many roles may require similar qualifications, such that a single individual may play all of them if desired. However, other roles may require more specialized knowledge and be unique to the workflow network.

### 3.3 Challenge III: Create New Disciplinary Branches at the Intersection of Domains

As data science evolves along the mode of differentiation and expansion, new roles for data scientists will emerge at the seed points of emerging sub-disciplines. These novel, more specialized disciplines will most likely emerge at the intersection between different domains and data science. It is at these intersections where unique problems present themselves that may be initially addressed with the tool set of data science. For instance, the intersection of data science and epidemiology has given rise to the field of *digital epidemiology* (Salathé et al., 2012). Data science methods may be deployed for early disease detection with Twitter-data (Bodnar and Salathé, 2013), harnessing and leveraging an entirely novel data type for epidemiological purposes (Paul et al., 2014). In such novel applications, concepts and theories of the domain will most likely merge with methods and tools from data science, entering a self-adjusting feedback loop and spawning a novel discipline.

The challenge will lie in providing a robust tool set for a broad set of potential researchable phenomena. We expect that most of these new data science

disciplines will integrate knowledge on the statistics of adaptive complex systems (Jensen, 2009; Thurner et al., 2018; Ladyman et al., 2013). This includes challenges such as how to analyze systems that are governed by non-linear feedback loops and in which interactions between components are strong and cannot be neglected, how to account for fluctuations, how to derive probability distributions that are broad and contain information about extreme events and how to account for differing scales of time evolution of single system components (Jensen, 2009).

New job profiles will develop at these intersections, e.g. the life science data scientists, health data scientists, environmental data scientists, digital epidemiologists, business data scientists, etc. Data science will transform into a field of multiple sciences of data with with their own blends of scientific theories and methods.

We expect that such approaches will prove most fruitful where they can leverage information theory (Ladyman et al., 2013), since it provides a common denominator between data science and complexity theory and thus naturally bridges both domains. Data science will then be utilized to perfect our analysis of complex data ("data produced by complex systems") by means of complex methods ("the methods appropriate to studying those systems") (Ladyman et al., 2013). How to recognize how complex a system is, and indeed, whether a system is complex, remains a subject of investigation (Ladyman et al., 2013). Nonetheless, significant research advances in the field of statistics of complex systems are being achieved (Shalizi, 2006; Gerlach and Altmann, 2019).

## 3.4 Challenge IV: Design Self-Organising Workflow Networks That Evolve into Stable States

Human top-down supervision of a workflow network will most likely be insufficient to ensure its seamless functioning over time. As complexity increases, the feedback loops and dependencies between parts of a data workflow and its environment will undermine the notion of complete predictability and thus, of top-down control of such data workflow systems (Thurner et al., 2018).

Thus, a major milestone to achieve this challenge will be to find viable descriptions of data workflow systems with humans in the loop (Wu et al., 2022; Deng et al., 2020) as well as supervisors, that are embedded in broader techno-social

contexts (Martin Jr et al., 2020). In our view, a possible route to approaching this challenge is to interpret data workflow systems as complex systems themselves (Ladyman et al., 2013), and specifically, networks (Newman, 2003). In other words, the challenge will consist in integrating our improving understanding of the behavior of complex systems that contain humans as agents with data science and its tool set.

On this major issue of appropriate modeling, we believe that much can be learned from research on techno-social systems from a networks standpoint, and in particular from economic systems modeling (Schweitzer et al., 2009) or research and development networks (König et al., 2012, 2019, 2011). These approaches are also employed to characterize techno-social systems, and particularly, to predict their behavior (Barabási and Albert, 1999; Newman, 2003; Barrat et al., 2008; Vespignani, 2009).

A particular difficulty is the accounting for the *reflexive aspect* of interdependent data workflow systems, that is, the circularity between cause and effect that arises from meta-level human agents – that are also localized in the "loop" – influencing the data workflow architecture on the grounds of its behavior. In other words, humans that form part of the data workflow will sometimes analyze its behavior in the role of a supervisor, and then decide to implement architectural changes to the network. Perhaps this aspect may be partially addressed by implementing insights from multiscale-networks (Vespignani, 2009), which evolve at multiple time and length scales. Supervisory influences may then be thought of as acting at longer time scales, and to only exert a structural effect after large amounts of experience have been accumulated.

