Is It Ops That Make Data Science Scientific?

Manuel Doemer and David Kempf

Abstract The primary challenge of data science as a new paradigm of scientific discovery is the rigorous and practically usable documentation of the data product development process. A more expanding and consequent implementation of principles from modern software development and operations will allow the field to mature as a truly scientific discipline.

1 Data Product Development

The large-scale recording of essential aspects of our world in the form of data (datafication, Cukier and Mayer-Schoenberger, 2013) has allowed for data science to emerge as a discipline that aims at generating value from data in

Manuel Doemer
Center for Artificial Intelligence, Zurich University of Applied Sciences, School of Engineering, Technikumstrasse 71, 8401 Winterthur, Switzerland
✉ manuel.doemer@zhaw.ch

David Kempf
Insitute of Computational Physics, Zurich University of Applied Sciences, School of Engineering, Technikumstrasse 9, 8401 Winterthur, Switzerland
✉ david.kempf@zhaw.ch

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

DOI: 10.5445/IR/1000150237
ISSN 2363-9881

the form of *data products* (Loukides, 2010, 2011; Braschler et al., 2019; Meierhofer et al., 2019).

Data encodes information, i.e. subjective descriptions that can be contextualised and interpreted within a framework of meaning by the parties exchanging and processing the information (see, for example, Zins, 2007 and reviews on the concept of the Data-Information-Knowledge-Wisdom hierarchy in Rowley, 2007; Frické, 2009). In this article, data exclusively refers to binary digital data in electronic form that can be processed and stored by common digital computers: Digital data consists of a finite string of discrete symbols, each of which can adopt only a fixed number of values from some alphabet, such as letters or digits. Binary digital data in particular, is represented by a finite string of binary digits (bits) each of which can adopt a value of either 0 or 1. This usage of the term specifically excludes analog data, e.g. data in the form of electrical analog signals, that can also be processed within the circuitry of common integrated computer chips, but exists on an infinitely continuous scale (Maloberti and Davies, 2016).
Computational processing and analysis of analog data therefore requires a preceding digitization procedure, which converts the analog data into a digital and eventually binary representation. This process involves lossy discretization, i.e. sampling in space and/or time of the analog signal, and quantization, i.e. the rounding and truncation of real numbers onto a discrete scale (Analog Devices Inc., 2004).

A data product aims at generating value from data (Figure 1) in the form of economic profit (Wang et al., 2021), societal benefit (Bansak et al., 2018) or scientific progress (Blei and Smyth, 2017). In an ideal world, both economic value and scientific progress lead to societal benefit. However, we find it necessary to highlight that current practice has not reached this point. Rather, in both the economic and scientific contexts, the immediate benefits from specific data products are typically restricted to relatively small, closed groups. At best we can hope that with time this value diffuses into societal spheres. However, further discussions on ethical and political aspects related to this observation are beyond the scope of the current article. In that sense the term is used as a general label for the target outcome of data science activities regardless of whether they are executed in a commercial, societal or scientific context. It emphasises the focus on meeting specific user needs, rather than striving primarily for

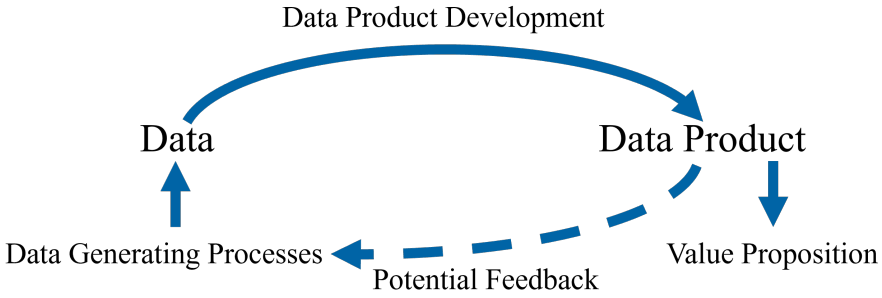

Figure 1: The objective of data science is to develop products that produce value from data. Knowledge about the data generating processes and the value proposition are crucial for success. Potential feedback might require particular considerations: The use of the data product can influence the context of the data streams or represent an additional, distinct data generating process on its own, which feeds into further development increments.

general and complete knowledge beyond the concrete problem statement. As a consequence, the data product is typically domain- or application-specific. This also means, that the appropriate deployment and integration of the product in the corresponding (business) processes and user communities is critical for its success. Subsequent operations of the solution might require substantial resources, especially in the case of complex software architectures.

