VOLUME 5 · ISSUE 04

BALSILLIE PAPERS

Proper Scoping of Data Science: A Guide for Policy Makers

M. Tamer Özsu

May 15, 2023

Data science has emerged as a promising field to leverage the large volumes of data collected from multiple sources. Many organizations have initiated data science projects and governments have produced policy statements. Unfortunately, there is significant variability in many of these policies, partially due to the lack of a consistent and clear understanding of the field. This paper provides a systematic framing of the field to assist policy makers in developing proper initiatives.

The last two decades have witnessed the increasing appreciation of data as a fundamental ingredient of many daily activities. *The Economist* has identified data as "the most valuable resource."[1] The World Economic Forum stated, "At the heart of the digital economy and society is the explosion of insight, intelligence and information — data."[2] "Data is the new oil"[3] is a phrase commonly heard. Although sometimes controversial,[4] similar opinions have been expressed repeatedly: for example, "data is the new currency,"[5] and "data is a commodity like gold."[6]

The increased appreciation of data as an important asset results from a technology-push/application-pull. The development, and increased deployment, of information and communication technologies generates ever-increasing volumes of different types of data (structured, text, audio, video, microblogs, and so forth). The increasing deployment of sensors and the move toward the Internet-of-Things have increased the need to deal with *streaming data*. The multiplicity of data sources, and their varying degrees of trustworthiness, have highlighted the importance of data veracity. This has led to the "big data revolution," with big data typically characterized by four Vs: volume, variety, velocity, and veracity.[7] The technological requirements of properly managing this data are challenging.

Concurrently, there is an application pull. Organizations are recognizing the importance of properly processing this data to obtain actionable insights. The real value of data is derived when it is properly processed. Consequently, we increasingly hear of the "data-driven" approach. The Organization for Economic Cooperation and Development (OECD) has identified "data-driven innovation" as having a central role in twenty-first century economies.[8] Brynjolfsson et al[9] proposed a measure for data-driven decision making to measure the impact of the use of data on a company's productivity. Their analysis of more than 170 publicly traded firms in the United States showed a 5-6% increase in those firms' output and productivity.

---

[1] *The Economist*, "The world's most valuable resource" (cover story), May 6, 2017, accessed March 21, 2023, https://www.economist.com/weeklyedition/2017-05-06.
[2] World Economic Forum, *A New Paradigm for Business Data*, July 29, 2020, accessed March 21, 2023, https://www.weforum.org/reports/new-paradigm-for-business-of-data/.
[3] Generally attributed to Clive Robert Humby. See Ritu Janegar, "Data is the new oil," *The Commerce Society*, Shri Ram College of Commerce, accessed March 21, 2023, https://comsocsrcc.com/data-is-the-new-oil.
[4] See, for example, Bernard Marr, "Here is why data is not the new oil," *Forbes*, March 2018, accessed March 21, 2023, https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/; Antonio Garcia March 21, 2023, https://www.wired.com/story/no-data-is-not-the-new-oil/
[5] Steven Burke, "CEO Antonio Neri: 10 boldest statements from HPE Discover 2021," *CRN*, June 22, 2021, accessed March 21, 2023, https://www.crn.com/slide-shows/storage/ceo-antonio-neri-10-boldest-statements-from-hpe-discover-2021.
[6] Matt Shephard, "Is Data the New Gold?" *CEO Today*, April 2018, accessed March 21, 2023, https://www.ceotodaymagazine.com/2018/04/is-data-the-new-gold/.
[7] For further reading, see Chapter 10 — "Big Data Processing" in M. Tamer Özsu and Patrick Valduriez, *Principles of Distributed Data Management* (Cham: Springer, 2020).
[8] OECD, *Data-Driven Innovation: Big Data for Growth and Well-Being*, October 6, 2015, https://doi.org/10.1787/9789264229358-en.
[9] Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" SSRN, April 24, 2011, http://dx.doi.org/10.2139/ssrn.1819486.