Humans, algorithms and data are increasingly becoming interdependent parts of an overall system, interacting within and across a hierarchy of different levels of emergent functionalities and abstraction (Vespignani, 2009). A key element of their future management will be their design and how this design ensures a useful and fair functioning of the network.

To address this crucial challenge, a possible route may be again to draw from other emerging disciplines such as complex systems theory. Complex networks are known to self-organize (Barabási and Albert, 1999; Newman et al., 2006; Newman, 2003; Barrat et al., 2008), giving rise to patterns that are often unintended by and independent of their planners and engineers. Thus, an improved description of how individual components are linked and how the system is embedded within its social context will magnify our capability to

realistically simulate possible evolutionary pathways of these systems. Time-tested methods such as those developed by Y. Sinai and colleagues in the field of ergodic theory, such as for example the Sinai-Kolmogorov-complexity measure (Sinai, 2009), may give crucial insights into the qualitative time-evolution of such systems (Koralov and Sinai, 2007; Sinai, 2009; Cornfeld et al., 2012). A growing body of theory and experience in the field of experimental algorithmics may be deployed to investigate how these systems do operate in reality (McGeoch, 2012). Major efforts are also underway to minimize the risk that data scientific products generate outcomes that deviate from ethical norms in what is termed the *alignment problem* (Christian, 2020). Recently, Martin et al. proposed a three sub-component system to address this issue, notably implementing complex systems approaches (Martin Jr et al., 2020).

In summary, we believe that the use of self-organization and evolution principles will lead to novel ways of organizing these systems, at both the single algorithm (e.g. Raghavan and Thomson, 2019) and system or network levels (Gross and Blasius, 2008; Pfeifer et al., 2007).

## 4 Conclusion

Based on the observation that data science as a scientific discipline is confronted with open questions regarding its practice and its theoretical foundation, we have argued for an integrative view of what we consider to be a central object of study, namely data workflows.

This object of study, is understood to be reflexive and self-referential: It acts upon itself and it contains itself as an object of study. This aspect is reflected in the role of the data scientists as they, like other human agents, become an integral part of the data workflow while simultaneously being called upon to contemplate their role from outside of the system.

Our line of argumentation comes with some caveats. We have reviewed a series of schools of thought in the philosophy of science that concern the evolution of science, and found that one in particular, appears to match our perception of the current mode of data science's evolution. We then proceed to give an account of what the consequences for data science would be based on such a correspondence. However, it is possible that data science may represent a completely novel paradigm of scientific evolution, not encompassed by any previously developed school of thought.

Nevertheless, we believe that the integrative view can serve the purpose of a sharpening and refinement as well as augmentation of the role of data science. Its adoption would inevitably increase the complexity of data science as a discipline. This has led us to derive four challenges about the development of data science, essentially expressing two lines of thought.

First, an increase in complexity of data workflows requires means to cope with it. In this respect, we have identified a) a need for a more abstract view of the elements of a data workflow (Challenge I) and b) the necessity to guide a general tendency toward diversification of disciplines towards the intersection between domains and data science (Challenge III).

Second, the increase in complexity is not just gradual. It is due to a qualitative transition that leads to considering data workflows as complex networks that include the interdependent activities of human and algorithmic protagonists. This must lead to c) a rethinking of the design of data workflows (Challenge II) and d) in the long-term to new design principles (Challenge IV).

Notably, techno-societal and biologically inspired bottom-up principles of development or learning in networks offer a complement to top-down design principles. These include rules of pattern formation through self-organization or local learning rules in neuronal networks. Likewise, phenomena of formation, cooperation or specialization of subsystems in the structure of an entire organism do not follow explicit blueprints in nature, but unfold on the basis of typically locally mediated mechanisms. The homeostasis of such systems of systems often requires a dynamic rebalancing of goals in the sense of an objective function that accounts for several critical goals, which may serve as a paradigm for the design of complex data workflows. This is particularly the case when data workflow products are embedded in social contexts in which notions of what is ethically acceptable or legal norms change with time (Martin Jr et al., 2020).

Such principles are an admission that the full human control of entire data workflows is not possible. This central limitation should refocus our attention onto managing the dynamics that govern the self-organization of data workflows. Data workflows are the combined result of infrastructure conditions and incentives. Their prescient design and optimization will set data workflows onto the course towards useful (see Sterman, 2000, Chapters 1 and 2 for examples) or harmful (see Sterman, 2000, for instance Part II, Chapter 7.2) operating equilibria.

