Big Data Infrastructure at the Crossroads

Support Needs and Challenges for Universities

Dylan Ruediger

Thea P. Atwood
Neelam Bharti
Bryan Briones
Patrick Campbell
Paula Carey
Daniel Castillo
Karen Ciccone
Cameron Cook
Danielle Cooper
Claire Curry
Justin De La Cruz
Will Dean
E.M. Dragowsky
Tom Durkin
Darnell Epps
Seth Erickson
Jen Ferguson
Erin D. Foster
Moriana Garcia
Zenobie S. Garrett
Ann Glusker
Ben Gorham
Jen Green
Hannah Gunderman
Jacalyn Huband
Jennifer Huck
Susan Ivey
Carolyn Jackson
Kelsey Jordan
Kate Kryder
Stephanie Labou
Mark Laufersweiler
Tracie Lewis
James Macalino
Tobin Magle
David Minor
Lana Munip
Rosaline Odom
Reid Otsuji
Jennifer Patiño
Tyler Pearson
Carissa Phillips
Sara Pugachev
Brian Quigley
David Rachlin
Melanie Radik
Vicky Rampin
Fred Rowland
Laura Sare
Rebecca M. Seifried
Adam Shambaugh
Sarah Siddiqui
Kate Silfen
Iyanna Sims
Bryan Sinclair
Margaret Smith
Gretchen Sneff
Mandy Swygart-Hobaugh
Paria Tajallipour
Julia Unis
John Vickery
Cynthia Vitale
Jeremy Walker
Huajin Wang
John Watts
Chris Wiley
Katie Wissel
Nicholas Wolf
Cindy Xuying Xin
Jen-Chien Yu
Roger Zender
Lee Zickel

ITHAKA S+R

Ithaka S+R provides research and strategic guidance to help the academic and cultural communities serve the public good and navigate economic, demographic, and technological change. Ithaka S+R is part of ITHAKA, a not-for-profit with a mission to improve access to knowledge and education for people around the world. We believe education is key to the wellbeing of individuals and society, and we work to make it more effective and affordable.

ITHAKA is interested in disseminating this brief as widely as possible. Please contact us with any questions about using the report: research@ithaka.org.

Executive Summary

Ithaka S+R's Research Support Services program explores current trends and support needs in academic research. Our most recent project in this program, "Supporting Big Data Research," focused specifically on the rapidly emerging use of big data in research across disciplines and fields. As part of our study, we partnered with librarians from more than 20 colleges and universities, who then conducted over 200 interviews with faculty. These interviews provided insights into the research methodologies and support needs of researchers working across a wide range of disciplines.

This report provides a detailed account of how big data research is pursued in academic contexts, focusing on identifying typical methodologies, workflows, outputs, and challenges big data researchers face. Full details and actionable recommendations for stakeholders are offered in the body of the report, which offers guidance to universities, funders, and others interested in improving institutional capacities and fostering intellectual climates to better support big data research. Our key findings are grouped into the following areas:

- **Tension and Interplay between Disciplinary and Interdisciplinary perspectives**. Big data research is an interdisciplinary enterprise conducted by practitioners working in institutional settings that are still organized around disciplines. Divergent incentive structures, cultures, and unequal access to funding can affect disciplinary participation in big data research projects. Moreover, widespread use of methodologies from the computer and data sciences—most importantly a clear trend towards machine learning—has created tension among researchers and raised questions about the relative importance of disciplinary perspectives.
- **Managing Complex Data**. In an era of relative data abundance, researchers often avoid the expense of generating new data and instead opt to work with existing data whenever possible. The work of acquiring, cleaning, and organizing data is typically the most labor-intensive aspect of big data projects.
- **Structures for Collaboration**. Big data research is almost always a collective endeavor involving students, faculty, staff, and colleagues, clients, and collaborators from in and beyond higher education. Labs are the core units for research, and within them, students (both undergraduate and graduate) make significant contributions to the research process. Researchers often also favor local, lab-based computing resources over centralized campus storage and computing options, including cloud computing services.
- **Sharing Knowledge**. Although peer-reviewed articles remain the most highly incentivized form of scholarly communication, researchers are broadly committed to the open sharing of research outputs, including data and code. However, academic sharing practices reflect a spectrum that extends well beyond formal sharing in open repositories that meet FAIR standards of findability, accessibility, interoperability, and reusability, encompassing many types of informal sharing with colleagues.[1] Barriers to formal sharing include widespread

[1] Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, no. 1 (2016): 1–9, https://doi.org/10.1038/sdata.2016.18.

perceptions that much data is either derivative, low quality, or gathered from sources that are inappropriate for open sharing.

