



20 Years of Data Science: An Editorial

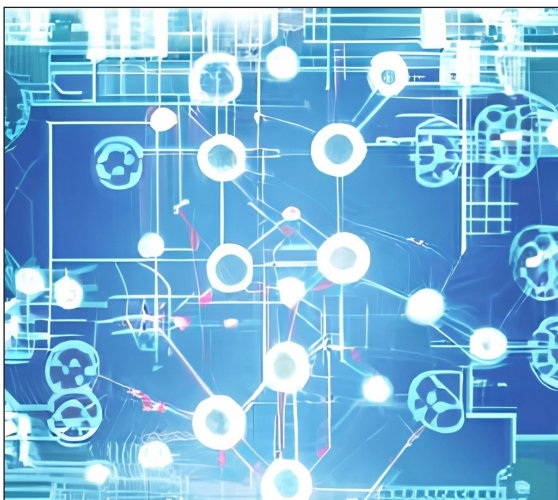
MARK A. PARSONS 

MATTHEW MAYERNIK 

*Author affiliations can be found in the back matter of this article

The *Data Science Journal* published its first issue in April 2002. It was a novel journal title for the time. Society was just beginning to recognize the potential of Jim Grey's fourth paradigm of data-intensive science (Hey et al. 2009). Francis 'Jack' Smith (2023), the journal's first editor-in-chief, notes that CODATA's Executive Board recognized that 'this title might be misunderstood. However, the majority felt that it was up to CODATA to ensure that it became understood.' Jack argues that over time the journal 'did just that'. We agree that the journal has contributed substantially to the 'science of data'. We also find, now living in the fourth paradigm with ever more pervasive and invasive data and rapidly growing computing power, that 'data science' is being used to refer to many types of scientific and technical work unforeseen when the journal launched.

As one example, 2022 and 2023 saw the explosion of data-driven text and image generators. These generators combine machine learning and other artificial intelligence techniques with large scale training data sets. Some of the computational techniques used in these systems existed 20 years ago, if in earlier forms. But the use of massive training data taken from the public internet has enabled the new ML/AI-based text and image generators to become far more effective than their predecessors. The image used in the DSJ landing page for this special collection, shown in Figure 1, was generated by the <https://www.crayon.com/> image generator, based on the prompt 'twenty years of data science'. Interestingly, all the images generated by crayon and depai.org had a similar aesthetic suggesting something of a bias. Bias in training data is a known problem for these types of systems, and presents an important emerging challenge for data science going forward.



COLLECTION:
20 YEARS OF
DATA SCIENCE

EDITORIAL CONTENT

 ubiquity press

CORRESPONDING AUTHOR:

Mark A. Parsons

University of Alabama in
Huntsville, USA
map046@uah.edu

TO CITE THIS ARTICLE:

Parsons, MA, Mayernik, M. 2023. 20 Years of Data Science: An Editorial. *Data Science Journal*, 22: 19, pp. 1–3. DOI: <https://doi.org/10.5334/dsj-2023-019>

Figure 1 Auto-generated image from the prompt 'twenty years of data science'.

In recognition of two dynamic decades of data science (however it may be conceived), we, the current editors-in-chief, seek to reopen the conversation on the core topics within data science – past, present, and future. This collection includes reflections and pragmatic considerations of past DSJ editors as well as more philosophical ideas from some forward-thinking scholars who engaged in a panel discussion on ‘20 Years of Data Science’ at International Data Week 2022.¹

The collection provides insight on the evolving definition of the field as well as some of the pragmatic considerations for fostering a formal discussion of a very dynamic discipline. Overall, the term ‘data science’ remains contested and ill-defined. The essays in this collection provide insight at a time when the topic has advanced enough to draw lessons for the next 20 years.

Matthew Mayernik (2023) provides an overarching context with a review he initially drafted before we conceived of this collection. He takes a long-term view that contrasts the evolution of data science and information science. He presents data science as an ‘interdiscipline’ and illustrates how that raises challenges. Interdisciplinarity inherently includes multiple perspectives, and interdisciplinarity has always been a focus for the journal. All the former editors make this point, and it is currently captured in the journal scope by requiring submissions to ‘generalise their significance.’²

Lindsay Poirier (2023) subtly challenges this view. She suggests that data scientists often (falsely) consider themselves somehow neutral or above the fray of disciplinary culture and action. She argues that while we consider the cultures of other disciplines, we need to confront our own culture with ‘thick description’ of culture and study of ‘data ethnography’. In short, Mayernik and Poirier both argue for more critical introspection of the field. Sarah Callaghan (2023) illustrates the consequences of myopia well with a story of ‘data scientists’ presenting unhelpful, sophomoric analysis of COVID-19 data early in the pandemic.

