Data Centrism and the Core of Data Science as a Scientific Discipline

Thilo Stadelmann and Tino Klamt and Philipp H. Merkt

Abstract Data science is one of the most significant developments in computing in the 21^{st} century. It is also described as a discipline in the making, drawing principles, methods and tools from established fields like computer science, statistics, science, business, politics, and any domain with adequate data. What are data science's underlying principles and techniques (models, methods) that are applicable across different use cases and fields of application? What novel aspect of science underlies this emerging discipline? We argue that it is *data centrism* – the reliance on data itself, in mindset, methods and products – that makes data science more than the sum of its parts, as this is not done in any other discipline.

Thilo Stadelmann
ZHAW Centre for Artificial Intelligence and ZHAW Data Science Laboratory, Winterthur, Switzerland
✉ stdm@zhaw.ch

Tino Klamt
University of Greifswald, Greifswald, Germany
✉ tino.klamt@stud.uni-greifswald.de

Philipp H. Merkt
Carl Remigius Medical School, Research Group Emergency Medicine, Idstein, Germany
✉ philipp.merkt@carl-remigius.de

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

DOI: 10.5445/IR/1000143637
ISSN 2363-9881

1 Introduction

Data science has been defined previously as *"a unique blend of principles and methods from analytics, engineering, entrepreneurship and communication that aim at generating value from the data itself"* (Stadelmann et al., 2019a). A similar notion was conveyed by Stadelmann et al. (2013) and later refined in (Stadelmann et al., 2019b) when by referring to the data scientist the authors actually defined the activity of doing data science as being determined by what is taken out of the contributing disciplines (see Figure 1).

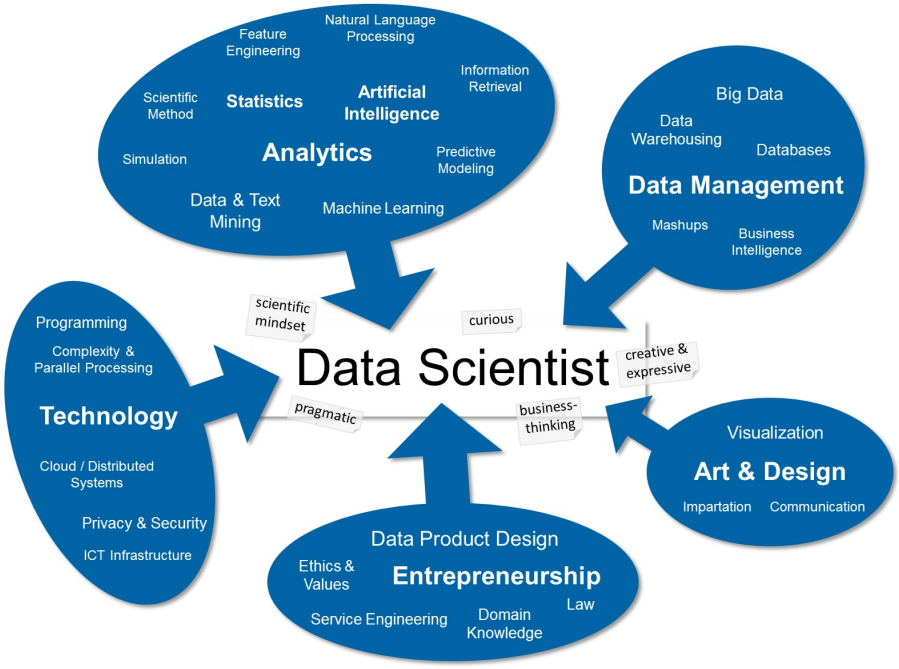

Figure 1: The definition of a data scientist and, by implication, of the activity of doing data science, according to Stadelmann et al. (2019b) (used with permission). In this paper, we argue that data science can *not* be defined merely as a unique cut of contributions from such contributing disciplines – it needs to have a scientific core of its own to warrant the designation of a discipline.