In the development process (upper curved arrow in Figure 1), data science practice draws pragmatically from a dynamic ecosystem of domain knowledge, applied mathematics and computer science (Conway, 2010). Consequently, the product's distinguished value derives from the specific blend of the ecosystem's components which results in more than simply the sum of its parts (Braschler et al., 2019; Stadelmann et al., 2013).

The domain expertise (or domain knowledge) refers, on the one hand, to the understanding of the data generating processes including the meaning of the data attributes, context and specific procedures of the data sampling, employed instruments and personnel, associated measurement errors and uncertainties, applied processing steps etc. On the other hand, it refers to the understanding of the value proposition, i.e. the way in which the data product generates value by addressing the specific users' needs, pains and gains (Osterwalder et al., 2014). Data products can give rise to ethical issues (Floridi and Taddeo, 2016; Christen et al., 2019), which have to be considered for a successful deployment in the

user communities and society at large. Furthermore, due to the major impact of data products on both the economy and private life, regulatory frameworks have been expanded in recent years. Accordingly, legal aspects have to be taken into account during the development process as early as possible (Widmer and Hegy, 2019).

The lower, dashed arrow in Figure 1 indicates a potential feedback from the data product to the data generating processes, which might influence the underlying distribution of the data, i.e. leading to concept drifts (Tsymbal, 2004; Souza et al., 2020; Widmer and Kubat, 1996). This can mean, that the data product has to be continuously updated to address the dynamically evolving concept, resulting in additional challenges, such as feedback loops in recommender systems (Chaslot, 2019; Yesilada and Lewandowsky, 2022). Furthermore, the deployed data product can represent a distinct data generating process on its own in subsequent development cycles: User tracking data or direct feedback allow to plan for additional features and functions to increase user satisfaction and uncover new insights. Operational data, such as logs and telemetry, can be used to improve the non-functional aspects of the solution, i.e. the quality by which the users experience the core functionalities of the product.

Due to the high complexity and unclear and rapidly changing requirements (Brodie, 2019) in typical data science use cases, agile principles, such as iterative and incremental development, are applied. The outcomes of minimal development investments are tested and user feedback informs the next activities until convergence. Inspired by software development practices, corresponding process frameworks have emerged in the data science community as well. For example, the "Knowledge Discovery in Databases" process (KDD, Fayyad et al., 1996) breaks the overall process down into data selection, pre-processing, transformation, data mining and interpretation/evaluation that leads to knowledge. The "Cross Industry Standard Process for Data Mining" (CRISP-DM, Shearer, 2000) adds special emphasis on the deployment after the business understanding, data understanding, data preparation, modelling and evaluation phases. Visualisations of such processes typically connect the last step back to the beginning into cyclical structures to stress the iterative progression.

Most recently, the focus on deployment challenges associated with big data, machine learning (ML) and artificial intelligence (AI) based products integrated in complex software architectures (Sculley et al., 2015) has led to the adoption of development operations (portmanteau *DevOps*) principles from software

engineering. DevOps aims at reducing the transitional periods between the commitment of a change to a system and the deployment of that change, while ensuring the overall quality of the whole development process (Bass et al., 2015). Corresponding extensions are coined as *Ops* monikers, e.g. data operations (portmanteau DataOps, Palmer, 2015; Vorhies, 2017), machine learning operations (portmanteau MLOps, Fursin, 2021; John et al., 2021; Makinen et al., 2021) and artificial intelligence operations (portmanteau AIOps, CXOtoday, 2017; Bowles, 2020) - just to name the most relevant ones in the context of data science. Similar ideas can be found in the CRISP-DM-based "Cross-Industry Standard Process Model for the Development of Machine Learning Applications with Quality Assurance Methodology" (CRISP-ML(Q)) (Studer et al., 2021).