- **Ethical Challenges**. The ethical dimensions of big data research remain contested, and some researchers are uncertain about best practices for ethical research conduct. Although IRB guidance is valued, some researchers expressed concerns that IRB regulations are not well adapted to new or evolving research methods.
- **Support and Training**. Researchers tend to favor informal training methods, such as internet tutorials, over formal training in big data methods. While such methods work well for solving immediate problems, they are less well suited to acquiring foundational knowledge, leaving the potential for blind spots in academic research.

Introduction

Big data has moved from the margins of academic research to the center of a growing number of disciplines. What constitutes big data remains a matter of debate. While "bigness" is important, size alone is not big data's defining characteristic. Rather, the term serves as a useful shorthand for research projects that take advantage of advances in computing power and new technologies for processing, storing, and retrieving data and use sophisticated tools for aggregating, combining, and interpreting the massive data sets that those technologies have enabled.

First coined in the corporate world of the Silicon Valley in the mid-1990s, the idea of "big data" as a distinctive type of research began showing up in academic contexts around the turn of the century.[2] In the intervening years—and especially over the past decade—big data, and the closely related data science methodologies that are often used to interpret it, has rapidly become a significant, perhaps normative, research method in many academic fields. Several recent literature reviews have documented the steep acceleration in big data's footprint in academic publishing and research.[3] The most comprehensive of these reviews located 36,000 articles based on or about big data research published between 2012-2017 alone.[4]

[2] Francis X. Diebold, "On the Origin(s) and Development of the Term 'Big Data,'" SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, September 21, 2012), https://doi.org/10.2139/ssrn.2152421; Steve Lohr, "The Origins of 'Big Data': An Etymological Detective Story," *Bits Blog*, February 1, 2013, https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

[3] Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management* 35 (April 1, 2015): 139, https://doi.org/10.1016/j.ijinfomgt.2014.10.007; Vivek Singh et al., "Scientometric Mapping of Research on 'Big Data,'" *Scientometrics* 105 (September 9, 2015), https://doi.org/10.1007/s11192-015-1729-9; Gabriel Camilo Lima, "The Growth of AI and Machine Learning in Computer Science Publications," Medium, January 30, 2019, https://medium.com/@thegcamilo/the-growth-of-ai-and-machine-learning-in-computer-science-publications-603d75467c38; Barbara Blummer and Jeffrey M. Kenton, "Big Data and Libraries: Identifying Themes in the Literature," *Internet Reference Services Quarterly* 23, no. 1–2 (2018): 15–40, https://doi.org/10.1080/10875301.2018.1524337. Jonathan Grossman and Ami Pedahzur argue that big data has, as yet, made only a modest impact on political science research: "Political Science and Big Data," *Political Science Quarterly* 135, no. 2 (Summer 2020): 225–57.

[4] Ehsan Mohammadi and Amir Karami, "Exploring Research Trends in Big Data across Disciplines: A Text Mining Analysis," *Journal of Information Science*, June 5, 2020, https://doi.org/10.1177/0165551520932855.

Big data research is resource intensive, in both obvious and less immediately apparent ways. Keeping up with the technological, human, and financial demands of data-intensive research is a core strategic challenge facing research universities. The computing infrastructure required to store, share, and analyze large datasets is expensive and, in at least one instance documented in the interviews conducted as part of this project, has become so energy intensive that individual labs are now stressing the capacity of their university's electrical grid. As data-driven research has proliferated, universities have invested heavily in research data services hosted by a wide range of campus units, creating a human and service infrastructure to support big data research.[5] These are expensive services, but are by no means the only labor costs associated with big data research, which depends on formal and informal collaboration between students, postdocs, research staff, faculty, IT and information professionals, librarians, as well as legal offices, IRBs, and other university offices. Though sometimes overshadowed by technological costs, the aggregate labor expenditures that make big data research possible are considerable.

> As the ready availability of data grows, and as incentive structures push research towards big data, the demands on research offices, university libraries, high-performance computing centers, graduate programs, individual labs, and other university units seem poised to accelerate.