## References

Bachelard G (1988) Der neue wissenschaftliche Geist, 1st edn. Suhrkamp, Frankfurt am Main.

Bachelard G (2015) Die Philosophie des Nein: Versuch einer Philosophie des neuen wissenschaftlichen Geistes, 3rd edn. No. 325 in Suhrkamp-Taschenbuch Wissenschaft, Suhrkamp, Frankfurt am Main.

Bachelard G, Lepenies W (2016) Die Bildung des wissenschaftliches Geistes: Beitrag zu einer Psychoanalyse der objektiven Erkenntnis, 3rd edn. No. 668 in Suhrkamp Taschenbuch Wissenschaft, Suhrkamp, Frankfurt am Main.

Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286(5439):509–512. DOI: 10.1126/science.286.5439.509.

Barrat A, Barthelemy M, Vespignani A (2008) Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511791383.

Bell G, Hey T, Szalay A (2009) Beyond the Data Deluge. Science 323(5919):1297–1298. DOI: 10.1126/science.1170411.

Bishop CM (2006) Pattern Recognition and Machine Learning, Vol. 4. Springer, New York. ISBN: 03-8731-073-8.

Bodnar T, Salathé M (2013) Validating models for disease detection using twitter. In: Schwabe D, Almeida V, Glaser H, Baeza-Yates R, Moon S (eds.), WWW '13 Companion: Proceedings of the 22nd International Conference on World Wide Web, pp. 699–702. DOI: 10.1145/2487788.2488027.

Braschler M, Stadelmann T, Stockinger K (eds.) (2019a) Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, Cham. DOI: 10.1007/978-3-030-11821-1.

Braschler M, Stadelmann T, Stockinger K (2019b) Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, Cham, pp. 17–29. DOI: 10.1007/978-3-030-11821-1_2.

Brodie ML (2019a) On Developing Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, Cham, pp. 131–160. DOI: 10.1007/978-3-030-11821-1_9.

Brodie ML (2019b) What Is Data Science? In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, Cham, pp. 101–130. DOI: 10.1007/978-3-030-11821-1_8.

Cao L (2018) Data Science: A Comprehensive Overview. ACM Computing Surveys 50(3):1–42. DOI: 10.1145/3076253.

Christian B (2020) The Alignment Problem: Machine Learning and Human Values. WW Norton & Company, New York. ISBN: 978-0-393868-33-3.

Cornfeld IP, Fomin SV, Sinai YG (2012) Ergodic Theory, Vol. 245. Springer Science & Business Media, New York. DOI: 10.1007/978-1-4615-6927-5.

Dao D (2022) GainForest - A sustainable smart contract for the natural world. URL: https://www.gainforest.net/gainforest.net/ [accessed 2022-06-29].

Davenport TH, Patil D (2012) Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review 90(10):70–76. URL: http://blogs.sun.ac.za/open-day/files/2022/03/Data-Scientist-Harvard-review.pdf.

De Roure D (2010) e-Science and the Web. IEEE Computer 43(5):90–93. DOI: 10.1109/MC.2010.133.

Deng C, Ji X, Rainey C, Zhang J, Lu W (2020) Integrating Machine Learning with Human Knowledge. iScience 23(11):1–27, Elsevier. DOI: 10.1016/j.isci.2020.101656.

Dhar V (2011) Prediction in Financial Markets: The Case for Small Disjuncts. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):1–22, ACM New York. DOI: 10.1145/1961189.1961191.

Dhar V (2013) Data Science and Prediction. Communications of the ACM 56(12):64–73, ACM New York. DOI: 10.1145/2500499.

Duhem P (1892) Quelques réflexions au sujet des théories physiques. Revue des Questions Scientifiques 31:139–177.

Duhem PMM (1996) Essays in the History and Philosophy of Science. Hackett Publishing Company, Indianapolis. ISBN: 978-0-872203-09-9.

ETH Zurich IoIBI (2020) COVID-19 Re. URL: https://ibz-shiny.ethz.ch/covid-19-re-international/ [accessed 2022-08-28].

Fleck L (2019) Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv, 12th edn. No. 312 in Suhrkamp-Taschenbuch Wissenschaft, Schäfer L, Schnelle T (eds.), Suhrkamp, Frankfurt am Main. ISBN: 978-3-518279-12-0.

Foster I, Kesselmann C (1999) The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco. ISBN: 978-1-558604-75-9.