Now, several years after the main wave of the data science hype, one could ask heretically: "what remains of this 'discipline in the making' (Brodie, 2019b) if

all there is in novelty is foremost a contribution to or from one of its constituting disciplines?" For example, when a data scientist develops a new analytical method, it will foremost be a novelty in the field of statistics or machine learning, not specifically in data science. "No scientific discipline" would be the correct answer, if there wasn't more than a selection of contributions from other fields – if there wasn't more to data science than the sum of its parts (Denning, 2005). Data science needs to contribute theories of its own that must be falsifiable (Popper, 1961) to warrant the designation of a science.

In this paper, we argue that there needs to be a scientific core of data science that is (or: is going to become) unique to data science, i.e., that is not the core issue in one of the contributing disciplines. We introduce our proposal for this core in Section 2, followed by an example from medical data analysis practice in Section 3 to illustrate the point. We then discuss limitations of this proposal in Section 4, which might indicate that this view is only partial, and draw conclusions in Section 5.

2 Data centrism

Naturally, this disciplinary core of data science has to materialize in aspects that transcend what was taken out of the contributing disciplines. It needs to amount to more than the adoption of singular methods and tools by

(1) designating a unique object (or: phenomenon) of study (Denning, 2013) as well as by

(2) containing an overarching principle under which this study is performed (Denning, 2005).

Regarding (a), we agree with previous definitions like (Dhar, 2013; Luna-Reyes, 2018; Braschler et al., 2019) and others that the object of study in data science is the creation of value from data. With respect to (b), it is our view that the overarching principle is "data centrism".

2.1 Data centrism and other disciplines

By data centrism we mean that data science, in contrast to the contributing disciplines, puts the highest value on data *itself*, by making the data itself central to the data-scientific mindset (source of inspiration), the conduct of doing data science (processes and methods) and its outcome (data products and predictions). We believe this aspect to be the core of data science because it firmly differentiates data science from related fields, as is demonstrated by the following exemplary consideration of such related fields.

Machine learning revolves around *learning from data* (not data itself): principles and methods to gain general knowledge out of finite data (Samuel (1959) put the highest weight on the learning outcome itself in his famous definition and neglected the input entirely). Despite the efforts of Andrew Ng to teach the field otherwise (Ng, 2021), this is still mainly a model-centric endeavour, i.e., conferences, sub-fields and projects revolve around model architectures as the centre pieces. Then, suitable data to satisfy the needs of the predominantly supervised modeling approaches has to be delivered for machine learners to usually take up the work. It is arguably the influence of data science that un- and semi-supervised methods are increasingly researched and used in recent years: Unsupervised learning was for a long time mainly equated to clustering (Mitchell, 1997). The rise of unsupervised learning as, e.g., spearheaded by Meta's Yann LeCun (LeCun and Misra, 2021), coincides with the rise of data-driven companies like Meta's Facebook and their needs as addressed by data science.

Statistics is concerned with *quantifying data*: its distribution, variability, the certainty of predictions, etc. Data thereby is the main object of analysis, while models again are the center of thinking and acting as well as the main outcome (Breiman, 2001). Specifically, the main stream of statistics revolves around certain modeling assumptions (e.g., linearity, normal error distribution $(0, \sigma^2)$) to which the data has to comply in order to permit claims to be made.

Data management cares for proper *processing of data* in an efficient, reliable and accessible fashion. Again, data is the object under focus, while algebra provides the theoretical backdrop for modelling, machine learning may provide means for optimizing queries (Heitz and Stockinger, 2019) and tools may provide support for data integration (Stonebraker et al., 2013; Stadelmann

et al., 2015). Data here (as before) is not the subject determining the course, but merely the object of study under the specific perspective of manageability.

Service engineering secures *value creation from data*: not just commercially, but for all stakeholders of the value chain, including providers and customers. It thus puts the pains and gains of all stakeholders at the center (Meierhofer et al., 2019), making data a natural resource rather than the centerpiece.