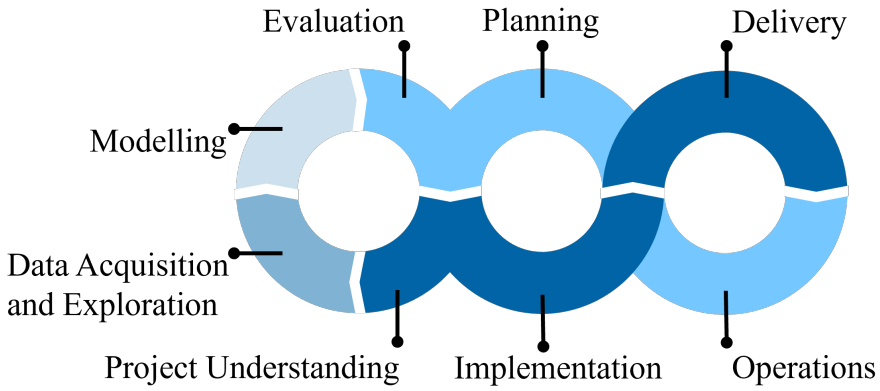

Figure 2: Illustration of the data product development process where the segments coloured in different shades of blue indicate the individual process phases and the white carets mark the corresponding transitions.

Aiming for a graphical illustration of the principles established so far, while providing a more detailed view on the data product development process (upper arrow in Figure 1) lets us propose the diagram in Figure 2 as a variation of the illustrations in other sources (Azure Architecture Center, 2021; Merritt, 2020; Visengeriyeva et al., 2021): The circle furthest to the left follows the CRISP-DM model with the phases *project understanding*, *data acquisition and exploration*,

modelling and *evaluation*. The primary goals of the *project understanding* phase are the comprehension of the problem statement, the benefits, the associated risks and required effort for a successful project. This results in the project goals and associated success criteria. During the *data acquisition and exploration* data sources are identified, followed by data acquisition and an exploratory analysis do determine their quality and suitability for solving the given problem statement. Depending on the specific methods used, data preparation steps are applied at the beginning of the *modelling* phase before pattern recognition, segmentation and prediction methods can be applied. The *evaluation* answers the question whether the modelling objectives are achieved by the results up to that point. A negative decision leads to a new iteration through the phases on the left or, in the worst case, to the abandonment of the project. In case of a positive decision, the work transitions into the deployment phases on the right, represented by the two interconnected circles in the form of the lying eight inspired by common illustrations of the DevOps process.

The results of the *evaluation* inform the *planning* phase, during which the team defines the capabilities of the system to be delivered to the users, how operational stability can be monitored and maintained and how user feedback on the quality of the solution can be collected. The deployment often requires substantial technical *implementation* efforts to provide the desired user experience around the modelling components of the solution. *Delivery* is the process of making the releases of the solution available to the users. This might include the appropriate infrastructure to enable the use of the product. Furthermore, the integration of the product in the corresponding (business) processes and user communities need to be addressed. The *operations* phase involves maintaining, monitoring, and troubleshooting the system. Monitoring can cover non-functional aspects related to the stability of the solution, but also any form of user feedback, environment and context changes affecting the data generating processes or the value proposition of the data product. The findings from all the preceding phases inform the *planning* of the next iteration.

Although inspired by modern software engineering principles, this description generalises deliberately beyond pure software delivery. In the case of the latter, more granular DevOps-activities can be specified, such as *code*, *build*, *test*, *package*, *release*, *configure* and *monitor*.

2 Data Science and the Scientific Method

The systematic development of knowledge, which is typically embedded in ontological frameworks established by communities of specialists (Hoyningen-Huene, 2008), is broadly accepted as a principal goal of scientific methodology. However, taking into account the variability of activities and practices over time and context in the field (Godfrey-Smith, 2003; Hepburn and Andersen, 2021), recent contributions to the philosophy of science argue that no distinct and static scientific method to achieve this goal can be identified (Weinberg, 1995). *Hey et al.* classify data science as the fourth paradigm (Hey et al., 2009; Bell et al., 2009) in the series:

