Foundations of Data Science: A Comprehensive Overview Formed at the 1st International Symposium on the Science of Data Science

Frank-Peter Schilling, Dandolo Flumini, Rudolf M. Füchslin, Elena Gavagnin, Armando Geller, Silvia Quarteroni and Thilo Stadelmann

Abstract We present a summary of the 1^{st} International Symposium on the Science of Data Science, organized in Summer 2021 as a satellite event of the 8^{th} Swiss Conference on Data Science held in Lucerne, Switzerland.

Frank-Peter Schilling · Thilo Stadelmann
ZHAW Centre for Artificial Intelligence and ZHAW Datalab, Winterthur, Switzerland
✉ scik@zhaw.ch
✉ stdm@zhaw.ch

Dandolo Flumini
ZHAW Institute of Applied Mathematics and Physics and ZHAW Datalab, Winterthur, Switzerland
✉ flum@zhaw.ch

Rudolf M. Füchslin
ZHAW Institute of Applied Mathematics and Physics and ZHAW Datalab, Winterthur, Switzerland
and European Centre for Living Technology, Venice, Italy
✉ furu@zhaw.ch

Elena Gavagnin
ZHAW Institute of Business Information Technology and ZHAW Datalab, Winterthur, Switzerland
✉ gava@zhaw.ch

Armando Geller
Scensei (Switzerland) GmbH, Zurich, Switzerland
✉ armando@scensei.com

Silvia Quarteroni
Swiss Data Science Center (SDSC) and EPFL, Lausanne and Zurich, Switzerland
✉ silvia.quarteroni@datascience.ch

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

DOI: 10.5445/IR/1000146422
ISSN 2363-9881

We discuss what establishes the scientific core of the discipline of data science by introducing the corresponding research question, providing a concise overview of relevant related prior work, followed by a summary of the individual workshop contributions. Finally, we expand on the common views which were formed during the extensive workshop discussions.

1 Introduction

The discipline of artificial intelligence was coined at the Dartmouth Conference (McCorduck, 1979; Nilsson, 2009); the discipline of Data Science was allegedly coined at LinkedIn and Facebook (Stadelmann et al., 2019b). If truth can be extracted from this abridged statement, it is the fact that data science as today's emerging discipline (Brodie, 2019b) has been largely shaped outside the walls of academia (Stadelmann et al., 2013), i.e., outside a scientific environment, but rather in business-driven settings. The goal of the recent 1^{st} Symposium on the Science of Data Science (Schilling et al., 2021) hence has been to discuss the canon of its underlying principles and techniques (models, methods) that are applicable across different use cases and fields of application, to answer the question what "science" underlies the discipline – if it actually is a discipline.

Put in simpler terms, the symposium revolved around the following hypothetical question: If, 15 years from now, one would compare the contents of the standard textbooks of statistics, computer science, AI and other "source disciplines" of data science on the one hand, with the contents of the then classic text book of data science (still to be written) on the other hand – what would be part of the data science textbook? What establishes the scientific core of data science that is not covered somewhere else? The symposium's goal thus was to launch an activity towards establishing a reference framework for data science.

The importance of this activity transcends common academic drives for order, rigour and scrutiny. First, by starting research labs, degree programs and whole departments, academia creates structures and molds careers that will stay for a long time. It is important that these developments are well-founded, non-redundant and long-lasting, and not just tailored to a sudden demand. Second, a lesson can be learned from projected similarities between data science's development on the one hand, and how computer science on the other hand

emerged out of the fields of mathematics and electrical engineering in the 1950s in Germany (Gunzenhäuser, 1988). At first being little more than the application of principles of these two source disciplines, computer science used the space it was granted as a new discipline to grow into completely new areas that might arguably not have been developed otherwise (see also the *Annals of the History of Computing*). Today, little of computer science's curricula overlap with mathematics or electrical engineering as a result of this emancipation that paved the way for much of what propelled (scientific and economic) progress in the last decades.

Similar to the way Denning (2005) argued for computer science, we thus think that data science has the potential to "[meet] every criterion for being a science, but it has a self-inflicted credibility problem" – the mainstream and media hype around it. In the remainder of this paper, we survey related work on the foundations of data science in Section 2. We then summarize the main contributions from the ISSDS'21 symposium in a synthesis-forming way in Section 3, pointing to a solution to the credibility problems. Last, but not least, we discuss the ensuing implications in Section 4, before the concluding remarks. This paper thus serves as a key and introduction to the individual contributions from the ISSDS'21 participants.

2 Related Work

Ever since the term "data science" came into existence around 60 years ago, there has been a debate on what exactly constitutes data science, how it differentiates itself from statistics and computer science, and whether it deserves the word science in its name. Can it be viewed as an academic discipline on its own that represents more than the sum of its constituent disciplines? In the following, we address these questions by giving a brief historic account as well as a, necessarily incomplete, summary of the current debate.