The list could be continued to include all major disciplines mentioned in Figure 1 as contributors to data science, but the pattern is already established, at least on an intuitive scale: These disciplines have data as an object of study (to varying degrees). In contrast, data science has data at the centre, as the subject (or: the driving force), and methods are employed that expect everything from the data itself (e.g., structure, patterns, supervision, value). Specifically, data science is the science of studying the data *as is*: it doesn't impose assumptions on the quality or quantity of data before its methods can be applied, but seeks methods that can make the most out of the data *that is available*. This is what is implied in having "value-creation from [actual] data" as the focus of the discipline. It includes both the current data at hand, but also data that can realistically be produced by improved data acquisition and preprocessing methods.

2.2 The effects of data at the centre

The unique point of view upheld by data science, hence, and in contrast to any of the contributing disciplines, is the one that looks for supreme value *in* the data itself (and not just *out* of it, as one ingredient). The distinction is subtle, but crucial: "in" the data means that data is the main ingredient, the centerpiece, at the same time ultima ratio and conditio sine qua non. On the other hand, by "out" of data we mean that data is a mere resource in the pursuit of some further end. The difference can be likened to a private horse owner who sees value *in* a horse (e.g., relational value), in contrast to a farmer of old who saw value *out* of a horse (as a means to pull a plow). Let's exemplify how data science implements this principle with a couple of examples.

Empiricism is the driving force in data science: in contrast to pre-conceived models of reality, data science reinforces the *mindset* to establish theories out of the patterns that arise from potentially vast amounts of data (i.e., empirical evidence rather than human intuition) (Hey et al., 2009). The effect of this is

that data science models tend to become complex and opaque, as they didn't originate in a simple human idea, but emerged in a data-driven way. Deep learning methods are a good example for this, and the recent trend to research and apply *explainable and trustworthy methods* (Samek et al., 2019; Amirian et al., 2021) can be seen as a direct reaction to the data science mindset: If the data itself is determining the model, the discipline responsible for this development, as a next step, has to provide methods that make this machine-conceived models again amenable to human intuition, decision and control.

Learning from less (e.g., less data with as little as possible human-provided interpretations/supervision) can guide the learning of decision-making functions out of mere observations and probably also should do so in order to avoid human-introduced biases (Glüge et al., 2020; Wehrli et al., 2021). While this naturally employs machine learning methods, it is the mindset of data science (seeking a solution that relies on data alone instead of human annotations) that prompts the selection of *un- and semi-supervised methods* and not vice versa. Additionally, such methods are also applied by data scientists to gain models (and out of them value) from obviously imperfect data sets. It is again the data science mindset that asks "what can be done to exploit the actual data best" rather than "who can bring me better data or labels to train my method" (Hollenstein et al., 2019; Simmler et al., 2021).

Data products are outcomes (digital services, physical products or anything in between) that have data at their core (Loukides, 2011) and not just as an ingredient. While again certain methods from the contributing disciplines are necessary conditions for them to function, a prime candidate being service engineering (Meierhofer et al., 2019), only by adding the data itself the sufficient conditions for value generation are met. Hence, they derive their added value from the added data.

2.3 Data centrism in the literature

This list in Section 2.2 could (and should) be extended as well to establish the pattern more strongly. However, intuitively, what the list resembles is the same mindset reinforced several times in the 2020-2021 issues of Andrew Ng's "The Batch" (DeepLearning.AI editorial team, 2021) of thinking data-centric rather than {model, user, customer, theory, application, . . . }-centric. Similar

arguments are provided for example by Della Corte and Della Corte (2021) and Gerdes (2021).

Putting data at the centre of *thinking* (i.e., assuming data necessary (Jeffreys and Jeffreys, 1988) for the realization of the expected added value, and data plus data science methods sufficient), has been already hinted at in Hey et al. (2009) for applications in the sciences, and is of course discussed in contributing disciplines like machine learning (Ng, 2021; Ng et al., 2021). Data centrism has further been discussed (and partially been dismissed) as a guiding principle for physical computer network organisation (Shenker, 2003) and server design (Siegl et al., 2016), database (Haas et al., 2011) and middleware development (Chen et al., 2008) as well as the build-up of whole embedded (Alvarez-Coello et al., 2021) and enterprise software architectures (Rajabi and Abade, 2012).