In some fields, including many of the most resource-intensive ones, sponsor funding generates revenue to offset these expenses, but as John Lombardi, former president of the Louisiana State University System, and others have noted, "research is essentially a money-losing proposition."[6] As the ready availability of data grows, and as incentive structures push research towards big data, the demands on research offices, university libraries, high-performance computing centers (HPCs), graduate programs, individual labs, and other university units seem poised to accelerate. Supporting this research is now central to the mission of research universities. Assessing the efficacy of existing infrastructures and identifying key needs of researchers is essential as universities develop plans to support big data research over the long term.

Ithaka S+R has a longstanding commitment to identifying trends in research agendas across academic fields and tracking changes in the structures that facilitate—or hinder—research. Recent Ithaka S+R studies have included a comprehensive survey of research data services provided by libraries, institutes, and high-performance computing (HPC) units in US colleges and universities, and a detailed exploration of the role of senior research officers at major research universities.[7] Likewise, our studies of data communities—informal or formal groups of

[5] Jane Radecki and Rebecca Springer, "Research Data Services in US Higher Education: A Web-Based Inventory," *Ithaka S+R*, November 18, 2020, https://doi.org/10.18665/sr.314397; "ARL/CARL Joint Task Force on Research Data Services: Final Report," July 16, 2021.

[6] John V. Lombardi, *How Universities Work* (Baltimore: Johns Hopkins University Press, 2013), 27; Christopher Newfield, *The Great Mistake: How We Wrecked Public Universities and How We Can Fix Them* (Baltimore: Johns Hopkins University Press, 2016), Ch 2; Karen A. Holbrook and Paul R. Sanberg, "Understanding the High Cost of Success in University Research," *Technology & Innovation* 15, no. 3 (December 18, 2013): 269–80, https://doi.org/10.3727/194982413X13790020922068.

[7] Radecki and Springer, "Research Data Services in US Higher Education: A Web-Based Inventory"; Jane Radecki and Roger C. Schonfeld, "Academic Research Budgets," *Ithaka S+R*, February 25, 2021, https://sr.ithaka.org/publications/academic-research-

scholars who voluntarily share data to advance research on topics of mutual concern—have explored decentralized data sharing networks that reach across institutional and disciplinary borders.[8]

> Big data research is almost always interdisciplinary in orientation, and the support needs of researchers in different disciplines overlap more often than might be expected.

"Supporting Big Data Research" builds on this foundation. Unlike previous projects, which have focused on research practices in single disciplines, "Supporting Big Data Research" takes a landscape view of big data research practices as a whole. In doing so, we are not dismissing the importance of disciplinary perspectives: indeed, as our findings make clear, disciplines shape the questions, methods, data types, and tools used in individual research projects. Moreover, disciplinary practitioners face different challenges, including unequal access to the resources required to sustain big data research and cultural norms that hinder data sharing. However, in practice big data research is almost always interdisciplinary in orientation, and the support needs of researchers in different disciplines overlap more often than might be expected. While universities should take disciplinary differences into account when developing support services, the greatest opportunities and efficiencies lie in identifying needs and opportunities that are broadly applicable.

Universities host core components of the big data infrastructure, but the larger ecosystem that sustains big data extends beyond academia. Funders, particularly in STEM fields, have made substantial investments in big data projects, while publishers have developed policies to facilitate data sharing to promote open science. "Supporting Big Data Research" offers ample evidence that these incentive structures have shifted research cultures and that an openness towards sharing data and code has taken hold in many disciplinary communities, despite practical barriers that cause actual sharing practices to lag behind the ethical embrace of the principle.

Given the importance of big data research and the closely related issue of data sharing to researchers and stakeholders in the research system, the stakes involved in creating infrastructures capable of sustaining big data research are high. The findings from this project suggest that universities are meeting many current needs, even as they highlight systematic challenges that will need to be addressed as data intensive research proliferates. This study explores research practices and emphasizes researcher's perspectives on their support needs. We anticipate that the findings and recommendations that follow will be useful to university research officers, libraries, computing centers, IT and information professionals, and faculty and staff who engage in big data research as well as publishers, funders, and others with stakes in research infrastructures.