The expectations of the value of data have led to significant activity at the governmental level. The United States has a number of initiatives including the White House Open Data Initiative,[10] the Department of State Data Strategy,[11] and the more recent National Artificial Intelligence Task Force.[12] The European Union's data strategy[13] includes several initiatives around data science. Canada has released a federal digital charter[14] that addresses the importance of data in the Canadian economy and establishes a vision for a data economy built around ten principles: universal access; safety and security; control and consent; transparency, portability and interoperability; open and modern digital government; a level playing field; data and digital for good; strong democracy; free from hate and violent extremism; and strong enforcement and real accountability. The importance of this charter is its interdisciplinarity, which dovetails well with data science.

Despite the interest and activity around data, and the persistent reference to data science as being important (even critical), the field is not well defined or understood. Part of the difficulty is the interchangeable and confusing use in popular press of the terms "big data," "data analytics," and "data science." This confusion usually extends to the technical literature as well. Another difficulty is the lack of differentiation between artificial intelligence (AI), machine learning (ML), data mining (DM), and data science (DS). This has played a significant and negative role in establishing proper DS programs and policies which, in turn, has negatively impacted the establishment of a proper DS ecosystem. Often, investments in AI are confused with DS.

The objective of this paper is to establish a coherent and internally consistent framework for DS. This should bring some clarity to the discussions around this topic and enable policy makers to make reasonable judgements regarding the establishment of a thriving DS ecosystem. DS policies and AI policies should be complementary and should reinforce each other. This paper is *not* a discussion of how DS can be applied to public policy design and related applications — there is a significant body of work in that space.[15,16] This paper is also *not* a primer on the DS tools for those working on public policy

---

[10] White House, "Open Government Initiative," May 2013, accessed March 21, 2023, obamawhitehouse.archives.gov/open.
[11] Matthew Gravis, "US Department of State launches first-ever data strategy," September 2021, accessed March 21, 2023, https://www.state.gov/dipnote-u-s-department-of-state-official-blog/u-s-department-of-state-launches-first-ever-data-strategy/.
[12] White House, "The Biden administration launches the national artificial intelligence research resource task force," press release, June 10, 2021, accessed March 21, 2023, https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.
[13] European Union, "A European Strategy for data," May 2022, accessed March 21, 2023, https://digital-strategy.ec.europa.eu/en/policies/strategy-data.
[14] Government of Canada, "Canada's digital charter: Trust in a digital world," 2020, accessed March 21, 2023, https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter-trust-digital-world.
[15] Michela Arnaboldi and Giovanni Azzone, "Data science in the design of public policies: dispelling the obscurity

https://doi.org/10.1016/j.heliyon.2020.e04300.
[16] Amar Numanović, "Data Science: The Next Frontier for Data-Driven Policy Making?" 2017, accessed March 23, 2023, http://www.policyhub.net/en/experience-and-practice/212.

applications — that is also a different discussion.[17,18] It is about how policy makers should think about DS as they develop policies and programs that properly promote the field.

What Is Data Science?

The origin of the term "data science" is not entirely clear. Since data is central to both statistics and computing, both communities have a long history of activity in this domain. Its separation from "data analytics," a term first introduced by Tukey in 1962,[19] has been a topic of significant discussion and concern as it leads to statisticians questioning whether their activities were not data science all along.[20]

A comprehensive working definition that captures the essence of the field would be the following:

> *Data science is a data-based approach to problem solving by analyzing and exploring large volumes of possibly multi-modal data, extracting knowledge and insight from it, and using information for better decision making. It involves the process of collecting, preparing, managing, analyzing, explaining, and disseminating the data and analysis results.*

A few salient points would benefit from highlighting. The first is the distinction between data analysis (or analytics) and data science. The modern understanding of data science is broader than data analytics and encompasses the latter as one of its activities. More on this in the next section.