1. Experimental science
2. Theoretical science
3. Computational science
4. Data science

Newton's law of motion published in the 17th century (Newton, 1687) marks the transition from purely experimental to include theoretical science as well. With developments around computational science in the 20th century, the properties of theoretical systems which do not allow for closed-form solutions have become tractable by means of computational methods. However, computational modelling and simulations might still constitute a special form of experimentation, which can be used to test and refine hypotheses in an iterative way, following the ideal of the hypothetico-deductive model (Hempel, 1966). In contrast, data science work typically starts from the data itself (compare the arguments around data centrism brought forward by Stadelmann, Klamt, and Merkt (Stadelmann et al., 2022)) rather than a hypothesis. This means, that data scientists often are not involved in the design or execution of the experiments and, therefore, have only incomplete knowledge on the data generating process(es). They work with observational data when it is impossible (at least within the constraints of the respective project) to conduct controlled experiments in the first place. Such data-driven (inductive) approaches focus primarily on discovering correlations and making predictions. Establishing causal relationships and finding mechanistic explanations for the observed phenomena is often of lower priority or might not be included in the project goals at all (Anderson, 2008). For

example, in predicting the breakdown of wind turbines (Zgraggen et al., 2021), the primary goal is to produce a warning signal with sufficient lead time to ensure the just-in-time availability of the required replacement parts, rather than understanding the mechanics producing the faults. Other examples can be found in molecular biology and neighbouring fields (Leonelli, 2012; Strasser, 2012). Nonetheless, data science in general is very pragmatic in the choice of methods and therefore data-driven approaches might not necessarily be completely free of hypotheses.

Brodie (2019) argues that data science is not a science because of its focus on specific use cases. This argument could be brought forward for other engineering fields or applied sciences as well, which are generally accepted to adhere to scientific principles (Cambridge English Dictionary, 2021). We see no specific reasons to separate data science from those disciplines.

However, the paradigm shifting characteristic of data science comes with specific challenges around the increasingly complex data acquisition, processing and modelling pipelines, which are often broken down into distinct steps conducted by different parties who do not necessarily disclose all the employed tools and configurations of the computational environments. Without a seamless documentation of the development process from data generation and acquisition to processing, analyzing and modelling the data, to interpreting the results and deploying the data product, traceability of the information and repeatability can not be guaranteed, which does not comply with standards in the scientific community (National Academies of Sciences Engineering and Medicine, 2019; Liu et al., 2022). Furthermore, incomplete knowledge on the nature of the data, its context and generating process(es) might lead to erroneous interpretations of statistical quantities as well as relying on spurious correlations (Calude and Longo, 2016; Blalock, 1971).

The issues related to the problem of incomplete documentation can be illustrated on a building data set released for public use by Microsoft Research in 2018 (Microsoft, 2018). It consists of 125 million building footprints for the United States extracted from optical satellite imagery by means of computer vision methods.

The value of the dataset for further analysis and modelling in various scientific, societal and economical contexts/applications is apparent. For example, it could

be combined with a wealth of other geographical features in OpenStreetMap (OSM, OpenStreetMap contributors, 2022). However, a number of problems appear:

1. No information is given on how the remote sensing data has been generated (e.g. sensor resolution) other than that the Bing Maps imagery is a composite of different sources.
2. The capture dates of the employed imagery is not consistently revealed across the dataset.
3. No information is given about the applied geo-referencing technique.
4. There is only a high-level reference to the network architecture and polygonisation algorithm used for model training and the extraction of the building footprints.

Such information would be critical to build a reliable data product from, for example, combining the Microsoft building footprint dataset with information in OSM.

3 Ops to the Rescue

According to Section 2 the primary issue of data science in its development to a truly scientific discipline seems to be a gigantic housekeeping problem: Consistently mapping all the data being used and produced during the development process onto a data model maintaining the links between products, raw data and configurations.