---

budgets/; Oya Y. Rieger and Roger C. Schonfeld, "The Senior Research Officer: Experience, Role, Organizational Structure, Strategic Directions, and Challenges," *Ithaka S+R*, December 1, 2020, https://doi.org/10.18665/sr.314490.

[8] Danielle Cooper and Rebecca Springer, "Data Communities," *Ithaka S+R*, May 13, 2019, https://doi.org/10.18665/sr.311396.

Methods

The goal of "Supporting Big Data Research" was to provide a high-level assessment of the research practices and support needs of big data researchers working in a wide range of disciplines and institutions. To this end, the project was conducted in collaboration with local research teams, which ensured a high volume of data collection, something particularly important given the diversity of big data research projects. Participation in the project was open to any institution of higher education able to meet project specifications such as timeline and research capacity. Ultimately, the project cohort included 23 colleges and universities (three of them participating under the umbrella of the Atlanta University Consortium). The majority of participants represented major research universities (a full list of participating institutions can be found in Appendix I).

Each participating institution fielded a local research team composed primarily of librarians, many of them with professional responsibilities for data services, digital scholarship, or liaison duties to fields where big data research is common. A few teams included information or IT professionals employed by other university units. Ithaka S+R trained each team to conduct semi-structured interviews at a remote workshop. Following the workshop, each team conducted approximately 11 interviews (with a range of 8-16) with researchers on their campuses, using an interview guide designed by Ithaka S+R. (See Appendix II for the interview guide). Due to disruptions from COVID-19, two project teams have adopted extended timelines—their interviews are neither included in the numbers cited above nor in the analysis which follows.

The local project teams interviewed researchers from across the humanities, social and behavioral sciences, theoretical and applied STEM fields, as well as professional programs such as law and business. After completing their interviews, project teams coded interview transcripts using a grounded approach, and produced local reports summarizing their findings and recommendations to campus leaders. Many of these reports are now publicly available, though teams had the option to keep their local reports private if they believed keeping them private would further consensus building at their institution. These local reports provide important complements to this capstone report. Appendix I includes links to those reports that were made public.

Ithaka S+R developed a sample of 50 transcripts from the 213 interviews conducted by project teams. Tables 1 and 2 summarize our sample, which was developed to be representative of the overall population's distribution as measured by subject field, academic rank of the interviewee, and institution. The sampled transcripts were coded for analysis in NVivo using a grounded approach. In some cases, key word searches were conducted across the entire data set to explore specific topics in greater depth. This study makes no claims to be statistically representative of big data research as a whole. Nevertheless, the disciplinary distribution of our sample is broadly in sync with trends in disciplinary distribution of big data research revealed in several recent literature reviews. The most comprehensive of these, which surveyed 36,000 big data publications published between 2012 and 2017 found that approximately 50 percent of big data publications came from computer science, which is underrepresented among the interviews

conducted for this project.[9] However, recent literature reviews suggest that growth in big data research is concentrated in other academic disciplines. Ehsari Mohammadi and Amir Karami, in particular, have stressed the "growing influence [of big data] in academic disciplines outside of computer science," with particularly rapid growth in urban informatics, business, education, medical and health fields, and the computational social sciences, each of which are represented in interviews conducted for the project.[10]

Table 1: Academic Rank of Interviewees Included in the Sample

Rank	Percentage
Assistant Professor	24%
Associate Professor	24%
Professor	36%
Non-Tenure-Track	2%
Postdoc/Researchers/Research Staff/Other	14%
Total	**100%**

[9] Singh et al., "Scientometric Mapping of Research on 'Big Data'"; Mohammadi and Karami, "Exploring Research Trends in Big Data across Disciplines."

[10] Mohammadi and Karami, "Exploring Research Trends in Big Data across Disciplines," 13.

Table 2: Departmental Affiliation of Interviewees Included in the Sample

Department Affiliation	Percentage
Biology and Biochemistry	10%
Business/Finance/Marketing	4%
Chemistry	2%
Computer Science	12%
Demography	2%
Economics	2%
Engineering	16%
Environmental Science	4%
Geography and Geoscience	6%
Health Sciences/Medicine	6%
Humanities/Fine Arts	6%
Information Sciences/Education	4%
Law	2%
Mathematics	2%
Physics and Astronomy	6%
Planning and Policy Studies	2%
Political Science	4%
Psychology/Neuroscience/Behavioral Sciences	4%
Public Health	4%
Sociology/Anthropology	2%
Total	**100%**

To protect their privacy, the identities of the interviewees were not shared with Ithaka S+R, and they remain anonymous in this report, but we thank them for their participation. Roger Schonfeld and Cynthia Hudson Vitale provided essential feedback on draft versions of this report, and we are grateful for their insights and comments. Above all, we wish to thank the 21

project teams that took part in this research, without whom this report would not have been possible.