2.1 Historic Roots

The first use of the term "data science" as a new scientific field goes back to the early 1960s, when Peter Naur introduced the term (interchangeably

with "datalogy") (Sveinsdottir and Frøkjær, 1988), while John Tukey (1962) described a new scientific field he called "data analysis". In 1974, the term "data science" appeared in Naur's book "Concise Survey of Computer Methods" (Naur (1974), p. 30):

> *" Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences. "*

It builds upon the IFIP[1] definition of data as "a representation of facts or ideas in a formalised manner capable of being communicated or manipulated by some process" (Gould, I.H. (ed.), 1971). Naur had a conception of data science rooted in computer science, while Tukey used the term in reference to statistics, two perspectives which are also alluded to in David Hand's two kinds of *big data exercise* (Hand, 2016).

The discussions in the scientific community then continued through the 80s and 90s. In 1985, C.F. Jeff Wu (1986) used the term "data science" as an alternative name for statistics. Later, in his inaugural lecture "Statistics = Data Science?" at the University of Michigan (Wu, 1997), he summarized statistics as a trilogy of data collection, data modelling and analysis, together with problem solving and decision making. He highlighted the most relevant future directions as dealing with large and complex data (data mining), employing a data-driven, empirical approach, as well as the representation of knowledge, and finally suggests that it is time for statistics to make a bold move, namely to rename itself to data science.

Already in 1992, at a statistics symposium in Montpellier, France, the emergence of data science as a new discipline was acknowledged (Escoufier et al., 1995, pp (vii)-(viii)):

> *" The authors propose ways to formalize* data analysis*. ... Such an approach gives birth to a new science with data at its core. Its nature, numerical, qualitative or symbolic, determines the type of operations possible with them. Their origin, whether exhaustive collection or sample, conditions the objective expected in their analysis. It seems justified to coin the term* data science *for this particular activity. "*

[1] International Federation for Information Processing

The first international conference which had the term "data science" in its name took place in 1996 in Kobe, Japan, where Chikio Hayashi (1998) argued for data science as a new, interdisciplinary concept with three phases: data design, collection, and analysis.

In his paper "Statistical Modeling: The Two Cultures", Leo Breiman (2001) discussed two approaches to extract value from data: (i) Predictive modeling, i.e. the ability to predict outcomes to future input data to a model, and (ii) inference, i.e. to extract some information about the underlying model which generates the data. Breiman argued that statistics as a discipline so far was almost exclusively focused on inference, and highlighted the importance of predictive modeling (the prime example being machine learning) when using data to solve problems.

In 2001, William S. Cleveland (2001) introduced data science as an independent scientific discipline based on his proposal of:

> *"... a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called data science."*

His "action plan" discussed six technical areas by which to extend statistics, namely (i) multidisciplinary investigations, (ii) models and methods for data, (iii) computing with data, (iv) pedagogy, (v) tool evaluation and (vi) theory.

These six technical areas of data science introduced by Cleveland were updated and generalized by David Donoho (2017) into the "6 divisions of Greater Data Science" as follows:

(i) Data exploration and preparation (exploratory data analysis, data cleaning);

(ii) Data representation and transformation (databases, feature extraction);

(iii) Computing with data (computer languages like R and Python, cluster and cloud computing, workflows and packages);

(iv) Data modeling (both generative and predictive cultures, c.f. Breiman (2001));

(v) Data visualization and presentation (plotting tools, dashboards);

(vi) Science about data science.

About the last division, the science about data science, Donoho writes (Donoho (2017), p. 756):

> *" Data scientists are doing science about data science when they identify commonly-occurring analysis/processing workflows, for example using data about their frequency of occurrence in some scholarly or business domain*; *when they measure the effectiveness of standard workflows in terms of the human time, the computing resource, the analysis validity, or other performance metric, and when they uncover emergent phenomena in data analysis, for example new patterns arising in data analysis workflows, or disturbing artifacts in published analysis results. The scope here also includes foundational work to make future such science possible – such as encoding documentation of individual analyses and conclusions in a standard digital format for future harvesting and meta analysis. "*

Donoho gives meta- and cross-study analyses as examples for this "science about data science", and concludes that it will grow dramatically in significance in the future, in particular because of the paradigm of reproducible, open science.