Friedman JH (1998) Data Mining and Statistics: What's the Connection? Symposium on the Interface between Computer Science and Statistics, Vol. 29, pp. 3–9. URL: http://www.stats.org.uk/Friedman1997.pdf.

Galison P, Einstein A, Poincaré H (2004) Einstein's clocks, Poincaré's maps: Empires of Time, 1st edn. Norton, New York, NY. ISBN: 978-0-393326-04-8.

Gerlach M, Altmann EG (2019) Testing Statistical Laws in Complex Systems. Physical Review Letters 122(16):168301, APS. DOI: 10.1103/PhysRevLett.122.168301.

Gross T, Blasius B (2008) Adaptive coevolutionary networks: a review. Journal of the Royal Society Interface 5(20):259–271. DOI: 10.1098/rsif.2007.1229.

Heilbron JL (2003) The Oxford Companion to the History of Modern Science. Oxford University Press. DOI: 10.1093/acref/9780195112290.001.0001.

Jensen HJ (2009) Probability and Statistics in Complex Systems, Introduction to. In: Meyers RA (ed.), Encyclopedia of Complexity and Systems Science, Springer, New York, NY, pp. 7024–7025. DOI: 10.1007/978-0-387-30440-3_419.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589. DOI: 10.1038/s41586-021-03819-2.

König MD, Battiston S, Napoletano M, Schweitzer F (2011) Recombinant knowledge and the evolution of innovation networks. Journal of Economic Behavior & Organization 79(3):145–164, Elsevier. DOI: 10.1016/j.jebo.2011.01.007.

König MD, Battiston S, Napoletano M, Schweitzer F (2012) The efficiency and stability of R&D networks. Games and Economic Behavior 75(2):694–713, Elsevier. DOI: 10.1016/j.geb.2011.12.007.

König MD, Liu X, Zenou Y (2019) R&D Networks: Theory, Empirics, and Policy Implications. Review of Economics and Statistics 101(3):476–491. DOI: 10.1162/rest_a_00762.

Koralov L, Sinai YG (2007) Theory of Probability and Random Processes. Springer Science & Business Media, Heidelberg. DOI: 10.1007/978-3-540-68829-7.

Kuhn TS (1962) The Structure of Scientific Revolutions. The University of Chicago Press, Chicago. ISBN: 978-0-226458-11-3.

Ladyman J, Lambert J, Wiesner K (2013) What is a complex system? European Journal for Philosophy of Science 3(1):33–67. DOI: 10.1007/s13194-012-0056-8.

Leek J (2013) The key word in "Data Science" is not Data, it is Science. Simply Statistics 12. URL: https://simplystatistics.org/posts/2013-12-12-the-key-word-in-data-science-is-not-data-it-is-science.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A, et al. (2011) Big data: The next frontier for innovation, competition, and productivity. Tech. Rep., McKinsey Global Institute. URL: https://www.mckinsey.com/~/media/mckinsey/businessmckinsey20insights/big20next20innovation/mgi_big_data_full_report.pdf.

Martin Jr D, Prabhakaran V, Kuhlberg J, Smart A, Isaac WS (2020) Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. arXiv preprint arXiv:2006.09663. URL: https://arxiv.org/abs/2006.09663.

Masterson V (2021) How 3 tech start-ups plan to tackle forest restoration. URL: https://www.weforum.org/agenda/2021/04/reforestation-start-ups-technology-trillion-trees/ [accessed 2022-06-29].

McGeoch CC (2012) A Guide to Experimental Algorithmics. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511843747.

Meinshausen N (2007) Relaxed Lasso. Computational Statistics & Data Analysis 52(1):374–393, Elsevier. DOI: 10.1016/j.csda.2006.12.019.

Newman ME (2003) The Structure and Function of Complex Networks. SIAM Review 45(2):167–256, SIAM. DOI: 10.1137/S003614450342480.

Newman ME, Barabási ALE, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton University Press. ISBN: 978-0-691113-57-9.

Niiniluoto I (2019) Scientific Progress. In: Zalta EN (ed.), The Stanford Encyclopedia of Philosophy, Winter 2019 edn. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2019/entries/scientific-progress/.

Paul MJ, Dredze M, Broniatowski D (2014) Twitter Improves Influenza Forecasting. PLOS Currents Outbreaks. DOI: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

Perlich C, Provost F, Simonoff J (2003) Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. Journal of Machine Learning Research 4:211–255. URL: https://www.jmlr.org/papers/volume4/perlich03a/perlich03a.pdf.