A second, and related, point is the relationship between DS, ML, and DM. The term "data science" is frequently colloquially used to mean data analytics using ML/DM. DS is not a subfield of ML/DM, nor is it synonymous with these. More broadly, it is not the case that DS is a subtopic of AI — a common claim originating from confusion on boundaries. AI and DS are conceptually different fields that overlap when ML/DM techniques are used in data analytics, but that otherwise have their own broader concerns.

Third, DS is different from "big data processing." A good analogy is that big data is like raw material; it has considerable promise and potential if one knows what to do with it. DS gives it purpose, specifying how to process it to extract its full potential and to what end. It does this typically in an application-driven manner, allowing applications to frame the study objective and question.

Fourth, there is a process view of DS. Any DS activity starts with a problem definition based on which data needs are specified, data sources are identified, data are collected, integrated, and prepared for analysis. The analysis generates some (possibly actionable) insights, or it may identify a need for the refinement of the problem definition or to adjustments to data needs and sources. This process with the feedback loops is usually called the "data science lifecycle."

Finally, note the use of the term "data-based" in the definition rather than the more common "data-driven." The latter has frequently been interpreted as "data should be the main (only?) basis for decisions" since "data speaks for itself." This is wrong — data certainly holds facts and can reveal a story, but it only speaks through those who interpret it and who can potentially introduce biases. Therefore, data should be

---

[17] Jeffrey C. Chen, Edward A. Rubin, and Gary J. Cornwall, *Data Science for Public Policy* (Cham: Springer, 2021).
[18] Ken Steif, *Public Policy Analytics* (Boca Raton: CRC Press, 2022).
[19] John W. Tukey, "The future of data analytics," *The Annals of Mathematical Statistics* 33 (1962): 1–67.
[20] Marie Davidian, "Aren't *We* Data Science?" *AMSTAT News: The Membership Magazine of the American Statistical Association*, July 1, 2013, accessed March 23, 2023, https://magazine.amstat.org/blog/2013/07/01/datascience/.

one of the inputs to decision making, but not the only one. Furthermore, "data-driven" has come to mean that it is possible to take data and analyze it by using automated tools to generate automated actions. This also is problematic. Although DS has significant potential, and successful DS applications are plenty, there are sufficient misuses of data to give us pause and concern: Google's algorithmic detection of influenza spread using social media data is one prominent example[21]; the Risk Needs Assessment Test used in the US justice system is another[22]. Therefore, "data-based" is the preferable phrase that signals that DS deployments are aids to the decision maker, not decision makers themselves.[23]

Data Science Scoping

Scoping of DS is important to establish the pillars of the field and to define the DS ecosystem. DS has four pillars that make up its core: data engineering, data analytics, data protection, and ethics (Figure 1). As noted earlier, although the term "data science" is frequently used to refer only to data analysis, the scope is wider, and the contributing elements of the field need to be fully recognized to best leverage its benefits.