2.2 Contemporary Definitions

In his comprehensive overview, Longbing Cao (2017) presents two definitions of data science as well as of data products (p. 43:8):

> *" Definition* 2.1 (*Data Science*): *A high-level statement is*: *data science is the science of data or data science is the study of data.*
> *Definition* 2.2 (*Data Science*): *From the disciplinary perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments* (*including domains and other contextual aspects, such as organizational and social aspects*) *in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology.*
> *Definition* 2.3 (*Data Products*): *A data product is a deliverable from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system. The ultimate data products of value are knowledge, intelligence, wisdom, and decision. "*

In Stadelmann et al. (2019a), it is argued that when defining data science, either a top-down or a bottom-up approach can be followed: The top-down view understands the field as the study of approaches to generate value from

data and building data products, while in the bottom-up view, data science is an interdisciplinary research field with a new, holistic way to deal with data, integrating competencies from computer science, statistics, AI, data mining, but also entrepreneurship. Their definition of data science reads (p. 18):

> *" Data science refers to a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aims at generating value from the data itself. "*

In contrast, Ley and Bordas (2018) first coin data science as "statistics 2.0", a rebirth of statistics in the big data era which has imposed new challenges and opened new research directions. They continue along similar lines as above by stating that being interdisciplinary by nature (statistics, computer and computational sciences, mathematics), it usually combines top-down (model-driven) and bottom-up (data-driven) approaches.

However, Weihs and Ickstadt (2018) point out that the role of statistics and statistical modeling of observational data in data science is often under-estimated compared with e.g. computer science. Michael L. Brodie (2019b) gives the following definition of data science (p. 104):

> *" Data Science is a body of principles and techniques for applying data analytic methods to data at scale, including volume, velocity, and variety, to accelerate the investigation of phenomena represented by the data, by acquiring data, preparing and integrating it, possibly integrated with existing data, to discover correlations in the data, with measures of likelihood and within error bounds. Results are interpreted with respect to some predefined (theoretical, deductive, top-down) or emergent (fact-based, inductive, bottom-up) specification of the properties of the phenomena being investigated. "*

He considers data science a new paradigm, different from the scientific method[2] in terms of (i) data, (ii) methods, (iii) models and (iv) outcomes:

(i) The data are often of observational nature, rather than being collected as in a controlled experiment as in the natural sciences.

[2] In the context of this paper, with the term "scientific method" we refer to the common underlying iterative process across natural sciences, which employs empirical methods as fundamental means to validate predictions, derived from newly-formulated hypotheses about a specific research question (Galilei, 1638; Newton, 1687; Popper, 1959, page: 480).

(ii) Methods are typically domain- and data-specific, even though based on general (e.g. machine learning) approaches.

(iii) Models are created on demand and ad-hoc, and changed or updated frequently, rather than being proposed and refined over many years.

(iv) Regarding outcomes, "the scientific method is used to discover causal relationships between a small number of variables that represent the essential characteristics of the natural phenomena being analyzed" while data science is mainly used to discover correlations.

2.3 Current Debates

In Brodie's view, empirical science and data science have another fundamental difference: The scientific method uses deductive reasoning, is hypothesis- or theory-driven, and works top-down, while data science is typically data-driven, uses inductive reasoning and works bottom-up (in contrast to the top-down views of data science highlighted by Ley and Bordas (2018) and Stadelmann et al. (2019a), see Section 2.2). Since a scientific discipline would require "fundamental principles and techniques applicable to all relevant domains", rather than being domain-, model- and method specific, Brodie does not consider data science a science yet, but rather "an immature, emerging domain that will take a decade to mature".

Regarding the development of data science as a discipline, Brodie (2019a) suggests that this process will be driven by the virtuous cycle of research, development and delivery (RD&D) underlying applied science, as will be the development of data science applications and education.

F. Jack Smith (2006) compares computer science and data science with respect to their recognition as an academic discipline, stating that for both, they are often perceived as being merely within the realm of tools used by technicians. Smith remarks that an important indicator for the establishment of an academic discipline is the dissemination of scientific articles through peer-reviewed journals, which so far had been lacking in the area of data science, but which has started to change in the 2000s, as it did for computer science already from the 1950s. Making the connection to the field of Artificial Intelligence (AI) and machine learning which is at the core of many data science

problems, Michael I. Jordan notes that we are witnessing the creation of a new branch of engineering which should be developed in a human-centric way (Jordan, 2019a,b).

A similar point is made by Blei and Smyth (2017) who present a holistic view of data science. It includes not only the statistical and computational perspectives, but also a human perspective, where the latter involves domain knowledge and data understanding, the ability to fuse methods from both the statistical and computational domains, as well as the task to interpret and visualize the results in their context.

Taking again a more sceptical view, Irizarry (2020) claims that the definitions of data science as given above generally lack consensus on the fundamental principles and the author proposes, in agreement with Jeannette Wing (2019), that "data science is an umbrella term to describe the entire complex and multi-step processes used to extract value from data".

Provost and Fawcett (2013) discuss data science from the perspective of its application to the business world, and state that the "the ultimate goal of data science is improving decision making", supported by data engineering and processing including big data technologies which they however do not consider to be part of data science. Provost and Fawcett advise not to confuse the description of the day-to-day tasks of a data scientist (at the technical level, which often involves a large amount of data processing) with a formal definition of data science as an academic discipline.

