(Viewpoint)

Thorny Problems in Data (-Intensive) Science

Michael J. Scroggins**,** Irene V. Pasquetto, R. Stuart Geiger, Bernadette M. Boscoe, Peter T. Darch, Charlotte Cabasse-Mazel, Cheryl Thompson, Milena S. Golshan, Christine L. Borgman, Christine.Borgman@ucla.edu (Corresponding Author)[1]

Introduction

As science comes to depend ever more heavily on computational methods and complex data pipelines, many non-tenure track scientists find themselves precariously employed in positions grouped under the catch-all term "data science." Over the last decade, we have worked in a diverse array of scientific fields, specializations, and sectors, across the physical, life, and social sciences; professional fields like medicine, business, and engineering; mathematics, statistics, and computer and information science; the digital humanities; and data-intensive citizen science and peer production projects inside and out of the academy [3,7,8,15]. We have used ethnographic methods to observe and participate in scientific research, semi-structured interviews to understand the motivations of scientists, and document analysis to illustrate how science is assembled with data and code. Our research subjects range from principal investigators at the top of their fields to first-year graduate students trying to find their footing. Throughout, we have focused on

[1] Scroggins, Pasquetto, Boscoe, Golshan, Borgman are affiliated with the Center for Knowledge Infrastructures, Department of Information Studies, UCLA. Geiger and Cabasse-Mazel are with the Berkeley Institute for Data Science, UC-Berkeley. Darch and Thompson are with the School of Information Sciences, University of Illinois.

the multiple challenges faced by scientists who, through inclination or circumstance, work as data scientists.

The "thorny problems" we identify are brambly institutional challenges associated with data in data-intensive science. While many of these problems are specific to academe, some may be shared by data scientists outside the university. These problems are not readily curable, hence we conclude with guidance to stakeholders in data-intensive research.

The Janitors of Science

Within data-intensive science, it is a truth universally acknowledged that a dataset in need of analysis must first be cleaned. This dirty job falls to the data scientist. Though the computational machinery of science has allowed new forms of scientific inquiry – and new kinds of scientists – to be developed, the machinery is fickle and only accepts pristine datasets. Yet the process of cleaning datasets is often hidden or rendered invisible by disciplinary and organizational divisions [14]. While even the simplest dataset must be massaged prior to use, the problem multiplies when instrument calibration degrades or automated pipelines are changed without notice. One interviewee suffered an instrument malfunction during a remote sensing experiment. Unknowingly, one in an array of sensors failed out of calibration range during a field study, but the automated pipeline continued to generate data, which had to be painstakingly cleaned in the following weeks. In scientific fields that produce comparatively small amounts of data, cleaning is often

done manually in a spreadsheet, and problems spotted visually, but with bigger data comes bigger spills that require bigger cleanups.

Continuing Education in Science

Early champions of "big data" infamously predicted an "end of theory" [1], arguing that with enough data and computation, all research questions become simply an abstract problem of data processing. In contrast to this anti-disciplinary discourse, we see academic data scientists struggling to master the subject expertise necessary to make competent decisions about how to capture, process, reduce, analyze, visualize, and interpret research data. Domain scientists work closely with data scientists to model scientific problems, relying on common understanding to develop a team's data pipeline and computational infrastructure. As a result, the integrity of the research process can rest with data scientists. In such settings, data scientists must develop "interactional expertise" [5] by learning how to speak the jargon and conceptual vocabulary of a given discipline, and, more cogently, learning to ask the right questions of disciplinary scientists. Interactional expertise is not a skill that is readily taught in formal settings, particularly in traditional disciplinary degree programs. In response, data scientists gain interactional expertise in the fields in which they work by tactics such as making vocabulary lists of disciplinary jargon, quizzing colleagues in the hallway before a meeting, attending department seminars, taking classes, and reading literature of multiple domains.

The Overwhelmingness of Openness

Data-intensive science is increasingly tied to practices of, and policies for, "open science" [12,13]. Open science spans open access publications, open datasets, open analysis code, open source software tools, and much more. The concept spreads over a myriad of tools, platforms, frameworks, and practices that change often. Conflicts arise between tools that are built on open source ecosystems and controlled by a mix of public and private entities, ranging from file formats to high-performance computing infrastructures. Managing so many overlapping mechanisms can be overwhelming, especially when data scientists are hired to take the burden of maintaining infrastructures off the backs of domain researchers [11]. Today's scientific training may provide solid fundamentals for early career work, but rarely provides the skills necessary to keep pace with a fast-changing, complex ecosystem. Research groups face difficult tradeoffs between migrating to new tools and maintaining old packages, versions, and formats that work well enough – and are often embedded in legacy systems that must be maintained. These tradeoffs can place data scientists in uncomfortable mediating positions, similar to when they must translate between different disciplines.

Scarcity of Career Paths

Despite the rapidly growing need for data scientists in scientific research collaborations, these roles can lack specific job descriptions, and therefore a career path [7]. Data scientists are often part of a research personnel pool that moves from project to project within a university. Few of these jobs lead to faculty positions or other secure career tracks. Even in scientific enterprises that invest in computational infrastructure for data, we rarely find career advancement systems that include data-specific tracks. Those exceptions we have encountered occur outside university departments, such as large-scale, globally distributed research projects with significant division of labor. The scarcity of career paths for those with combined expertise in a scientific domain and information technology results in a profound loss of research capacity for universities. Whether individuals entered academic data science jobs as a career choice or as a byway en route to a faculty post, the lack of perceived upward mobility is resulting in departures for industry or other sectors.