Defining Big Data

Size is an essential part of what defines big data, but there is, as yet, no consensus definition of that term. Even size can be measured in several ways, including the number of distinct observations, number of variables, and total amount of data accumulated over the course of a project (often measured in gigabytes, but in some fields in terabytes or even petabytes). Doug Laney's influential 2001 definition of big data focuses on the convergence of the "Three Vs"—volume, velocity, and variety—highlighting the speed at which new data is generated in venues like social media, the resulting challenge of collecting and processing data that is constantly in motion, and the diversity of different, often unstructured data types that are often combined in big data research. Reflecting the priorities of the business sector from which it sprang, Laney's characterization of big data is oriented towards real-time decision making in data-rich contexts rather than the more deliberate pace of academic research.[11]

> Size is an essential part of what defines big data, but there is, as yet, no consensus definition of that term.

Other definitions (of which there are many) focus relatively less attention on the properties of data and more on the infrastructures, methodological frameworks, and tools necessary to interpret huge datasets. Kate Crawford and danah boyd, for example, have suggested that "big data is less about the size of the data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets." By these standards, big data sits at the intersection of massive data sets, human labor, and the technical apparatuses required to process data.[12] Francis Diebold, one of the term's earliest academic adopters, has recently proposed that big data is creating a new interdisciplinary perspective that is leading the production of knowledge into "wildly new places, unimaginable only a short time ago." In its fullest sense, big data is not just size combined with tools, it is also an emerging epistemological stance spurred by what boyd and Crawford have described as a "computational turn in thought and research."[13]

The interviews conducted as part of "Supporting Big Data Research" turned up ample evidence that these and other definitions of big data circulate widely in research communities. In the absence of a consensus definition, this project was designed to cast a wide net by utilizing an inclusive, minimalist definition of big data to capture as wide a range of researchers and

[11] Doug Laney, "Deja VVVu: Gartner's Original 'Volume-Velocity-Variety' Definition of Big Data," August 25, 2021, https://community.aiim.org/blogs/doug-laney/2012/08/25/deja-vvvu-gartners-original-volume-velocity-variety-definition-of-big-data.

[12] danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society,* 15, no. 5 (2012): 664,666, https://doi.org/10.1080/1369118X.2012.678878.

[13] Diebold, "On the Origin(s) and Development of the Term 'Big Data,'" 5; boyd and Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," 666; David O'Sullivan, "Big Data … Why (Oh Why?) This Computational Social Science?," in *Thinking Big Data in Geography*, ed. Jim Thatcher, Josef Eckert, and Andrew Shears, New Regimes, New Research (University of Nebraska Press, 2018), 21–38, https://doi.org/10.2307/j.ctt21h4z6m.7.

research practices as possible. We encouraged project teams to consider big data research as involving "large, diverse datasets that are often collected in real time," and interpreted by researchers in a wide range of disciplines using specialized methods such as "machine learning, data/text mining, modeling, and other data science techniques."

Disciplinarity and Interdisciplinarity in Big Data Research

Big data research is a highly interdisciplinary enterprise conducted by practitioners trained in disciplinary settings. While research questions are diverse, the widespread adoption of methodologies drawn from computer and data sciences can create tension between researchers and raise questions about the relative value and status of disciplinary perspectives.