Pfeifer R, Lungarella M, Iida F (2007) Self-Organization, Embodiment, and Biologically Inspired Robotics. Science 318(5853):1088–1093. DOI: 10.1126/science.1145803.

Popper K (2014) Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge. ISBN: 978-0-415285-94-0.

Raghavan G, Thomson M (2019) Neural networks grown and self-organized by noise. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc Fd, Fox E, Garnett R (eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., Vol. 32. URL: https://proceedings.neurips.cc/paper/2019/file/1e6e0a04d20f50967c64dac2d639a577-Paper.pdf.

Rheinberger HJ (2006a) Epistemologie des Konkreten. Studien zur Geschichte der modernden Biologie. Suhrkamp, Frankfurt a. M. ISBN: 978-3-518293-71-3.

Rheinberger HJ (2006b) Experimentalsysteme und epistemische Dinge. Suhrkamp, Frankfurt a. M. ISBN: 978-3-518294-06-2.

Rheinberger HJ (2008) Historische Epistemologie zur Einführung, 2nd edn. Junius, Hamburg. ISBN: 978-3-885066-36-1.

Rigden JS (2005) Einstein 1905: The Standard of Greatness. Harvard University Press. ISBN: 978-0-674015-44-9.

Saha D, De S (2022) Practical Self-Driving Cars: Survey of the State-of-the-Art. Preprints. DOI: 10.20944/preprints202202.0123.v1.

Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, et al. (2012) Digital Epidemiology. Public Library of Science San Francisco. DOI: 10.1371/journal.pcbi.1002616.

Salathé M, Althaus CL, Neher R, Stringhini S, Hodcroft E, Fellay J, Zwahlen M, Senti G, Battegay M, Wilder-Smith A, et al. (2020) COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. Swiss Medical Weekly 150(1112):1–3, EMH Media. DOI: 10.4414/smw.2020.20225.

Schickore J (2018) Scientific Discovery. In: Zalta EN (ed.), The Stanford Encyclopedia of Philosophy, Summer 2018 edn. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2018/entries/scientific-discovery/.

Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR (2009) Economic Networks: The New Challenges. Science 325(5939):422–425. DOI: 10.1126/science.1173644.

Shalizi CR (2006) Methods and Techniques of Complex Systems Science: An Overview. In: Complex Systems Science in Biomedicine, pp. 33–114. Springer, Boston, Deisboeck TS, Kresh JY (eds.). DOI: 10.1007/978-0-387-33532-2_2.

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. (2017) Mastering the game of Go without human knowledge. Nature 550(7676):354–359. DOI: 10.1038/nature24270.

Sinai Y (2009) Kolmogorov-Sinai entropy. Scholarpedia 4(3):2034. DOI: 10.4249/scholarpedia.2034.

Sterman JD (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill, Boston. ISBN: 978-0-072311-35-8, URL: http://hdl.handle.net/1721.1/102741.

Thurner S, Hanel R, Klimek P (2018) Introduction to the Theory of Complex Systems. Oxford University Press. DOI: 10.1093/oso/9780198821939.001.0001.

Vespignani A (2009) Predicting the Behavior of Techno-Social Systems. Science 325(5939):425–428. DOI: 10.1126/science.1171990.

Wind E (2000) Das Experiment und die Metaphysik, 1st edn. Suhrkamp Taschenbuch Wissenschaft, Buschendorf B (ed.), Suhrkamp. ISBN: 978-3-518290-78-1.

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L (2022) A survey of human-in-the-loop for machine learning. Future Generation Computer Systems 135:364–381. DOI: 10.1016/j.future.2022.05.014.

Yao A, Gall J, Leistner C, Van Gool L (2012) Interactive Object Detection. In: Chellappa R, Matas J, Quan L, Shah M (eds.), 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3242–3249. DOI: 10.1109/CVPR.2012.6248060.

Zeigler BP (1970) Towards a Formal Theory of Modeling and Simulation I. Tech. Rep., University of Michigan, Ann Arbor, MI.

Zeigler BP (1971) Towards a Formal Theory of Modelling and Simulation II. Tech. Rep., University of Michigan, Ann Arbor, MI.

Zeigler BP, Muzy A, Kofman E (2019) Theory of Modeling and Simulation: Discrete Event and Iterative System Computational Foundations, 3rd edn. Academic Press, San Diego, CA. DOI: 10.1016/C2016-0-03987-6.