2.4 State of the Art

Looking forward, Jeannette Wing (2020) formulates three meta-questions about data science as a discipline:

(i) What is/are the driving deep question(s) of data science, similar to the questions about the origin of life in biology or the origin of the universe in astrophysics?[3]

(ii) What is the role of the domain in the field of data science, i.e. is the inclusion of the domain specific to data science?

[3] The authors would add the unsolved P versus NP problem in theoretical computer science here.

(iii) What makes data science a science, i.e. what makes it more than the sum of its constituent disciplines computer science and statistics?

She then discusses 10 research challenge areas in data science, among others scientific understanding of learning algorithms, causal reasoning, trustworthy AI, privacy and ethics.

In summary, three diverging main themes emerge from the historic and current discussion regarding data science as a scientific discipline:

(i) Data science is often considered an extension/update of statistics ("statistics 2.0"), which is upscaled to meet the new challenges of the big data era, and it is shifting its focus from inference to prediction.

(ii) Data science is an interdisciplinary field, built upon varying selections of fields but mostly upon statistics and computer science, while adding data understanding and domain knowledge as a new perspective.

(iii) Data science can be approached from both top-down (model-driven, generating value, building data products) and bottom-up (data-driven) perspectives. It may be viewed as a new paradigm, which is different from the traditional scientific method which employs controlled experiments.

Thus, at the time of writing, no consensus seems to have formed yet on the question whether data science can be considered an independent academic discipline.

3 Aspects of Data Science

In the following, we present a summary of the main ideas presented in the individual workshop contributions which were received. More details can be found in the individual articles contained in the same volume of this journal.

3.1 A New Scientific Paradigm?

Four workshop contributions (Doemer and Kempf, 2022; Heitz and Schumann, 2021; Ott et al., 2022; Stadelmann et al., 2022) tried to answer the question about the scientific nature of data science as a discipline.

Doemer and Kempf (2022) argue that Data Science can be viewed as a new paradigm in scientific practice, in addition to experimental, theoretical and computational science. As discussed in Section 2, data science can be viewed as either a data-driven (inductive), or hypothesis-driven (deductive) approach. However, the paradigm-shifting nature of data science comes with a problem still awaiting a practical solution: The increasingly complex setups producing huge amounts of data and information at various levels (e.g., meta data in addition to observational data). These are typically of merely observational nature for the data scientist, in contrast with those classically obtained through experiments (in line with the scientific method's principles), i.e. generated under controlled conditions and setup. As a potential solution to this problem, which would provide a basis that allows data science to be consistent with the requirements of transparency, traceability and reproducibility demanded by the scientific method, the authors suggest the adoption of tools, frameworks and platforms provided by *"XOps"* ("X for IT Operations") approaches, where X can be e.g. ML (Machine Learning), Data or AI. For example, MLOps is a set of best practices that aims to deploy and maintain ML models in production reliably and efficiently.

Heitz and Schumann (2021) state that data science consists of *two elements, one based on engineering and the other based on science*. The science element is concerned with creating insights based on phenomena/data measured in the real world, while the engineering element is concerned with creating value ("data products") by making use of derived insights. On the scientific side, it is argued that the way in which the insights are derived from the data must follow scientific principles such as empirical evidence, validity (e.g. in terms of statistical significance)[4] and reproducibility. Such a scientific process is then able to make predictions that however need not necessarily be accompanied by a causal model, in contrast to (Brodie, 2019b). On the other hand, the engineering

[4] Regarding aspects of model validation besides accuracy and significance, see also the discussion in section 2.2.3 of Oberkampf and Roy (2010).

element involves anything that changes the course of the world, as opposed to plain knowledge creation. It includes not only the data product itself, but also decision making, which often depends on external factors (e.g. contribution to profits) not related to data analytics as well. Finally, the engineering side should also include ethical considerations, such as algorithmic fairness.

Ott et al. (2022) make the case for a *systemic view* of the data science workflow, which extends the "classical" workflow (comprising data collection, cleaning, visualization, model building, evaluation and impact/value creation) with various stakeholders such as data scientists, business owners, domain experts and users giving feedback, as well as with societal influences and impacts, that all influence the outcome of a data science project. Four hypotheses towards this perspective are developed:

(i) There is a need for more abstraction and automation in the data workflow and pipeline engineering process.

(ii) Humans play an active role in the data science workflow (e.g. in active learning, or in identifying bias).

(iii) Data science will diversify at the intersection of domains (e.g. life science, health, economics and business etc.).

(iv) With increasing complexity, data science workflows evolve into complex networks, which can be studied and organized with the help of complex systems science.