However, the furthering of the data-centric mindset as the core of a scientific discipline on a broader scale within the data-related community, with the *subsequent* consideration within the contributing disciplines to data science in recent years (Lau et al., 2018; Nwokeji et al., 2015; Ng et al., 2021), is arguably the effect and contribution of data science. This view is shared by Leonelli (2019) and Fekete et al. (2021). However, while we are concerned here with a proper *delineation* of the fields of science and technology such as the ones identified by (Braschler et al., 2019) as being contributors to data science (cp. Figure 1), Leonelli presents a philosophical analysis of data-centric research, and Fekete and colleagues are concerned with data science teaching.

3 An example from practice

To illustrate the contrasting approaches in data science and related disciplines, an example is presented from resilience research. It is a prototypical example of a use case that could build on multiple highly different data sources, which would require different methodology to exploit them, leading to different research outcomes in terms of type and scope.

The example research is concerned with increasing the resilience of emergency workers from heterogeneous professional backgrounds such as fire brigade, rescue service, police, military and NGOs, to stressful situations. This comprises answering the two questions of (a) how to effectively and efficiently (i.e., practically possible for professionals in service) measure stress under realistic conditions, and (b) how to increase the resilience to such stress by interventions

like individual trainings. The setup for this research in a first phase is as follows (with the prospect to scale up to larger samples in the next phase): over a period of 72 hours, a group of ca. 20 participants are cast into a series of role-playing scenarios belonging to a fictitious foreign catastrophe situation (Merkt and Wilk-Vollmann, 2021). In these scenarios, they face constantly increasing challenges of asymmetric threat (cp. Figure 2) while data is being recorded. Specifically, all radio traffic is recorded, physiological parameters are taken (heart rate; blood pressure; blood gas analysis for lactate, base excess, glucose; and neurophysiological biomarkers like cortisol and α-amylase), and questionnaires for subjective assessment of the stress level are taken based on standardized interview settings.

Figure 2: Example of a catastrophe scenario as used in the described resilience research (Merkt and Wilk-Vollmann, 2021): Role play is used to create realistic, stressful crisis situations; data is collected during and after the scenarios from the participants to reflect their stress level (picture shows one of the authors). Copyright © by Stefan Mikolon (used with permission).

Typical resilience research would focus on structured questionnaires as data sources to account for human factors in the dealing with stress (Merkt et al., 2020), evaluating them using a qualitative research approach based on Grounded Theory (Adolph et al., 2011). The advantage of these methods lies in the inductive development of categories and theories. This means that the heterogeneous and complex situations within catastrophe scenarios that cannot be standardized beforehand could be dealt with very individually. As part of the qualitative content analysis according to Mayring (2015), which

is based on the Grounded Theory, the inductive theory formation is specified by a concrete methodological analysis process. The core of this process is the coding of individual statements, aiming at assigning the interview content to different categories. These categories, in turn, are validated as part of a reliability test on the basis of various statistical measures, after which an evaluation and interpretation takes place. This is the strength of the qualitative, social science approach, which is based on a formal, structured process of data acquisition.

However, when this resilience research project enters the next phase, it has to scale up to thousands of participants, not only in controlled settings of role-playing scenarios, but in emergency operations in practice. As there is simply no way of getting structured, standardized questionnaire data from all subjects in practice, a data-centric approach rooted in data science is a valid alternative: Subjects are equipped with few easily manageable sensors and post-hoc stress analysis is attempted with the data that these deliver. Additionally, communication under stress reveals a lot about the communicators' stress level, so it is worthwhile to decode the radio communication using AI-based emotion recognition (Biondi et al., 2017). While qualitative methods might in principle deliver more meaningful results based on smaller samples, such methods are excluded by the use case. Only a data science approach with its mindset of "creating value from the actual data" can lead to any result, where "actual" data is the data either readily available or at least realistically producible.