Current DevOps practices in IT operations already successfully address the same fundamental need by means of tools and processes assuring consistency, traceability, transparency and reliability, which in the end results in the desired delivery speed and quality (Forsgren et al., 2014, 2018; Linders, 2018; Overby, 2018). DevOps platforms provide the toolchains to enable security and regulatory compliance, effective collaboration, workflow automation and continuous improvement during the individual process phases *planning*, *development*, *delivery* and *operations*. The data associated to the configurations and individual executions of the pipelines is logged and contextualised. Each release is linked to the corresponding code changes (source code management

tools), build logs (continuous integration tools), executed test cases and results (continuous testing tools), software artifacts (container image registries), release approvals (change management and release automation systems), infrastructure configuration (infrastructure as code tools), and monitoring logs (application performance and end-user experience). The toolchains allow to trace back all relevant information when required, at least within the given access restriction on the particular platform.

Therefore, some critical components addressing the need for a complete documentation of data product development are established and the rising awareness among data scientists with respect to a more rigorous implementation of DevOps-principles to ensure consistency and stability in increasingly shorter release cycles for new ML features also acts in favor of scientific discipline. However, the challenge remains grand: As a next step it seems up to the community to establish a consensus on the format and operational aspects of a platform technology that supports collaboration across individual data science activities.

While the proposal for a concrete realisation is beyond the scope of the current article, a number of requirements can be identified to guide future research and development activities: The corresponding technical solution needs to enable compatibility and interoperability of the Dev- and DataOps-platforms. For example, a common data model is needed to generically represent the definitions of data processing and modelling pipelines, their runtime configurations and outputs. A standard interface should allow the transfer of such information between the different platforms. Furthermore, the immense amount and heterogeneity of the data associated with these objects will require innovative approaches to assist its exploration and interpretation in order to efficiently trace back information from previous steps whenever required so that products could be reliably built on top of each other in subsequent ventures by different teams. Other challenges are related to the protection of sensitive data while still making them available for analysis and modelling and allowing for reproducibility of the results. Privacy-enhancing technologies, such as fully homomorphic encryption schemes (Gentry, 2009; Xiao et al., 2012), aim at enabling efficient computation on encrypted data without disclosing sensitive information.

4 Concluding Remarks

In summary, the present article follows Hey et al. (2008) in claiming a *new paradigm of scientific discovery* for the field of data science where the activities start from data, rather than with the design of principled (computational) experiments to generate it. Data science is framed as an engineering discipline with the aim of developing use case-specific data products following a set of core principles:

- User-centricity.
  - Incorporating domain knowledge.
  - Embracing incremental development cycles.
  - Understanding deployment as integral part of the value proposition.
  - Transparent managing of the trade-off between accuracy and speed.
- Pragmatic balance between data- and hypothesis-driven approaches.
- Effective communication.
- Collaboration.

The apparent issues related to the seamless and practically usable documentation of the data product development process can be resolved by a more rigorous implementation of DevOps principles. Following *collaboration* as a core principle of data science activities, it seems up to the community to stipulate a consensus on the format and operational aspects of a platform technology that allows to consistently link data, models and products, along with a technical realisation.

References

Analog Devices Inc. (2004) The Data Conversion Handbook, 1st edn. Kester W (ed.), Newnes, Oxford. ISBN: 07-5067-841-0.

Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. WIRED. URL: https://www.wired.com/2008/06/pb-theory/.

Azure Architecture Center (2021) Machine learning operations (MLOps) framework to upscale machine learning Lifecycle with Azure Machine Learning. URL: https://docs.microsoft.com/en-us/azure/architecture/example-scenario/mlops/mlops-technical-paper.

Bansak K, Ferwerda J, Hainmueller J, Dillon A, Hangartner D, Lawrence D, Weinstein J (2018) Improving refugee integration through data-driven algorithmic assignment. Science 359(6373):325–329, American Association for the Advancement of Science. DOI: 10.1126/science.aao4408.

Bass L, Weber I, Zhu L (2015) DevOps: A Software Architect's Perspective, 1st edn. Addison-Wesley Professional, New Jersey. ISBN: 978-0-134049-84-7.

Bell G, Hey T, Szalay A (2009) Beyond the Data Deluge. Science 323(5919):1297–1298. ISSN: 0036-8075, DOI: 10.1126/science.1170411.

Blalock HM (1971) Causal Models in the Social Sciences, 1st edn. Aldine-Atherton, Chicago. ISBN: 978-0-202300-76-4.