Managing Infrastructures for the Long Term with Short-Term Funding

Scientific infrastructures accrete over long periods of time. Laboratories are constructed, equipment acquired, staff hired and trained, software and tools developed, journals and conferences launched, and new generations of scientists educated and graduated. Data scientists are increasingly responsible for maintaining the continuity of essential

knowledge infrastructures, yet projects may outlast individual grants, leaving data scientists to operate in conditions of uncertainty about the long-term future of the infrastructure they build [9]. This uncertainty poses complex challenges, both in terms of anticipating the needs of future users and of sustainability. In some scientific fields, the project life cycle unfolds on the scale of decades, in distinct stages such as initial conception, setting scientific goals, designing data management systems, constructing instruments and facilities, collecting data, processing data through pipelines, and releasing "science ready" data to the community. Builders of scientific infrastructure must make decisions in the present that will affect what data are collected and made available for decades, opening up some potential avenues of inquiry and closing down others [2]. Data-intensive science is plagued by the tyranny of small decisions; choices optimal in the short term may create a thorny nest of complications five or ten years later.

Untangling Thorny Problems

The data-intensive science problems we outlined above are intertwined with the organizational and funding of science within the university system [6]. They only exist, and can only be addressed, within these larger institutional and political constraints. The specific circumstances of data science activities vary widely between and within the physical, life, biomedical, and social sciences; engineering, humanities, and other fields. Scientific practices in all of these fields are in flux, requiring new tools and infrastructures to handle data at scale, and grappling with new requirements for open science. Some individuals choose data science jobs in universities, but often the job finds them. Learning data science may be an investment that leads to a productive career, but

all too often, time spent as the “data person” or “computer person” on the science team is labor not spent on dissertations, publications, or the scientific research that launches a tenure-track career.

These scientific environments have high personnel turnover rates, with individuals working in data science capacities through sequential post-doctoral fellow or grant-funded research scientist positions, or leaving for jobs in the corporate sector. Labor statistics are unlikely to capture the growth or turnover rate of these positions in science because the work is hidden behind so many different job titles. It is difficult to assess the damage to scientific progress when trusted data scientists move on to other institutions, as the losses may become apparent only months or years later. No matter how well code is documented, no paper trail can substitute for the rich domain expertise and tacit knowledge of those who conducted the science [4,10].

By bringing attention to these thorny problems, we aim to promote further discussion of the role of data science work both inside and outside of data-intensive science. Our list of problems is by no means exhaustive and our proposed remedies by no means complete. We offer our vignettes in the spirit of diagnosis and invite data scientists working in other fields, disciplines, and industries to contribute their own sets of thorny problems and solutions. We have written from the point of view of academic science as one permutation of data science, a term which escapes easy definition even as it advances. Much work remains.

ACKNOWLEDGEMENTS

This research was supported by grants to the University of California, Los Angeles from the Alfred P. Sloan Foundation (#201514001, C.L. Borgman, PI) and grants to the University of California, Berkeley from the Gordon and Betty Moore Foundation (Grant GBMF3834) and the Alfred P. Sloan Foundation (Grant 2013-10-27), as part of the Moore-Sloan Data Science Environments.

REFERENCES

1. Chris Anderson. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired 16*. Retrieved August 26, 2010 from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
2. Karen S. Baker, Ruth E. Duerr, and Mark A. Parsons. 2015. Scientific Knowledge Mobilization: Co-evolution of Data Products and Designated Communities. *International Journal of Digital Curation* 10, 2. https://doi.org/10.2218/ijdc.v10i2.346
3. Christine L. Borgman. 2015. *Big data, little data, no data: Scholarship in the networked world*. MIT Press, Cambridge, MA.
4. Geoffrey C. Bowker. 2005. *Memory Practices in the Sciences*. MIT Press, Cambridge, Mass.
5. Harry M. Collins and Robert Evans. 2007. *Rethinking Expertise*. University of Chicago Press, Chicago.
6. Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science* 41, 5: 667–690. https://doi.org/10.1177/0306312711413314
7. R. Stuart Geiger, Charlotte Mazel-Cabasse, Chihoko Y. Cullens, Laura Norén, Brittany Fiore-Gartland, Diya Das, and Henry Brady. 2018. Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey. *SocArXiv*. https://doi.org/10.17605/OSF.IO/XE823
8. Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska,

and Aleksandra Slavkovic. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology* 10, 4: e1003542. https://doi.org/10.1371/journal.pcbi.1003542

9. Steven J. Jackson, David Ribes, Ayse Buyuktur, and Geoffrey C. Bowker. 2011. Collaborative Rhythm: Temporal Dissonance and Alignment in Collaborative Scientific Work. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (CSCW '11), 245–254. https://doi.org/10.1145/1958824.1958861
10. Bruno Latour and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton, N.J.
11. Charlotte P. Lee, Paul Dourish, and Gloria Mark. 2006. The Human Infrastructure of Cyberinfrastructure. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (CSCW '06), 483–492. https://doi.org/10.1145/1180875.1180950
12. Nadine Levin, Sabina Leonelli, Dagmara Weckowska, David Castle, and John Dupré. 2016. How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice. *Bulletin of Science, Technology & Society* 36, 2: 128–141. https://doi.org/10.1177/0270467616668760
13. National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press, Washington D.C. Retrieved August 9, 2018 from https://doi.org/10.17226/25116
14. Jean-Christophe Plantin. 2018. Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science. *Science, Technology, & Human Values*: 0162243918781268. https://doi.org/10.1177/0162243918781268
15. Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE* 8, 7: e67332. https://doi.org/10.1371/journal.pone.0067332