Converging Methods

Unsurprisingly, most researchers interviewed for this project had graduate training in a traditional academic discipline and work in settings still largely organized by discipline. The research questions they asked were often rooted in disciplinary perspectives and concerns. For this reason, the range of research questions underlying big data research is dizzyingly diverse. The tools researchers use over the lifecycle of a project can also be rooted in disciplinary traditions—for example, some social scientists still prefer STATA for data analysis and visualization. Recognizing these differences is important, as it has real impacts on universities' capacities to support big data research. However, as big data techniques spread across disciplines, centralizing tendencies towards certain tools and methodologies are emerging. Many of the most common tools used for statistical analysis—for example, Python, R, STATA—are often viewed as different means to the same end, to be used interchangeably or based solely on the preferences and knowledge of individual researchers. As one researcher noted, "it really makes little difference which language or library we choose, we will get similar results usually." Even so, our interviews suggest a consolidation across many fields towards Python and R, tools that a biologist described as two nodes (along with Linux) in the "holy trilogy" of data science. More consequentially, a few key methodologies—web scraping, machine learning, natural language processing, statistical analysis—are at the heart of much big data research regardless of academic field. Researchers expected even further methodological consolidation towards AI as a standard approach. Speaking for many of their colleagues, one interviewee observed that "the movement towards deep learning methodologies is inevitable in almost every single field right now." The consolidation of research methodologies towards those drawn from the computer and data sciences is creating tension between disciplinary and interdisciplinary modes of inquiry and ways of framing research questions and agendas. Many of these tensions revolve around questions about whether the growing reliance on machine learning is shifting research methodologies away from hypothesis-driven inquiry towards the search for predictive correlations.

Disciplinary Tensions

In practice, big data methods often complement established disciplinary modes of inquiry. However, critics have expressed fears that data science and AI could become generic models for inquiry that could override long-standing discipline-specific ways of knowing. Several researchers interviewed for this project echoed these concerns. For example, an assistant professor in a business school, speaking about the challenges facing the field, remarked that "a lot of researchers learned one method of analysis, and rather than trying to genuinely understand their data and where it came from as well as what they're trying to do with it, most of the time researchers don't know what to do with it. They take the same hammer and just keep applying it to all different data sets." Others described the reception of their research as contested by traditional disciplinary practitioners, or worried that their field was beginning to drift too far from their roots towards AI methods, which one physicist described as the "flavor of the month." The reasons behind these worries were seldom fully explained, but the general stakes are clear: methodologies shape research questions and frame conclusions. Data science is a powerful lens, but one that may compete with disciplinary epistemologies—by pushing qualitatively oriented fields towards quantitative methods, for example. The outputs and internal logic of algorithmic methods can be opaque even to expert users. Relatively few researchers interviewed for this project voiced strong worries about these trends: this is unsurprising given that all project interviewees were actively conducting big data research, and thus presumably found this mode of inquiry interesting and fruitful. Even so, they demonstrate concerns—even among practitioners—that relying on a small set of approaches from the computer and data sciences risks simplifying the epistemological horizons of many fields and encourages already pronounced tendencies to elevate quantitative evidence over qualitative evidence.

> "A lot of researchers learned one method of analysis, and rather than trying to genuinely understand their data and where it came from as well as what they're trying to do with it, most of the time researchers don't know what to do with it. They take the same hammer and just keep applying it to all different data sets."

One particularly important aspect of these controversies is the idea that big data could come to function as a supra-disciplinary perspective capable of overriding not only existing disciplinary orientations but of superseding the hypothesis-driven, theory-based mode of inquiry that has served as the foundation of modern scientific methods. The idea that big data might undermine the centrality of theory to the scientific method was famously evoked by *Wired* magazine's Chris Anderson in an important early reflection on the potential effects of big data on academic research.[14] Anderson's speculations about the "end of theory" in favor of a science built on the

[14] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, April 12, 2021, https://www.wired.com/2008/06/pb-theory/; Sauro Succi and Peter V. Coveney, "Big Data: The End of the Scientific Method?," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 377, no. 2142 (April 8, 2019), https://doi.org/10.1098/rsta.2018.0145; Mario Carpo, "Big Data and the End of History," *Perspecta* 48 (2015): 46–59; Rob Kitchin,

discovery of correlations hasn't yet come to fruition (at least in academic big data research), and he has since backed away from the most speculative parts of his essay. However, the idea that, given big enough datasets, correlations might supplant causation circulates in popular discourse about the implications of big data research and within research communities.[15] Our interviews suggest that data, rather than theory, is sometimes a starting point for research. A professor of public health, for example, identified big data with a reversal of standard methodologies: rather than creating bespoke data to test a hypothesis, big data approaches started by assessing what existing data sets might be able to reveal. A biologist put it even more succinctly, describing how computing has become the "starting point" for their research: "we try to let the data tell us where we should go."