Stadelmann et al. (2022) propose their answer to the question of the scientific core of data science, which distinguishes it from its contributing disciplines and is not already part of one of them. For the authors, this overarching, unique principle is *data centrism*, i.e. putting data at the center and subject of study, something which is not the case for the contributing disciplines, neither for statistics and machine learning, nor for computer science or service engineering. The unique principle in data science is to create value out of *actual* data (but not ignoring tools and methods to improve data acquisition), and it is argued that recent trends such as explainability (e.g., explainable AI or XAI (see also Melchior, 2022)) and trustworthiness, but also learning from less supervision, are grounded in the data centrism of data science. Finally, the authors argue that, besides its core of data centrism, data science includes several new areas

of research which are not dealt with in the contributing disciplines per se, such as MLOps (see also Doemer and Kempf, 2022), or whose current surge can be attributed to a mindset shift originating in the use cases and culture shaped by data science, like applied semi- and weakly-supervised learning (Simmler et al., 2021), or explainable AI. One example of the latter are explanations of deep neural networks whose necessity arises out of data-driven applications in safety-critical sectors like healthcare (Jin et al., 2022), while other aspects of explainability have been dealt with since longer (Keil and Wilson, 2000).

3.2 Explainability, Rationality and Trust

Furthermore, Melchior (2022) and Füchslin and Flumini (2022) discuss special topics in data science, namely the important issues of transparency and explainability as well as various ways of defining and automating the decision making process.

Melchior (2022) focuses on the notions of transparency, explainability and interpretability in data science in the context of machine learning models, which is an issue in particular for deep learning architectures where hand-crafted features are replaced with many layers of deep neural networks. Recently, explainable AI (XAI) has become a subject of research, in particular for applications with strong safety or ethical requirements, but also in the case of fundamental/natural sciences, where ML is used for knowledge discovery. It is argued that in order to achieve explainability and interpretability, domain experts and data scientists have to work together. Several concrete technical examples are given for the inclusion of domain knowledge in a deep learning model in order to facilitate learning of interpretable features, such as autoencoders or generative models, invertible flow networks or graph neural networks. The latter are particular promising in view of unifying symbolic and connectionist AI approaches.

Füchslin and Flumini (2022) give a definition of rational decision making (structured, inductive, verifiable, grounded) and it is claimed that the former is more appropriately complemented by arationality rather than irrationality. In summary, rational decision making exhibits two main features. Firstly, it is based on some sort of generally accepted scheme of reasoning (in mathematics expressed in an axiomatic manner) and some data/variables. Secondly, the

process of reasoning can be expressed in a language that enables to make the reasoning transparent and comprehensible to a (sufficiently well-educated) other individual and uses terms/variables with a meaning that relates them to the objects one reasons about. Whereas irrationality lacks both of these features, arationality captures the concept of decision making that leads to sensible results but uses processes that one may or may not be able to describe mechanistically, but without the possibility to attribute a meaning to the data representing the process steps in between input and output. Arational decision making includes for instance the notion of intuition. Discovering new proofs of mathematical theorems and generally generating insights and mathematical theories is typically based on a conjecture-proof workflow that includes arationality. According to the authors, an artificial mathematician must therefore also include arational decision making, which is provided for instance by deep neural networks, providing an implementation of AI, as the authors state in their title, as *Arational Intelligence*.

3.3 Education vs Technical Skill

Finally, the contribution of Helmer (2022) is concerned with teaching data science and the corresponding curriculum. As particular challenges, the author mentions the very diverse background of students with different levels of technical skills, the difficulty in providing a suitable computing environment for labs and exercises (local vs cloud based) given the short life cycles of relevant tools and frameworks, and the selection of appropriate use cases and datasets. It is argued that the curriculum should be structured according to the elements and layers of a typical data lifecycle model, in order to provide a structure and frame for the theoretical foundations. Regarding practical approaches, the author suggests, largely in agreement with Irizarry (2020), to structure the curriculum into backend (data engineering) and frontend (data analysis, machine learning) parts, to build the knowledge for developing and maintaining data processing pipelines, to teach data science at the graduate (rather than undergraduate) level, and to consider theoretical foundations at least as important as practical examples. Focusing too much on ever-changing tools and frameworks would shift the curriculum too much towards training, as opposed to education. Finally,

it is argued that care should be taken not to standardize this still very young and fluid field too quickly.

In summary, the individual workshop contributions highlight different aspects of data science and address the research question of its scientific core from various complementary angles. They formed the basis for the common discussion, which is summarized in the following section.

4 Discussion

One of the potential controversial aspects when reasoning about the science in data science is the fundamental difference between experimental data, observational (field) data and citizen-based data, i.e. data collected in the context of citizen science projects. Unlike classic quantitative science, data science relies strongly on the latter two categories and not only on controlled experimental data. Observational and citizen-based data are both affected by the big problem of being potentially biased by humans in their selection or generation process. For example, people take most pictures with daylight.