4 Limitations

Focusing on a single aspect is necessary for any detailed study, and identifying the core of an emerging scientific discipline is no exception. We are convinced that data centrism as discussed above is of utmost importance to the scientific core of data science in the sense that it serves as a focal point in deciding what is data science and what is part of a contributing discipline. However, we do not see clearly enough yet if this is the scientific core itself, or some inner ring around it.

Particularly, the following duality illustrates that zooming in too much on data as a subject in data science rather than mere object of study can be misleading in the limit: Making data the "subject that determines the course" naturally assumes given data as the starting point of data science endeavours, and we have presented examples above that illustrate the importance of data science

in working with the data *one has*, the given data, to subsequently research and apply methods that make the most of it rather than dismissing it.

However, already the (real) use case in Section 3 shows that also a data-centric approach rooted in data science has to take into consideration the source, acquisition and quality improvement of data. It will develop adequate methods for this distinct from data acquisition methods in, e.g., qualitative analysis. But this case shows that equating data centrism with "creating value out of *given* data" falls short of the scope of data science and the power of the data-centric paradigm: Data science does contain methods, data-centric methods, to improve on the data by getting more adequate data. Such methods for example analyze the data at hand, realize shortcomings, and prompt users for specific improvements such as filling gaps in the coverage of the data set (guided by data, aimed at data – thus having data at the centre) or create new synthetic samples as in data augmentation (Shorten and Khoshgoftaar, 2019). We thus chose to refer to data science as the discipline dealing with *actual* data (cp. end of Section 3) rather than idealized data (idealization that happens, e.g., when assuming Gaussianty, as discussed by Li (2007)).

On a more fundamental level, having data science as a discipline that puts supreme value in actual data (rather than, e.g., human theories on the causes of this data) opens the door to all kind of problems inherited from this data: The data might be biased (Wehrli et al., 2021) and thus barely suitable to build models on it; it might, in the absence of any theory on its origin and requirements on its quality, give rise to models that find spurious patterns and consequently produce models of machine magical thinking (Diaconis, 2006). It might not find any value at all because the data, in combination with current methods, turns out to be insufficient to realize the added value. For all these – true, actual – risks of assumption-free data analysis, it is important to make data science not a replacement of other scientific disciplines, but an enrichment. If the more formal, less error-prone methods of statistics can be applied in a certain analysis, then this should be done; if causal analysis (Pearl, 2009) can be done and is important for the validity of the result, this should not be neglected. But if no other principle of analysis can be applied than data centrism, for practical or theoretical reasons, then it is important to have the best possible data science methods available. Mitchell (1997) proves that no learning is possible without assumptions; we argue that data science is home to those methods that deliberately work with the least possible amount of assumptions, which

sometimes is the only viable route to take. Of course, such approaches can only detect correlations in the data and make no statements about causality (Cap, 2019). But while correlation is not causation, correlation often is enough (Brodie, 2019b,a; Stockinger et al., 2019). Hence, furthering data science as a data-centric discipline adds something unique to the quiver of scientific methodologies. The skilled hunter will carefully chose the appropriate arrow for each situation.

5 Conclusion

If the scientific core of data science is constituted of those aspects that put data at the core of thinking, acting and expectation, and if, next, methodology from other fields is assembled around this core as need arises, the following tentative list of novel areas of research (and the respective works therein) can arguably be seen as being genuine first-class citizens of the discipline of data science – the non-borrowed part of it:

Machine Learning Operations (MLOps): The discipline of machine learning could live well without taking care of operational issues for several decades (Mitchell, 1997). It is since the advent of data science and hence the data-centric paradigm that methods are created and community is formed to care for the development process including the operation of the complete data product pipeline, and the various feedbacks between them (Mäkinen et al., 2021).

Applied semi- and weakly-supervised learning: While the research of methods on how to learn from little supervision is core machine learning terrain inspired by findings in neuroscience (Zador, 2019), the application of such findings to data problems in industry, health, finance, retail, etc. is the domain and contribution of data science.

Data product design: The data product Loukides (2010) already appeared to be one of the outstanding contributions of data science in one of the first major courses on the subject (Howe, 2014).