Even researchers who recognized the potential value of unexpected correlations were likely to be protective of traditional modes of inquiry. One researcher noted, for example, that "data-intensive science has become so important that some are suggesting it should be considered another [scientific] paradigm." Data science techniques, the researcher continued, were excellent at finding correlations in the data and could sometimes tie unexpected things together. Yet the value of these correlations was primarily measured by their ability to inform new hypothesis-driven research. A professor of marketing expressed concern about research focused on prediction rather than explanation and association rather than causation. Another dismissed machine learning, in particular, as too often involving little more than "dump[ing] data into some kind of neural network" in search of correlations. Their lab, the researcher continued, will remain rooted in hypothesis-driven experimentation. One researcher expressed outright contempt over data scientists grappling with problems germane to other disciplines, pointing out that they had seen a huge number of papers on public health issues related to the COVID-19 pandemic written by researchers trained in computer science or data science, but who had "no background in either biology or public health or epidemiology." The researcher worried that government officials were making public health decisions based on predictions created by unqualified researchers. "It's great," they continued, "for fields to have people come in from other disciplines, that's how fields grow. I think that's fantastic, but I think you need a certain level of humility for that. Most of the people [writing those papers], I have not seen that level of humility."

It's easy, and sometimes probably fair, to read these concerns as academic turf battles, but tensions over disciplinary research perspectives can texture the collaborative relationships on which big data research depends. Several of the researchers interviewed for the project described being essentially dependent on the expertise of data scientists, to whom they have outsourced key components of their research agenda. A physical therapist, who emphasized that they were trained as a clinician and had no training in data science, described their collaborative relationships by saying that they took raw data to colleagues in the data sciences "so that we can

"Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* 1, no. 1 (April 2014): 1–12, https://doi.org/10.1177%2F2053951714528481.

[15] E.g., Viktor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2013).

have them tell us what's good about our data, what's not." Asked about the biggest challenges that working with big data presented, the researcher responded by saying that "I don't know how to do it."

> Faculty from the humanities and qualitative social sciences reported feeling particularly burdened by the necessarily cross-disciplinary features of big data research.

This was a thankfully extreme example of inequitable collaboration; however, the intellectual challenges of working with big data and the difficulty of balancing disciplinary and data science perspectives towards research are something that many researchers struggled with.[16] The resulting tensions may be especially acute in fields oriented towards qualitative research, whose disciplinary ways of knowing risk further marginalization in already hostile academic climates as the normativity of quantitatively-focused big data research grows. Many of those fields have access to fewer financial resources to support data-intensive research. Faculty from the humanities and qualitative social sciences reported feeling particularly burdened by the necessarily cross-disciplinary features of big data research. Because they continued to be evaluated under disciplinary standards that don't make room for data science or computer science expertise, it was especially difficult for them to justify investing the time to develop those skills. Moreover, big data research projects in these fields often developed without robust grant funding and/or staff support, making barriers to entry especially high, and the temptation to outsource the analytical parts of their work potentially acute. Even in comparatively well-resourced fields, the complexity of big data methods and tools can create deep uncertainty across the research process. The goal of research, an engineer noted, is to take raw data and transform it through "different levels of abstraction or processing" into actionable knowledge. Ideally, the research process should flow from "data to insight, to knowledge, to decision." The question, though, is "how do we get there?"

Managing Complex Data

Data management is essential to successful big data research. Though often largely invisible to outsiders, the work of acquiring, cleaning, processing, and organizing data is the most labor-intensive aspect of most big data projects.

Creating and Accessing Data

Big data projects are hungry for inputs. Relatively few researchers depend entirely on primary data that they have gathered firsthand. Most interviewed for this project reported relying either entirely or predominantly on existing, secondary datasets created by others, or used a combination of secondary and primary data. Cost factors heavily into researchers' decisions about which type of data to use, and generally pushes them towards existing data, especially

[16] For a detailed analysis of disciplinary tensions with data science perspectives, see Kitchin, "Big Data, New Epistemologies and Paradigm Shifts."

data that is freely available. Social science and humanities faculty were particularly likely to describe cost as major factors in their choice of data. For example, a political scientist described their research as shaped by their awareness of costs: “things are either free or you can’t get them, or that’s how it works in my head.” But even researchers in STEM fields take cost into account while designing research projects. “There are lots of data sets that are available that you can purchase that are quite expensive and out of my realm,” said a professor of public health. And regardless of field, generating primary data was a significant expense that researchers avoided if possible.