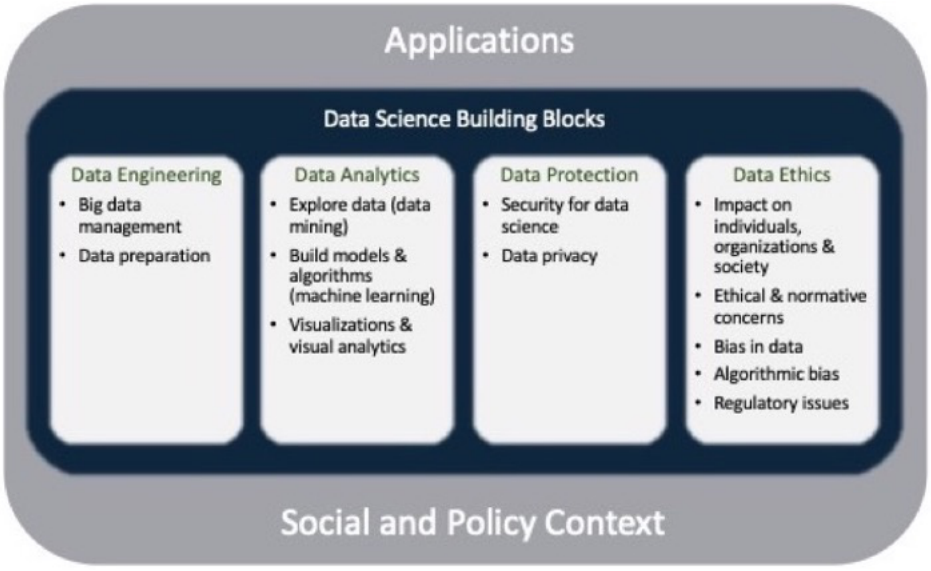

Figure 1. The Data Science Ecosystem

The core is in close interaction with application domains that have the dual function of informing the appropriate technologies, tools, algorithms, and methodologies that should be useful to develop, and leveraging these capabilities to solve their problems.

[21] David Lazer and Ryan Kennedy, "What We Can Learn from the Epic Failure of Google Flu Trends," *Wired*, 1 October 2015, https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.
[22] Nathan James, "Risk and Needs Assessment in the Federal Prison System," Congressional Research Service Report R44087, 10 July 2018, https://sgp.fas.org/crs/misc/R44087.pdf .
[23] Other suitable terms that have been used are "data-enhanced" or "data-enabled."

DS application deployments are highly sensitive to the existing social and policy context and these influence both the core technologies and the application deployments.

Data Engineering

The role of data (in particular, that of big data) was discussed previously. Data engineering in DS addresses two main concerns: the management of big data, including the computing platforms for its processing; and the preparation of data for analysis.

Managing big data is challenging but critical in DS application development and deployment. These data characteristics are quite different from those that traditional data management systems are designed for, and require new systems, methodologies, and approaches. What is needed is a data management platform that provides appropriate functionality and interfaces for conducting data analysis, executing declarative queries, and sophisticated search.

Data preparation[24,25] is typically understood as the process of dataset selection, data acquisition, data integration, and data quality enforcement. Applying appropriate analysis to the integrated data will provide new insights that can improve organizational effectiveness and efficiency and result in evidence-informed policies. However, for this analysis to yield meaningful results, the input data must be appropriately prepared and trustable. It really makes little difference how good the analysis model is; if the input data is not clean and trustable, then the results will not be of much value. The old adage of "garbage in, garbage out" is real in data science.

Data Analytics

Data analytics[26] is the application of statistical and ML techniques to draw insights from the data under study and to make predictions about the behaviour of the system under study.

Data analysis can target either inferencing or prediction. Inference is based on building a model that describes a system behaviour by representing the input variables and relationships among them. Prediction goes further and identifies the courses of action that might yield the "best" outcomes. This categorization can be made more fine-grained by identifying four classes of analysis: descriptive, diagnostic, predictive, and prescriptive. The latter two are sometimes called advanced analytics.

Data Protection

DS's reliance on large volumes of varied data from many sources raises important data protection concerns. The scale, diversity, and the interconnectedness of data (as, for example, in online social

[24] Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras, *Principles of Data Wrangling: Practical Techniques for Data Preparation*, O'Reilly Media, 2017.
[25] Ihab F. Ilyas and Xu Chu, *Data Cleaning* (New York: ACM Books, 2019).
[26] The literature on this topic is rich. The following is a good starting point for general reading: Anil Maheshwari, *Data Analytics Made Accessible*, accessed March 23, 2023, https://tinyurl.com/ywnjh8se.

networks) requires revisiting the data protection techniques that have been mostly developed for corporate data.[27,28]

It is customary to discuss the relevant issues under the complementary topics of data security and data privacy. The former protects information from any unauthorized access or malicious attacks, while the latter focuses on the rights of users and groups over data about themselves. Data security typically deals with data confidentiality, access control, infrastructure security, and system monitoring, and uses technologies such as encryption, trusted execution environments, and monitoring tools. Data privacy, on the other hand, deals with privacy policies and regulations, data retention and deletion policies, data subject access requirement policies, management of data use by third parties, and user consent; it also involves privacy-enhancing technologies.