Blei DM, Smyth P (2017) Science and data science. Proceedings of the National Academy of Sciences of the United States of America 114(33):8689–8692, Bickel PJ (ed.), National Academy of Sciences. ISSN: 1091-6490, DOI: 10.1073/pnas.1702076114.

Bowles J (2020) Moogsoft CEO Phil Tee on AIOps and service assurance in the age of digital transformation. In: diginomica. URL: https://diginomica.com/moogsoft-ceo-phil-tee-aiops-and-service-assurance-age-digital-transformation.

Braschler M, Stadelmann T, Stockinger K (2019) Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business, 1st edn. Springer International Publishing, Cham, chap. 2, pp. 17–29. ISBN: 978-3-030118-21-1, DOI: 10.1007/978-3-030-11821-1_2.

Brodie ML (2019) What Is Data Science? In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business, 1st edn. Springer International Publishing, Cham, chap. 8, pp. 101–130. ISBN: 978-3-030118-21-1, DOI: 10.1007/978-3-030-11821-1_8.

Calude CS, Longo G (2016) The Deluge of Spurious Correlations in Big Data. Foundations of Science 22(3):595–612, Springer Netherlands. ISSN: 1572-8471, DOI: 10.1007/s10699-016-9489-4.

Cambridge English Dictionary (2021) Engineering. Cambridge University Press, Cambridge. URL: https://dictionary.cambridge.org/dictionary/english/engineering.

Chaslot G (2019) The Toxic Potential of YouTube's Feedback Loop. WIRED. URL: https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/.

Christen M, Blumer H, Hauser C, Huppenbauer M (2019) The Ethics of Big Data Applications in the Consumer Sector. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science, 1st edn. Springer, Cham, chap. 10, pp. 161–180. ISBN: 978-3-030118-20-4, DOI: 10.1007/978-3-030-11821-1_10.

Conway D (2010) The Data Science Venn Diagram. URL: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Cukier K, Mayer-Schoenberger V (2013) The Rise of Big Data: How It's Changing the Way We Think About the World. Foreign Affairs 92(3):28–40. DOI: 10.1515/9781400865307-003.

CXOtoday (2017) Algorithmic IT Operations Drives Digital Business: Gartner. In: CXOtoday. URL: https://web.archive.org/web/20180128074703/http://www.cxotoday.com/story/algorithmic-it-operations-drives-digital-business-gartner/.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3):37–54. DOI: 10.1609/aimag.v17i3.1230.

Floridi L, Taddeo M (2016) What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2083), The Royal Society. ISSN: 1364-503X, DOI: 10.1098/RSTA.2016.0360.

Forsgren N, Kim G, Kersten N, Humble J (2014) 2014 State of DevOps Report. Tech. Rep., Puppet Labs, IT Revolution Press and ThoughtWorks. URL: https://services.google.com/fh/files/misc/state-of-devops-2014.pdf.

Forsgren N, Humble J, Kim G (2018) Accelerate: The Science of Lean Software and DevOps: Building and Scaling High Performing Technology, 1st edn. IT Revolution Press, Portland. ISBN: 978-1-942788-33-1.

Frické M (2009) The Knowledge Pyramid: A Critique of the DIKW Hierarchy. Journal of Information Science 35(2):131–142. ISSN: 0165-5515, DOI: 10.1177/0165551508094050.

Fursin G (2021) Collective knowledge: organizing research projects as a database of reusable components and portable workflows with common interfaces. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379(2197). DOI: 10.1098/rsta.2020.0211.

Gentry C (2009) Fully Homomorphic Encryption Using Ideal Lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, NY, USA, STOC '09, pp. 169–178. ISBN: 978-1-605585-06-2, DOI: 10.1145/1536414.1536440.

Godfrey-Smith P (2003) Theory and Reality: An Introduction to the Philosophy of Science, 1st edn. University of Chicago Press, Chicago. ISBN: 978-0-226300-61-0, DOI: 10.7208/9780226300610.

Hempel CG (1966) Philosophy of Natural Science, 1st edn. Prentice Hall, Upper Saddle River. ISBN: 01-3663-823-6.