Explainable Artificial Intelligence (XAI): Few other fields with a strong technical core have managed to incorporate overarching (societal) concerns into the discipline itself as well as data science has. Be it under the terms of explainable artificial intelligence, data ethics, {DataScience, AI}4Good or others, these developments wouldn't come out of the neighboring disciplines of AI or ethics without the mindset promoted by data science – data centrism. Only

data centrism promotes methods that seek value from the data itself without deferring to humans for modeling decisions, which in turn creates the demand for new methods and frameworks for transparency, interpretation and ethical acting.

Future work will include a more thorough analysis of data centrism: Its origins and current traces, and if this confirms the view suggested here of data centrism being the scientific core of the discipline – and hence kingmaker of data science.

References

Adolph S, Hall W, Kruchten P (2011) Using grounded theory to study the experience of software development. Empirical Software Engineering 16(4):487–513. DOI: 10.1007/s10664-010-9152-6.

Alvarez-Coello D, Wilms D, Bekan A, Gómez JM (2021) Towards a data-centric architecture in the automotive industry. Procedia Computer Science 181:658–663. DOI: 10.1016/j.procs.2021.01.215.

Amirian M, Tuggener L, Chavarriaga R, Satyawan YP, Schilling FP, Schwenker F, Stadelmann T (2021) Two to Trust: AutoML for Safe Modelling and Interpretable Deep Learning for Robustness. In: Heintz F, Milano M, O'Sullivan B (eds.), Trustworthy AI - Integrating Learning, Optimization and Reasoning, Springer, Cham, pp. 268–275. ISBN: 978-3-030739-59-1, DOI: 10.1007/978-3-030-73959-1_23.

Biondi G, Franzoni V, Poggioni V (2017) A deep learning semantic approach to emotion recognition using the IBM watson bluemix alchemy language. In: Stankova E, Gervasi O, Apduhan BO, Murgante B, Misra S, Ryu Y, Taniar D, Tarantino E, Rocha AMA, Torre CM (eds.), International Conference on Computational Science and Its Applications, Springer, Trieste, pp. 718–729. DOI: 10.1007/978-3-319-62398-6_51.

Braschler M, Stadelmann T, Stockinger K (2019) Data science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 17–29. DOI: 10.1007/978-3-030-11821-1_2.

Breiman L (2001) Statistical modeling: the two cultures. With comments and a rejoinder by the author. Statistical Science 16(3):199–231. DOI: 10.1214/ss/1009213726.

Brodie ML (2019a) On Developing Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 131–160. ISBN: 978-3-030118-21-1, 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. Springer, Cham, pp. 101–130. DOI: 10.1007/978-3-030-11821-1_8.

Cap CH (2019) Risks and Side Effects of Data Science and Data Technology. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 79–95. ISBN: 978-3-030118-21-1, DOI: 10.1007/978-3-030-11821-1_6.

Chen G, Li M, Kotz D (2008) Data-centric middleware for context-aware pervasive computing. Pervasive and mobile computing 4(2):216–253. DOI: 10.1016/j.pmcj.2007.10.001.

DeepLearning.AI editorial team (2021) The Batch. Published via: `https://www.deeplearning.ai/the-batch/`. [Online; accessed 23-June-2021].

Della Corte D, Della Corte KA (2021) The data-centric lab: A pharmaceutical perspective. In: Kohei A (ed.), Future of Information and Communication Conference, Springer, Cham, pp. 1–15. DOI: 10.1007/978-3-030-73103-8_1.

Denning PJ (2005) Is computer science science? Communications of the ACM 48(4):27–31. DOI: 10.1145/2447976.2447988.

Denning PJ (2013) The Science in Computer Science. Communications of the ACM 56(5):35–38. DOI: 10.1145/2447976.2447988.

Dhar V (2013) Data science and prediction. Communications of the ACM 56(12):64–73. DOI: 10.1145/2500499.

Diaconis P (2006) Theories of Data Analysis: From Magical Thinking Through Classical Statistics. In: Hoaglin DC, Mosteller F, Tukey JW (eds.), Exploring Data Tables, Trends, and Shapes. John Wiley & Sons, Ltd, New York, chap. 1, pp. 1–36. ISBN: 978-1-118150-70-2, DOI: https://doi.org/10.1002/9781118150702.ch1.

