COLLECTION:
20 YEARS OF DATA SCIENCE

ESSAY

ubiquity press

Two Journals and a Pandemic: Reflections on Being a Data Science Editor-in-Chief

SARAH CALLAGHAN

ABSTRACT

This essay is a collection of reflections from my experiences in the past few years, first as editor-in-chief (EiC) of the Data Science Journal and then launching Patterns as its first EiC during the global COVID-19 pandemic. I discuss what I learned and how I worked with these journals, and I close with some hopes for the next 20 years of data science.

CORRESPONDING AUTHOR:

Sarah Callaghan

University of Oxford, GB

sorcha.ni@gmail.com

KEYWORDS:

Data science; academic publishing; open access; pandemic

TO CITE THIS ARTICLE:

Callaghan, S. 2023. Two Journals and a Pandemic: Reflections on Being a Data Science Editor-in-Chief. *Data Science Journal*, 22: 14, pp. 1–4. DOI: https://doi.org/10.5334/dsj-2023-014

When I was asked to take on the role of EiC for the *Data Science Journal* (*DSJ*) in 2015, I was very honoured indeed. I was not new to academic publishing: as part of my day job at the time, I had managed projects and led a series of projects on how to give researchers academic credit for publishing their research data. As part of that work, I had collaborated with colleagues from the Royal Meteorological Society and Wiley to launch the *Geoscience Data Journal* and had been an associate editor for that journal for several years.

Starting off as EiC of the *DSJ* was less a case of taking over an established system than a relaunch, although the journal name was well known within the field. The journal management system was changed at the point that I took over, effectively starting again from scratch. This meant that in addition to delivering the revised aims and scope of the journal and appointing associate editors, I needed to build a reviewer database. This took time, though I give lots of credit to Ubiquity Press for their excellent support and communication when we worked together.

I have always been a strong proponent of open access. I feel it is the just and fair way to publish research, and I am always so grateful to all the members of the team behind the scenes that facilitate that, from the associate editors and peer reviewers to the journal management staff. Many of these people are providing time and effort for free as part of their commitment to the nature of academic communication, and I am so appreciative of those efforts. The publications in the *DSJ* are labours of love and freely given expertise and effort. Yes, to make them gold open access, there needs to be some form of article publication charge, but the *DSJ* is very reasonable in that respect.

I learned so much from being EiC of the *DSJ*—primarily, that managing the peer review process is a lot of work, and it can take a lot of time. Because we were a small, community-run journal staffed by volunteers, we were not considered big enough to be indexed in the commercial journal indexes, so the authors who came to us to submit their articles knew us on a personal level. I learned a lot about the everyday business of article publication and even dealt with my first (and only) retraction as an EiC.

After four years of being EiC for the *DSJ*, I was approached by Cell Press (part of Elsevier), as they wanted to launch a new journal of data science that was so new it did not even have a name yet. This showed that the whole subject of data science was becoming more important to the wider world, and the fact that the large, commercial academic publishers were launching new journals in this topic was a good sign.

I took on that role in May 2019 and dived into what it meant to be a full-time EiC of what is now a high-profile, high-impact, gold open access data science journal called *Patterns*. For nearly the first year of that role, I travelled around the world, soliciting papers, appointing an advisory board, attending conferences, and planning the aims and scope for *Patterns*, given a rapidly changing and innovating space. We went from my appointment and nothing to the first issue of *Patterns* in 11 months. It was fast, exhilarating, and occasionally a little bit nerve-wracking.

And then the COVID-19 lockdown happened in the UK. I went from lots of travelling and meeting people to doing everything via Zoom and email. Conferences were cancelled, and the number of publications being submitted to journals exploded, particularly on the topic of COVID.

COVID brought with it a significant challenge to data science as a field. The sharing of pandemic data was wonderful to see, as was how so many researchers of all types immediately volunteered their time and effort to do what they could to help. But I did happen to notice a trend in all of this: some data scientists, not knowing anything about how diseases spread, would take the freely available data and extrapolate (sometimes in very simple, linear regression ways) to 'prove' things about the rate of cases. These results were often presented in very simplistic ways, boiled down into a single graph for easy sharing on social media, and I found them naive and a bit panic inducing. There were several times that I had family and friends personally reaching out to me to explain what the available data and these curves actually meant in ways that they could understand.

At *Patterns*, we also rapidly reached the point where we would get quickly written submissions on COVID using the authors' favourite machine learning algorithm, which unfortunately did not really tell us anything new about either the pandemic or the data science methods used. We were not unique in this, however.

Stepping back from the pandemic, and taking a longer view, in the past 20 years, data science has grown immensely, though it still does not appear as a subject area in Scopus or as a subject category in Web of Science. Instead, articles are more likely to be categorised as a subset of computer science, which works well for artificial intelligence (AI) and machine learning or other, less obviously useful, topics, like decision sciences. This makes it very difficult to categorise what data science is from an academic standpoint.

When I relaunched the *DSJ*, I had a narrower view of what I thought data science was than I do now. Back then, because of my job history and research preferences, I defined 'data science' for myself as 'the science of managing the data produced from academic research'. I was less interested in data science algorithms and the like, considering them to be more computer science than relating to data. Thankfully, I soon learned the importance of being broader in scope for the *DSJ*, mainly after being given guidance from the editorial board.