Hepburn B, Andersen H (2021) Scientific Method. In: Zalta EN (ed.), The Stanford Encyclopedia of Philosophy, summer 2021 archive edn. Metaphysics Research Lab, Stanford University, Stanford. ISSN: 1095-5054, URL: https://plato.stanford.edu/archives/sum2021/entries/scientific-method/.

Hey J, Yu J, Agogino AM (2008) Design Team Framing: Paths and Principles. In: Proceedings of the ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 4, pp. 409–420. DOI: 10.1115/DETC2008-49383.

Hey T, Tansley S, Tolle K (eds.) (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery, 1st edn. Microsoft Research, Redmond. ISBN: 978-0-982544-20-4.

Hoyningen-Huene P (2008) Systematicity: The Nature of Science. Philosophia 36(2):167–180. ISSN: 0048-3893, DOI: 10.1007/S11406-007-9100-X.

John MM, Olsson HH, Bosch J (2021) Towards MLOps: A Framework and Maturity Model. In: Baldassarre MT, Scanniello G, Skavhaug A (eds.), 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, Piscataway, New Jersey, pp. 334–341. ISBN: 978-1-665427-06-7, DOI: 10.1109/SEAA53835.2021.00050.

Leonelli S (2012) Introduction: Making sense of data-driven research in the biological and biomedical sciences. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43(1):1–3, Pergamon. ISSN: 1369-8486, DOI: 10.1016/J.SHPSC.2011.10.001.

Linders B (2018) Q&A on the Book Accelerate: Building and Scaling High Performance Technology Organizations. In: InfoQ. URL: https://www.infoq.com/articles/book-review-accelerate/.

Liu J, Carlson J, Pasek J, Puchala B, Rao A, Jagadish HV (2022) Promoting and Enabling Reproducible Data Science Through a Reproducibility Challenge. Harvard Data Science Review 4(3):1–22. DOI: 10.1162/99608f92.9624ea51.

Loukides M (2010) What is data science? In: Radar. URL: https://www.oreilly.com/radar/what-is-data-science/.

Loukides M (2011) The evolution of data products. In: Radar. URL: https://www.oreilly.com/radar/evolution-of-data-products/.

Makinen S, Skogstrom H, Laaksonen E, Mikkonen T (2021) Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? In: Bosch J, Crnkovic I, Holmström H, Lwakatare LE (eds.), 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN), IEEE, Piscataway, New Jersey, pp. 109–112. ISBN: 978-1-665444-71-2, DOI: 10.1109/WAIN52551.2021.00024.

Maloberti F, Davies AC (2016) A Short History of Circuits and Systems, 1st edn. River Publishers, Gistrup. ISBN: 87-9337-971-4.

Meierhofer J, Stadelmann T, Cieliebak M (2019) Data Products. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science: Lessons Learned for the Data-Driven Business, 1st edn. Springer International Publishing, Cham, chap. 4, pp. 47–61. ISBN: 978-3-030118-21-1, DOI: 10.1007/978-3-030-11821-1_4.

Merritt R (2020) What Is MLOps? In: NVIDIA Blog. URL: https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/.

Microsoft (2018) USBuildingFootprints. In: GitHub. URL: https://github.com/microsoft/USBuildingFootprints.

National Academies of Sciences Engineering and Medicine (2019) Reproducibility and Replicability in Science, 1st edn. National Academies Press, Washington. ISBN: 03-0948-616-5, DOI: 10.17226/25303.

Newton I (1687) Philosophiae Naturalis Principia Mathematica, 3rd edn. Watchmaker Publishing. ISBN: 978-1-603863-79-7.

OpenStreetMap contributors (2022) OpenStreetMap. In: OpenStreetMap. URL: https://www.openstreetmap.org/.

Osterwalder A, Pigneur Y, Bernarda G, Smith A, Papadakos T (2014) Value Proposition Design: How to Create Products and Services Customers Want, 1st edn. Wiley, Hoboken, New Jersey, United States. ISBN: 978-1-118968-05-5.

Overby S (2018) 7 takeaways to "Accelerate" your DevOps. In: TechBeacon. URL: https://techbeacon.com/devops/7-takeaways-accelerate-your-devops.