Data Science Ethics

The fourth building block of DS is ethics. In many discussions, ethics is bundled with a discussion of data privacy. The two topics certainly have a strong relationship, but they should be considered separate pillars of the DS core.

DS ethics has three dimensions: data, algorithms, and practice. The ethics of data refers to the ethical problems posed by the collection and analysis of large datasets and on issues arising from the use of big data in diverse sets of applications. The ethics of algorithms addresses concerns arising from the increasing complexity and autonomy of algorithms, their fairness, bias, equity, validity, and reliability. Finally, the ethics of practices addresses the questions concerning the responsibilities and liabilities of people and organizations in charge of data processes, strategies, and policies. The growing research in AI ethics tackles many of these issues.[29,30]

Social and Policy Context

As noted earlier, DS deployments are highly sensitive to the societal and policy contexts in which they are deployed. For example, what can be done with data differs in different jurisdictions. The context can be legal, establishing legal norms for DS deployments, or it can be societal in identifying what is socially acceptable. Furthermore, there are significant intersections between social science and humanities and the core issues in DS. There are four central concerns: (1) data *ownership*, access, and use, in particular in

---

[27] Elisa Bertino and Elena Ferrari, "Big Data Security and Privacy," in *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, Sergio Flesca, Sergio Greco, Elio Masciari, and Domenico Saccà (eds) (Cham: Springer International Publishing, 2018), 425–439, https://doi.org/10.1007/978-3-319-61893-7_25.
[28] José Moura and Carlos Serrão, "Security and Privacy Issues of Big Data," in *Cloud Security: Concepts, Methodologies, Tools, and Application* (IGI Global, 2019), 1598–1630, https://doi.org/10.4018/978-1-5225-8176-5.ch080.
[29] Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (Oxford Academic, 2020), https://doi.org/10.1093/oxfordhb/9780190067397.001.0001.
[30] Paul W. Grimm, Maura R. Grossman, and Gordon V. Cormack, "Artificial Intelligence as Evidence," *Northwestern Journal of Technology and Intellectual Property* 9, no. 1 (2021), https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2.

terms of how individual data is generated, who has access to it, who owns it, and, by extension, who profits from it; (2) ensuring diverse and equitable *representation* at all stages of the DS workflow (lifecycle); (3) *regulation and accountability* to ensure transparency and explainability of the analysis, the algorithms applied, and how they lead to specific outputs and recommendation, and fair distribution and sharing of benefits of DS; and (4) the integration of DS in the analysis and formation of *public policy*. Obviously, there is overlap between these and the data ethics concerns.

Data Science Is Interdisciplinary

DS is interdisciplinary, as noted multiple times in this paper. According to Choi and Pak, "Interdisciplinarity analyzes, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole."[31] Furthermore, "interdisciplinarity would arise in a near symmetrical way when two or more disciplines converge in a given field... This convergence can lead to a practical and theoretical integration of the disciplines involved, which would be unified."[32] This is indeed a primary characteristic of DS as a field — it combines and integrates a number of different fields (Figure 2).

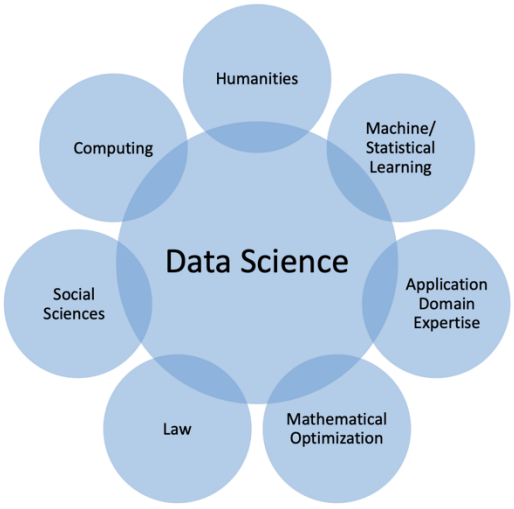

Figure 2. A Unifying View of Data Science

It is important to recognize this characteristic and consider DS as a unifying field that connects a number of others, some of which are STEM and some of which are not. Within this context, it is helpful to consider the stakeholders in DS. At the risk of oversimplification, the following are three different constituencies that have an interest in DS. The first group consists of STEM people who focus on foundational techniques and the underlying principles (e.g., computer scientists, statisticians, mathematicians). The second group also consists of STEM people, but those who focus on scientific and engineering DS applications (e.g., biologists, ecologists, earth and environmental scientists, health