Fekete A, Kay J, Röhm U (2021) A data-centric computing curriculum for a data science major. In: Cutter P, Monge A, Sheard J (eds.), Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, Association for Computing Machinery, New York, pp. 865–871. DOI: 10.1145/3408877.3432457.

Gerdes A (2021) A participatory data-centric approach to AI Ethics by Design. Applied Artificial Intelligence, pp. 1–19. DOI: 10.1080/08839514.2021.2009222.

Glüge S, Amirian M, Flumini D, Stadelmann T (2020) How (not) to measure bias in face recognition networks. In: Schilling FP, Stadelmann T (eds.), IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, Cham, pp. 125–137. DOI: 10.1007/978-3-030-58309-5_10.

Haas PJ, Maglio PP, Selinger PG, Tan WC (2011) Data is dead... without what-if models. Proceedings of the VLDB Endowment 4(12):1486–1489. DOI: 10.14778/3402755.3402802.

Heitz J, Stockinger K (2019) Join query optimization with deep reinforcement learning algorithms. arXiv preprint arXiv:1911.11689.

Hey AJ, Tansley S, Tolle KM, et al. (2009) The fourth paradigm: Data-intensive scientific discovery, Vol. 1. Microsoft Research Redmond, WA.

Hollenstein L, Lichtensteiger L, Stadelmann T, Amirian M, Budde L, Meierhofer J, Füchslin RM, Friedli T (2019) Unsupervised Learning and Simulation for Complexity

Management in Business Operations. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 313–331. DOI: 10.1007/978-3-030-11821-1_17.

Howe B (2014) Introduction to Data Science. Published via: https://www.classcentral.com/course/datasci-451. [Online; accessed 02-February-2022].

Jeffreys H, Jeffreys BS (1988) Section 1.036: Necessary: Sufficient. In: Jeffreys H, Jeffreys BS (eds.), Methods of Mathematical Physics, 3rd Edition. Cambridge University Press, pp. 10–11.

Lau FDH, Adams NM, Girolami MA, Butler LJ, Elshafie MZ (2018) The role of statistics in data-centric engineering. Statistics & Probability Letters 136:58–62. DOI: 10.1016/j.spl.2018.02.035.

LeCun Y, Misra I (2021) Self-supervised learning: The dark matter of intelligence. URL: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/. [Online; accessed 02-February-2022].

Leonelli S (2019) Data governance is key to interpretation: Reconceptualizing data in data science. Harvard Data Science Review 1(1). DOI: 10.1162/99608f92.17405bb6.

Li C (2007) Non-Gaussian, Non-stationary and Nonlinear Signal Processing Methods-with Applications to Speech Processing and Channel Estimation. Institut for Elektroniske Systemer, Aalborg Universitet.

Loukides M (2010) What is data science? O'Reilly Media, Inc. URL: https://www.oreilly.com/radar/what-is-data-science/. [Online; accessed 02-February-2022].

Loukides M (2011) The evolution of data products. O'Reilly Media, Inc. URL: https://www.oreilly.com/radar/evolution-of-data-products/. [Online; accessed 02-February-2022].

Luna-Reyes LF (2018) The search for the data scientist: Creating value from data. ACM SIGCAS Computers and Society 47(4):12–16. DOI: 10.1145/3243141.3243145.

Mäkinen S, Skogström H, Laaksonen E, Mikkonen T (2021) Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? arXiv preprint arXiv:2103.08942.

Mayring PAE (2015) Qualitative Inhaltsanalyse: Grundlagen und Techniken, 12. Auflage. Springer, Cham. ISBN: 978-3-407293-93-0.

Meierhofer J, Stadelmann T, Cieliebak M (2019) Data products. In: Applied Data Science. Springer, Cham, pp. 47–61.

Merkt PH, Wilk-Vollmann S (2021) Anspruchsvolle Übungslagen: Kommunikationsverhalten und Stressreaktionen. Rettungsdienst 1:14–18.