Palmer A (2015) From DevOps to DataOps. URL: https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/.

Rowley J (2007) The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science 33(2):163–180. ISSN: 0165-5515, DOI: 10.1177/0165551506070706.

Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo JF, Dennison D (2015) Hidden Technical Debt in Machine Learning Systems. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds.), Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, MIT Press, Montreal, Vol. 28, pp. 2503–2511. DOI: 10.5555/2969442.2969519.

Shearer C (2000) The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing 5(4):13–22, THE DATA WAREHOUSE INSTITUTE. URL: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf.

Souza VMA, dos Reis DM, Maletzke AG, Batista GEdAPAB (2020) Challenges in benchmarking stream learning algorithms with real-world data. Data Mining and Knowledge Discovery 34(6):1805–1858, Springer. ISSN: 1384-5810, DOI: 10.1007/S10618-020-00698-5.

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: Proceedings of the 9th European Computer Science Summit. URL: https://digitalcollection.zhaw.ch/handle/11475/4172.

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 8(2). DOI: 10.5445/IR/1000143637.

Strasser BJ (2012) Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43(1):85–87, Pergamon. ISSN: 1369-8486, DOI: 10.1016/J.SHPSC.2011.10.009.

Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, Müller KR (2021) Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Machine Learning and Knowledge Extraction 3(2):392–413. DOI: 10.3390/make3020020.

Tsymbal A (2004) The problem of concept drift: definitions and related work. Tech. Rep., Department of Computer Science, Trinity College Dublin, Dublin, pp. 1–7. URL: https://www.scss.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf.

Visengeriyeva L, Kammer A, Bär I, Kniesz A, Plöd M (2021) CRISP-ML(Q). The ML Lifecycle Process. URL: https://ml-ops.org/content/crisp-ml.

Vorhies W (2017) DataOps – It's a Secret. In: Data Science Central. URL: https://www.datasciencecentral.com/profiles/blogs/dataops-it-s-a-secret.

Wang R, Shivanna R, Cheng DZ, Jain S, Lin D, Hong L, Chi EH (2021) DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In: Leskovec J, Grobelnik M, Najork M (eds.), WWW '21: Proceedings of the Web Conference 2021, Association for Computing Machinery, New York, pp. 1785–1797. ISBN: 978-1-450383-12-7, DOI: 10.1145/3442381.3450078.

Weinberg S (1995) The Methods of Science...And Those by Which We Live. In: Shaw P, Iannone C, Short T, London H (eds.), What Do the Natural Sciences Know and How Do They Know It?, Symposium, National Association of Scholars, New Brunswick, Vol. 8, pp. 7–13.

Widmer G, Kubat M (1996) Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning 23(1):69–101, Springer. ISSN: 1573-0565, DOI: 10.1023/A:1018046501280.

Widmer M, Hegy S (2019) Legal Aspects of Applied Data Science. In: Braschler M, Stadelmann T, Stockinger K (eds.), Applied Data Science, 1st edn. Springer, Cham, chap. 5, pp. 63–78. ISBN: 978-3-030118-20-4, DOI: 10.1007/978-3-030-11821-1_5.

Xiao L, Bastani O, Yen IL (2012) An Efficient Homomorphic Encryption Protocol for Multi-User Systems. Cryptology ePrint Archive 193. URL: `https://eprint.iacr.org/2012/193`.

Yesilada M, Lewandowsky S (2022) Systematic review: YouTube recommendations and problematic content. Internet Policy Review 11(1):1–22. ISSN: 2197-6775, DOI: 10.14763/2022.1.1652.

Zgraggen J, Ulmer M, Jarlskog E, Pizza G, Huber LG (2021) Transfer Learning Approaches for Wind Turbine Fault Detection using Deep Learning. In: Do P, King S, Fink O (eds.), Proceedings of the European Conference of the PHM Society 2021, PHM Society. ISBN: 978-1-936263-34-9, DOI: 10.36001/phme.2021.v6i1.2835.

Zins C (2007) Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology 58(4):479–493. ISSN: 1532-2882, DOI: 10.1002/asi.20508.