Therefore, one of the goals of data science that distinguishes it from traditional science is to provide a rigorous methodology to handle data from the *real world* by accounting for the inevitable complexity (e.g., bias) or by modeling concepts which can not be directly observed and, therefore, need experiments or simulations (e.g., risk assessment). One proposal formed at the symposium then is the idea that an exemplary common trait in data science is the way insights are derived from (not controlled) data sources. While the insights derived belong primarily to the respective scientific domains, *how* these were derived pertains to data science and ultimately this constitutes one scientific aspect of the discipline.

Another common difficulty when arguing about the scientificity (i.e., referring to systematicity, logicality, certainty, and precision of knowledge (Xu, 2005)) of data science is solving the issue of explainability and trust. Basic founding principles of science are the quest for explainable models and reproducible experiments: Both of these elements are the basis to trust the ensuing results.

In data science, however, it is not always straightforward to rely on fully explainable models and, as already argued before, on controlled experiments, with the obvious result of doubts being cast on the amount of science present in this discipline. An interesting point of discussion in this context is how

the concept of *trust* is associated to the explainability of the model, therefore often to its simplicity, rather than to its correctness. However, a simple model delivering wrong results can not be trustworthy, hence this association is not always reasonable. This reveals the need for clarity around the concept of trust within and towards data science. Specifically, alternative ways for building trust – in the outcomes of data science, and by extension into the discipline itself – need to be found, which do not necessary rely on experiments or full explainability of models.

The following observation might serve as a starter in this direction: while science is intuition-inspired-then-fact-driven, data science is fully fact-driven-then-intuition-enabled. This stems from the observation that in science, theories are sparked by creativity (the intuition of an apple falling from a tree) and later confirmed by a fully rational (i.e., systematic and logical) process. In data science, theories are derived from data by rational models (e.g., number-crunching neural networks) but encoded in such incomprehensible ways (the network's weight matrix) that methods need to be built to bring human intuition back ex post to leverage the findings (XAI).

Whatever viewpoint individual participants took in the discussion, a small set of key words emerged as central elements to their statements on data science: *data*; *"the wild"* (i.e., real-world applications and use-cases); *pipeline*; and *data products*. While debaters couldn't unify behind a coherent picture of how these central issues are related, there was a consensus that:

(i) Data is central to data science (data evokes theories and not just confirms them).

(ii) Data science is about the real world, specifically its *messiness* (for which it provides methods and tools to deal with).

(iii) As data products are the natural results of data science (its "claims"), the *process* of creating them (the pipeline) plays an important part in constituting the field.

5 Conclusions

Picking up from where Wing (2020) asked her three meta-questions, considering and eventually deliberating more profoundly on what "a" data science actually

is through the lens of philosophy of science (Boyd et al., 1999; Losee, 2001) should be a fruitful endeavor, further shaping the ongoing debate. More specifically, reflecting on and addressing some of the following questions should improve clarity for practitioners and philosophers alike: What is the *purpose* of data science? For example, is it foremost about producing predictions, as many of todays real world applications suggest? Or is it also about creating explanations, too?

If it is also about creating explanations, then what is the *explanatory power* of data science? For example, is it really a black box, as some applications of neural networks would suggest? Is it simply about correlations, as some applications of statistical learning would suggest? Or can we actually learn something from data science about data generation and social mechanisms (Hedstrom, 2005), mid-range theories (Merton, 1949), causality (Pearl and Mackenzie, 2018) and so forth? If so, then this would imply that, at least theoretically, there is *"truth"* through data science. This again would imply that there is a role for rationality, intelligence and intuition in data science and underlying models. And if truth, then also a "measure" of what good science is (Moss and Edmonds, 2005).

And if there is truth, then what is the merit of data science? For example, does it help us to solve practical challenges pertaining to daily decision support tasks better, because it creates more precise *and* accurate predictions? Does data science contribute to conducting science better because it makes better use of an ever expanding repertoire of computational techniques and data repositories (e.g., data lakes)? Does it improve trust in science, because it increases explainability grounded in data?

Similar to the situation in Goethe's "sourcerer's apprentice", it may be that the spirits we summoned, we now cannot rid ourselves of again. No harm done. But at least we should know why we summoned a new scientific discipline. The questions raised above should help creating some clarity. Some attempts to answer them are found in the remaining contributions to this special issue on the 1^{st} Symposium on the Science of Data Science. Others are left for future work.

Acknowledgements The authors are grateful for the support of the ZHAW Datalab and the data innovation alliance in organizing ISSDS'21 as a satellite event of the 8^{th} Swiss Conference on Data Science, and for the participants of the symposium to share and co-shape each other's thoughts. Michael L. Brodie's input to an earlier draft of the symposium concept is very much appreciated.

References

Blei DM, Smyth P (2017) Science and Data Science. Proceedings of the National Academy of Sciences 114(33):8689–8692, National Academy of Sciences. DOI: 10.1073/pnas.1702076114.

Boyd R, Gasper P, Trout JD (eds.) (1999) The Philosophy of Science. MIT Press, Cambridge. ISBN: 978-0-262521-56-7.