For *Patterns*, I went wider. I wanted to bring people together to share data science solutions to cross-domain problems. I narrowed these down to three main groups of people: (1) 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; (2) the data-intensive domain researchers with the well-described data and real-world problems, who would quite like to see these new techniques; and (3) the data stewards and engineers, who build the infrastructure and agree on the policies to allow data science work to be done. This latter group includes those researchers involved with AI ethics and responsible data science—this was a key theme that we publicised both through research articles and opinion pieces, as I feel very strongly about these issues.

Data science also has very fuzzy boundaries when it comes to the academic/industrial split. It is not unusual to find big tech companies with stalls at computer science conferences like CVPR and NeurIPs, but what I found interesting was the number of data science conferences that were squarely aimed at businesses and their CEOs and CFOs. 'Data science' in this context was referring to business data and how one could get more insights out of one's web or marketing data in order to understand one's customers and get more sales. Yes, this is a valid expression of data science, and development of these tools has an enormous impact in the business sphere. But this data science is not what I was wanting to publish as EiC, and it did crowd the field and discussions. (It also made it very difficult to hire editors with data science experience, as the salaries data scientists can command for working for big tech companies far exceeded what was standard in the publishing domain.)

Data science, for me now, is the science of data and research on the methods and tools used to learn from data, though I appreciate that this definition is definitely incomplete. It does not 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.

There is a tendency in data science to view the data as just numbers—that is, that they can be analysed, visualised, and put into all sorts of algorithms without any real-world consequences. This is completely false, and this idea has caused a great deal of harm, and potential harm, to marginalised communities. Yes, it is an interesting challenge to take a photo of a person and turn it into a series of character/avatar artworks. The problem is that big AI models like these are inherently biased, based on the biased historical data on which they were trained (usually scraped from the Internet, which has its own problems). It might seem harmless that the outputs of these picture generators create images of men in suits and dressed as astronauts, but when a female face is presented, the suggestion for her clothing is far more likely to be a bikini (Heikkilä 2022). More dangerous are the AI algorithms that can supposedly determine the sex or race of someone based only on their face; besides being completely inaccurate, those are actively harmful, as oppressive regimes or extremists could easily use them to target people who do not fit the algorithm's rigid example of the gender binary or racial characteristics.

On an even more pragmatic note, the cost of these large-scale machine learning models is considerable, and not just in terms of monetary amounts. The amount of carbon dioxide emitted by these models is substantial (Strubell et al. 2019; Freitag et al. 2022) and is increasing as the models' complexity increases.

So, where do we go from here? What will data science look like in the next 20 years? I have never been good at crystal ball gazing and have always been an optimist, so my judgements are more of what I hope will happen.

I hope that in the next 20 years AI will be a fully integrated tool in the scientific process, but it will be as a support for researchers and experts. I do not want to have AI deciding whether or not a growth is cancerous based on scan results, but I do want it to be able to quickly flag up a suspicious case so it can be reviewed by an experienced medical professional. I hope that the use of AI to generate art will be viewed as an interesting but ultimately unnecessary duplication of human artists. I hope AI will be able to guide people to the answers they are looking for in the vast corpus of human literature without generating plausible but ultimately false answers.

I hope that data scientists will continue to think about the impact their work has on the wider social and environmental domains and that they will take more of a 'first do no harm' ethos when designing their tools and experiments. I hope that data scientists will pay more attention to the communities they derive the data from and that the whole act of data collection becomes decolonised so that those who provide data will also share in the benefits of its collection.

I hope that all research will become more open, transparent, verifiable, rigorous and ethical and that we acknowledge the humanity of the researchers who create it. I hope for a fairer research world, where diversity of experience and opinion is valued and where we have removed the perverse incentives that promote individuals to the detriment of the research community and our collective knowledge.

I am sure that in the next 20 years we will see new developments in data science that very few could have predicted, and I sincerely hope that these will benefit all of humanity. With good intentions, sharing our research and our experience can help make the world a better place for us all.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Sarah Callaghan orcid.org/0000-0002-0517-1031
University of Oxford, GB

REFERENCES

Freitag, C, Berners-Lee, M, Widdicks, K, Knowles, B, Blair, GS and **Friday, A.** 2022. The real climate and transformative impact of ICT: a critique of estimates, trends, and regulations. *Patterns*, 3(8): 100576. DOI: https://doi.org/10.1016/j.patter.2022.100576

Heikkilä, M. 2022. *The viral AI avatar app Lensa undressed me—without my consent*. MIT Technology Review. Available at https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/ [Last accessed 5 June 2023].

Strubell, E, Ganesh, A and **McCallum, A.** 2019. *Energy and policy considerations for deep learning in NLP*. arXiv: 1906.02243. DOI: https://doi.org/10.18653/v1/P19-1355

TO CITE THIS ARTICLE:
Callaghan, S. 2023. Two Journals and a Pandemic: Reflections on Being a Data Science Editor-in-Chief. *Data Science Journal*, 22: 14, pp. 1–4. DOI: https://doi.org/10.5334/dsj-2023-014

Submitted: 06 January 2023
Accepted: 24 April 2023
Published: 09 June 2023

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