[31] Bernard C. K. Choi and Anita W. P. Pak, "Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness," *Clinical and Investigative Medicine* 29, no. 6 (2006): 351–364.
[32] David Alvargonzález, "Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences," *International Studies in the Philosophy of Science* 25, no. 4 (2011): 387–403, https://doi.org/10.1080/02698595.2011.623366.

scientists). The final group are non-STEM people who focus on social, political, and societal aspects. It is important to be inclusive of all these stakeholders in discussions around DS, while at the same time establishing a recognizable core of the field. This is a hard balance to maintain.

Arguably, these observations beg the question of *who is a data scientist*? The DS ecosystem discussion identifies the topics of interest while the stakeholder classification provides an orthogonal specification. People engaged in data engineering, data analytics, and data protection fall into the first stakeholder group; people who work in data ethics can be either in the first or the third category depending on their interest; the third stakeholder group further includes those who study the social impact of DS and its policy context; finally, the second stakeholder group predominantly comes from the application domains. This characterization can be helpful in identifying the competencies that are required for a data scientist: in-depth knowledge of at least one of data engineering or data analytics pillars (expert level); working knowledge of the other three pillars; in-depth knowledge of at least one, but preferably multiple, application areas (almost expert level); and the ability to work in a team and effectively communicate, given the interdisciplinarity of the field.

Individual DS projects should ideally be populated to reflect the various stakeholder groups. For the success of such projects, it is important for the core technologists (first group) to have a good understanding of the driving application in addition to the technical challenges, while application experts (second group) should have a sufficient understanding of the technology to fruitfully engage in discussions regarding what solutions would be suitable and to have the wherewithal to deploy the technical solutions that are developed. Embedding into the solutions policy constraints and alertness to societal impact would require participation of experts from the third stakeholder group.

Conclusions

As increasing amounts of data are collected and stored at a pace that is more rapid than ever, the value of data as a central asset in an organization has grown. The field of DS is expected to enable us to leverage this data for meaningful and improved decisions and outcomes. Initial deployments of DS have demonstrated the potential benefits, but fuller exploitation is challenging without a clearer understanding of the scoping of the field and the identification of its relationship with related fields. Progress is further limited by the lack of a holistic approach to DS projects, the lack of knowledge exchange between experts in the sub-fields of DS, and the lack of tools, methods, and principles for understanding and translating these insights into improved decisions, products, systems, and policies. In this paper, a systematic study of DS is provided for policy makers to assist them in developing initiatives to fully exploit the benefits of DS.

Author's Note

This paper borrows liberally from M. T. Özsu, Data Science – A Systematic Treatment, *Communications of ACM*, 2023, https://doi.org/10.1145/3582491. Research supported in part of Natural Sciences and Engineering Research Council of Canada. Contact address: University of Waterloo, Cheriton School of Computer Science, Waterloo, Ontario N2L 3G1, Canada; tamer.ozsu@uwaterloo.ca

M. Tamer Özsu is a university professor in the Cheriton School of Computer Science, University of Waterloo. His research is on the data engineering aspects of data science. He is a Fellow of the Royal Society of Canada, Science Academy of Türkiye, American Association for the Advancement of Science, Asia-Pacific Association of AI, and Balsillie School of International Affairs, and a Life Fellow of IEEE and ACM.

BALSILLIE PAPERS

balsilliepapers.ca

ISSN 2563-674X

doi:10.51644/BAP54