Breiman L (2001) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199–231, Institute of Mathematical Statistics. DOI: 10.1214/ss/1009213726.

Brodie ML (2019a) On Developing Data Science. In: Applied Data Science. Springer, pp. 131–160. DOI: 10.1007/978-3-030-11821-1.

Brodie ML (2019b) What is Data Science? In: Applied Data Science. Springer, pp. 101–130. DOI: 10.1007/978-3-030-11821-1.

Cao L (2017) Data Science: A Comprehensive Overview. ACM Computing Surveys 50(3), Association for Computing Machinery, New York, NY, USA. DOI: 10.1145/3076253.

Cleveland WS (2001) Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Review 69(1):21–26. DOI: 10.1111/j.1751-5823.2001.tb00477.x

Denning PJ (2005) Is Computer Science Science? Communications of the ACM 48(4):27–31, ACM New York, NY, USA. DOI: 10.1145/1053291.1053309.

Doemer M, Kempf D (2022) Is It Ops That Make Data Science Scientific? Archives of Data Science, Series A (Online First) 8(2):1–12. DOI: 10.5445/IR/1000150237.

Donoho D (2017) 50 Years of Data Science. Journal of Computational and Graphical Statistics 26(4):745–766, Taylor & Francis. DOI: 10.1080/10618600.2017.1384734.

Escoufier Y, et al. (1995) Preface. In: Escoufier Y, et al. (eds.), Data Science and Its Applications. Academic Press, Tokyo. ISBN: 978-0-122417-70-2.

Füchslin RM, Flumini D (2022) AI as Arational Intelligence? Archives of Data Science, Series A (Online First) 8(2):1–10. DOI: 10.5445/IR/1000150236.

Galilei G (1638) Discorsi e dimostrazioni matematiche intorno a due nuove scienze attinenti la meccanica e i movimenti locali. Elzeviri, Leiden (NL).

Gould, I.H. (ed.) (1971) IFIP Guide to Concepts and Terms in Data Processing. North-Holland Publishing Company.

Gunzenhäuser R (1988) Entwicklung und Bedeutung der Informatik in den Hochschulen der Bundesrepublik Deutschland. In: Baur F (ed.), Nutzungsbilanz moderner Informations-und Kommunikationssysteme aus Anwendersicht/User Experience in the Application of Modern Information and Communication Systems. Springer, pp. 25–36. DOI: 10.1007/978-3-642-83515-5.

Hand DJ (2016) Editorial: Big Data and Data Sharing. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179(3):629–631. DOI: 10.1111/rssa.12185.

Hayashi C (1998) What is Data Science? Fundamental Concepts and a Heuristic Example. In: Hayashi C, Yajima K, Bock HH, Ohsumi N, Tanaka Y, Baba Y (eds.), Data Science, Classification, and Related Methods, Springer Japan, Tokyo, pp. 40–51. DOI: 10.1007/978-4-431-65950-1_3.

Hedstrom P (2005) Dissecting the Social. On the Principles of Analytical Sociology. Cambridge University Press. DOI: 10.1017/CBO9780511488801.

Heitz C, Schumann R (2021) Data Science: What is Science and What is Engineering? Oral Presentation at the Swiss Conference on Data Science, Lucerne 2021.

Helmer S (2022) Teaching Data Science. Constructing Pillars in a Fluid Field. Archives of Data Science, Series A (Online First) 8(2):1–8. DOI: 10.5445/IR/1000150240.

Irizarry RA (2020) The Role of Academia in Data Science Education. Harvard Data Science Review 2(1):1–8. DOI: 10.1162/99608f92.dd363929.

Jin D, Sergeeva E, Weng WH, Chauhan G, Szolovits P (2022) Explainable Deep Learning in Healthcare: A Methodological Survey from an Attribution View. WIREs Mechanisms of Disease n/a(n/a):e1548. DOI: 10.1002/wsbm.1548.

Jordan MI (2019a) Artificial Intelligence - The Revolution Hasnt Happened Yet. Harvard Data Science Review 1(1). DOI: 10.1162/99608f92.f06c6e61.

Jordan MI (2019b) Dr. AI or: How I Learned to Stop Worrying and Love Economics. Harvard Data Science Review 1(1). DOI: 10.1162/99608f92.b9006d09.

Keil FC, Wilson RA (eds.) (2000) Explanation and Cognition. The MIT Press, Cambridge. ISBN: 978-0-262112-49-9.

Ley C, Bordas SPA (2018) What Makes Data Science Different? A Discussion Involving Statistics 2.0 and Computational Sciences. International Journal of Data Science and Analytics 6(3):167–175. DOI: 10.1007/s41060-017-0090-x.

Losee J (2001) A Historical Introduction to the Philosophy of Science, 4th edn. Oxford University Press, Oxford. ISBN: 978-0-198700-55-5.