Lili Zhang (2023) understands this critique and proposes ideas to challenge any detachment or complacency in the field. She reviews past practice and urges the community to question basic concepts like infrastructure, technology, and education to take them to the next level of actual engaged implementation.

John Rumble (2023), the CODATA President when DSJ began, describes the central motivators for launching the journal. He notes how this communication venue is especially needed for discussions of scientific and research data: ‘The scientific data community requires more from data science than other communities’. This provocative statement reinforces the self-critical ideas from the other contributors. Many of the issues considered in DSJ (preservation, policy, ethics, reproducibility, quality, etc.) receive less emphasis within data science work focused on business and economic concerns or basic algorithmic advances.

Sarah Callaghan (2023) discusses how she has considered three audiences in her career as an editor in data science: ‘the computer and data scientists with the cool new algorithms, but not necessarily the good quality data or real world problems on which to test these solutions; the data-intensive domain researchers with the well-described data and real world problems, who would quite like to see these new techniques; and the data stewards and engineers, who build the infrastructure and agree the policies to allow data science work to be done.’ DSJ considers all of these audiences, and given its relationship with CODATA, focuses on the last. The journal is grounded in the pragmatics of data scientists doing their day job while finding ways to share both the abstract and practical lessons they learn with their colleagues. As Sarah notes, ‘It doesn’t really matter what domain the data come from, but it does matter how the data are treated. And perhaps, more importantly, how the people behind the data are treated.’

The nature of data science will continue to change, but the evolution of the field is driven by those involved. A casual review of DSJ’s publications over the years suggests that we have gradually added layers of abstraction and grapple now with more social concerns than technical concerns. The two most-cited papers in the journal over the long-term and most recently consider issues of data quality and data sovereignty (Cai and Zhu 2015; Carroll et al. 2020) – two very contextual issues.

1 <https://www.scidatacon.org/IDW-2022/sessions/392/>.

2 <https://datascience.codata.org/about>.

Ultimately it appears that the term ‘data science’ will remain contested and ill-defined, especially with the growth of generative AI and other data-intensive endeavors, but the focus of DSJ remains relevant. The provenance, ethics, and contextual understanding of data will always underpin good science and scholarship.


We encourage discussion of these and related issues in contributions to this special collection, to the journal at large, and within the overall professional conversation around data science.

COMPETING INTERESTS

Parsons and Mayernik are joint editors-in-chief of the *Data Science Journal*.

AUTHOR AFFILIATIONS

Mark A. Parsons  orcid.org/0000-0002-7723-0950
University of Alabama in Huntsville, USA

Matthew Mayernik  orcid.org/0000-0002-4122-0910
National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), Boulder, CO, USA

REFERENCES

- Cai, L and Zhu, Y.** 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 4(0): 2. DOI: <https://doi.org/10.5334/dsj-2015-002>
- Callaghan, S.** 2023. Two journals and a pandemic: Reflections on being a data science editor-in-chief. *Data Science Journal*, 22(1): 14. DOI: <https://doi.org/10.5334/dsj-2023-014>
- Carroll, S R, Garba, I, Figueroa-Rodríguez, O L, Holbrook, J, Lovett, R, Materechera, S,** et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1): 43. DOI: <https://doi.org/10.5334/dsj-2020-043>
- Hey, T, Tansley, S and Tolle, K.** (eds.) 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.
- Mayernik, M S.** 2023. Data science as an interdiscipline: Historical parallels from information science. *Data Science Journal*, 22: 16, 1–18. DOI: <https://doi.org/10.5334/dsj-2023-016>
- Poirier, L.** 2023. Attending to the cultures of data science work. *Data Science Journal*, 22: 6, 1–7. DOI: <https://doi.org/10.5334/dsj-2023-006>
- Rumble, J.** 2023. Thoughts on starting the CODATA *Data Science Journal*. *Data Science Journal*, 22(1): 13. DOI: <https://doi.org/10.5334/dsj-2023-013>
- Smith, F J.** 2023. The launch of the *Data Science Journal* in 2002. *Data Science Journal*, 22(1): 11. DOI: <https://doi.org/10.5334/dsj-2023-11>
- Zhang, L.** 2023. Looking back to the future: A glimpse at twenty years of data science. *Data Science Journal*, 22(1): 7. DOI: <https://doi.org/10.5334/dsj-2023-007>

TO CITE THIS ARTICLE:

Parsons, MA, Mayernik, M. 2023. 20 Years of Data Science: An Editorial. *Data Science Journal*, 22: 19, pp. 1–3. DOI: <https://doi.org/10.5334/dsj-2023-019>

Submitted: 14 June 2023

Accepted: 15 June 2023

Published: 30 June 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.