Merkt PH, Wilk-Vollmann S, Wolz C (2020) Forschung in der Notfall- und Katastrophenmedizin. Taktik+Medizin 4:28–31.

Mitchell TM (1997) Machine Learning. McGraw-Hill, New York. ISBN: 978-0-071154-67-3.

Ng AY (2021) MLOps: From Model-centric to Data-centric AI. Published via: https://www.youtube.com/watch?v=06-AZXmwHjo.

Ng AY, Aroyo L, Coleman C, Diamos G, Reddi VJ, Vanschoren J, Wu CJ, Zhou S (eds.) (2021) Online Proceedings of the NeurIPS'21 Data-Centric AI Workshop. URL: https://datacentricai.org/.

Nwokeji JC, Clark T, Barn B, Kulkarni V, Anum SO (2015) A Data-centric Approach to Change Management. In: Hallé S, Mayer W (eds.), 2015 IEEE 19th International Enterprise Distributed Object Computing Conference, Adelaide, Australia.

Pearl J (2009) Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge.

Popper KR (1961) The Logic of Scientific Discovery. Basic Books, Inc., New York.

Rajabi Z, Abade MN (2012) Data-centric enterprise architecture. International Journal of Information Engineering and Electronic Business 4(4):53. DOI: 10.5815/ijieeb.2012.04.08.

Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (2019) Explainable AI: Interpreting, explaining and visualizing deep learning, Lecture notes in computer science, Vol. 11700. Springer, Cham. DOI: 10.1007/978-3-030-28954-6.

Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM Journal of Research and Development 3(3):210–229.

Shenker S (2003) The data-centric revolution in networking. In: Freytag JC, Lockemann P, Abiteboul S, Carey M, Selinger P, Heuer A (eds.), Proceedings 2003 VLDB Conference, Elsevier, New York, p. 15. DOI: 10.1016/B978-012722442-8/50010-0.

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. Journal of Big Data 6(1):1–48. DOI: 10.1186/s40537-019-0197-0.

Siegl P, Buchty R, Berekovic M (2016) Data-centric computing frontiers: A survey on processing-in-memory. In: Jacob B (ed.), Proceedings of the Second International Symposium on Memory Systems, pp. 295–308. DOI: 10.1145/2989081.2989087.

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: Geiger M, Bürki GH (eds.), 8th Swiss Conference on Data Science, IEEE, Lucerne. DOI: 10.1109/SDS51136.2021.00012.

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: van Deursen A, Ghezzi C (eds.), 9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013.

Stadelmann T, Cieliebak M, Stockinger K (2015) Toward automatic data curation for open data. ERCIM News 2015(100):32–33.

Stadelmann T, Braschler M, Stockinger K (2019a) Introduction to Applied Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 3–16. DOI: 10.1007/978-3-030-11821-1_1.

Stadelmann T, Stockinger K, Bürki GH, Braschler M (2019b) Data Scientists. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science. Springer, Cham, pp. 31–45. DOI: 10.1007/978-3-030-11821-1_3.

Stockinger K, Braschler M, Stadelmann T (2019) Lessons Learned from Challenging Data Science Case Studies. In: Stockinger K, Braschler M, Stadelmann T (eds.), Applied Data Science. Springer, Cham, pp. 447–465. ISBN: 978-3-030118-21-1, DOI: 10.1007/978-3-030-11821-1_24.

Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik SB, Pagan A, Xu S (2013) Data Curation at Scale: The Data Tamer System. In: Sixth Biennial Conference on Innovative Data Systems Research, CIDR 2013, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings, www.cidrdb.org. URL: `http://cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf`.

Wehrli S, Hertweck C, Amirian M, Glüge S, Stadelmann T (2021) Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics, pp. 1–14. DOI: 10.1007/s43681-021-00108-6.

Zador AM (2019) A critique of pure learning and what artificial neural networks can learn from animal brains. Nature Communications 10(1):1–7, Nature Publishing Group. DOI: 10.1038/s41467-019-11786-6.