McCorduck P (1979) Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. CRC Press. ISBN: 978-1-568812-05-2.

Melchior M (2022) Incorporating Domain Knowledge for Learning Interpretable Features. Archives of Data Science, Series A (Online First) 8(2):1–14. DOI: 10.5445/IR/1000150142.

Merton RK (1949) On Sociological Theories of the Middle Range. In: Social Theory and Social Structure. Simon & Schuster, The Free Press, New York, pp. 39–53.

Moss S, Edmonds B (2005) Towards Good Social Science. Journal of Artificial Societies and Social Simulation 8(4):1–15. URL: https://www.jasss.org/8/4/13/13.pdf.

Naur P (1974) Concise Survey of Computer Methods. Studentlitteratur, Lund, Sweden. ISBN: 91-4407-881-1.

Newton I (1687) Philosophiae naturalis principia mathematica. J. Societatis Regiae ac Typis J. Streater.

Nilsson NJ (2009) The Quest for Artificial Intelligence. Cambridge University Press. DOI: 10.1017/CBO9780511819346.

Oberkampf WL, Roy CJ (2010) Verification and Validation in Scientific Computing, 1st edn. Cambridge University Press. DOI: 10.1017/CBO9780511760396.

Ott T, Horn C, Garcia V (2022) The Sciences of Data – Moving Towards a Comprehensive Systems Perspective. Archives of Data Science, Series A (Online First) 8(2):1–9. DOI: 10.5445/IR/1000150241.

Pearl J, Mackenzie D (2018) The Book of Why. Basic Books. DOI: 10.5555/3238230.

Popper KR (1959) The Logic of Scientific Discovery. The Logic of Scientific Discovery, Basic Books, Oxford, England. ISBN: 978-0-415278-43-0.

Provost F, Fawcett T (2013) Data Science and Its Relationship to Big Data and Data-Driven Decision Making. Big Data 1:51–59. DOI: 10.1089/big.2013.1508.

Schilling FP, Flumini D, Füchslin RM, Stadelmann T (2021) ISSDS 2021: 1st Intl. Symposium on the Science of Data Science. URL: `https://sds.data-innovation.org/sds2021-1st-international-symposium-on-the-science-of-data-science/`. accessed: 2022-07-02.

Simmler N, Sager P, Andermatt P, Chavarriaga R, Schilling FP, Rosenthal M, Stadelmann T (2021) A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications. In: 8th Swiss Conference on Data Science, IEEE. DOI: 10.1109/SDS51136.2021.00012.

Smith FJ (2006) Data Science as an Academic Discipline. Data Science Journal 5:163–164. DOI: 10.2481/dsj.5.163.

Stadelmann T, Stockinger K, Braschler M, Cieliebak M, Baudinot G, Dürr O, Ruckstuhl A (2013) Applied Data Science in Europe: Challenges for Academia in Keeping up with a Highly Demanded Topic. In: 9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013.

Stadelmann T, Braschler M, Stockinger K (2019a) Data Science. In: Applied Data Science. Springer, pp. 17–29. DOI: 10.1007/978-3-030-11821-1.

Stadelmann T, Braschler M, Stockinger K (2019b) Introduction to Applied Data Science. In: Applied Data Science. Springer, pp. 3–16. DOI: 10.1007/978-3-030-11821-1.

Stadelmann T, Klamt T, Merkt PH (2022) Data Centrism and the Core of Data Science as a Scientific Discipline. Archives of Data Science, Series A (Online First) 8(2):1–16. DOI: 10.5445/IR/1000143637.

Sveinsdottir E, Frøkjær E (1988) Datalogy – The Copenhagen Tradition of Computer Science. BIT Numerical Mathematics 28(3):450–472. DOI: 10.1007/BF01941128.

Tukey JW (1962) The Future of Data Analysis. The Annals of Mathematical Statistics 33(1):1–67. DOI: 10.1214/aoms/1177704711.

Weihs C, Ickstadt K (2018) Data Science: The Impact of Statistics. International Journal of Data Science and Analytics 6(3):189–194. DOI: 10.1007/s41060-018-0102-5.

Wing JM (2019) The Data Life Cycle. Harvard Data Science Review 1(1):1–6. DOI: 10.1162/99608f92.e26845b4.

Wing JM (2020) Ten Research Challenge Areas in Data Science. CoRR abs/2002.05658. DOI: 10.48550/arXiv.2002.05658.

Wu CFJ (1986) Future Directions of Statistical Research in China: A Historical Perspective. Application of Statistics and Management 1:1–7. URL: https://archive.org/details/future_directions.

Wu CFJ (1997) Statistics = Data Science? URL: http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.

Xu ZL (2005) Science and Scientificity. Genomics Proteomics Bioinformatics 3(4):197–200. DOI: 10.1016/s1672-0229(05)03026-3.