When available, researchers found data repositories to be “really, really important” resources, valued because they centralized data and simplified discovery. Repository availability varied by field and subject area, so many researchers were forced to locate useful datasets from widely scattered locations, such as GitHub, public-access databases, and the websites of individual laboratories. Many researchers expressed hopes that more repositories would become available, or that their libraries would purchase more subscription databases, because the decentralized discovery process was time-consuming and challenging. Absent central resources, many researchers relied on the open web. Asked how they found the data required for their work, one computer scientist answered, “Google search. I mean, obviously, sometimes we know where we can get data from, but basically very often it’s just search. . . we just go to the internet and try to find something.” Another quipped that they “like to joke with my friends who do natural science work” by saying that “my idea of a good day in the field is a new website.” Our interviews suggest that this is a version of field work that many natural scientists are familiar with.

Some researchers generate their own data, a practice that allows them to tailor it to their needs, but which also introduces new complexities. “Collecting data from the real world,” said a computer scientist, “is super hard, because you're working in the real world, which means that you have all of the uncontrolled, and context dependent aspects of what's going on in the real world.” A biologist, whose work involved undersea exploration, described in excruciating detail the expensive and elaborate procedures necessary to perform their work, which involved submarines, boats, labs, and support teams racing against the clock and the budget to gather data that must then be transformed and sequenced before its value can be assessed. This was, in some respects, a rather baroque example, but other researchers—in geosciences, physics, for example—described similar expensive efforts to generate research data.

Because generating new data is expensive and finding exactly the data one needs to answer a specific research question can be impossible, some researchers have experimented with new methods for data gathering. A few described having designed research agendas around available datasets, in effect working backwards from the data rather than beginning with a hypothesis and then finding data to investigate it. More frequently, researchers described having adjusted their research in response to limitations in the available data. A few expressed an interest in using simulated data, a process that allows for greater control over variables while reducing error and redundancy. Proponents of simulated data suggested that “curating it is usually relatively straightforward, just because you have control over the process from beginning to end.” A geoscientist added that simulated data was also cheaper to generate than other types of

experimental data, making it more cost-effective to create and process because it "generates less noise as it is easier to control variables."

One of the defining features of the digital era is that data—once a scarce resource—has become relatively abundant and readily available. While some researchers still reported that the data they need does not exist, they were more likely to feel overwhelmed by the quantity of data available to them, especially because the quality of data hasn't necessarily improved in lockstep with its quantity. "Sometimes it's too much data," said one researcher, echoing a sentiment familiar to many of their colleagues. Our interviews suggest that, in the absence of perfect data, many researchers will choose to use proximate data, even if it requires alterations to their research agenda.

Cleaning, Wrangling, and Managing Data

Once researchers acquire data, it needs to be cleaned and processed for analysis. Often assigned to junior members of the research term, cleaning and wrangling data are usually the most labor-intensive aspects of big data projects. A professor of public health reckoned that 80 percent of the total time expended on a project was devoted to cleaning data, while a physicist described analysis as the most fun and rewarding part of their work, "but that's maybe 10 percent of the job ultimately, right?" Though frequently considered low-level work, these tasks are typically confined to the shadows while analysis and publication take center stage. Yet the accuracy of major research projects depends on how well data is cleaned and how seamlessly different datasets are integrated. Many researchers interviewed as part of this project insisted on the intellectual integrity of the work, noting that it often requires fine judgment and content knowledge. Cleaning data, said one researcher, is "actually a lot of work, and it's quite complicated."

Cleaning and wrangling are perhaps especially important when working with secondary data, where errors can be particularly difficult to catch. An agricultural economist described the risks: "if you're gathering the data yourself, you know all the warts, and if you're just using somebody else's data, the warts are not as obvious." This danger, inherent to secondary data, is aggravated by the trend towards ever bigger data sets, whose bulk makes careful examination difficult. Big data sets, the researcher had found, could easily hide "errors in data or incorrect coding or things that showed up that were coded as missing. . . and you don't realize that's been coded as missing, and instead you're all of a sudden treating that as a real number." Poor documentation, spotty metadata, and a lack of information about how data has been processed are major sources of opacity in the research process. "Provenance," remarked a computer scientist, is a frequent challenge with using existing data, requiring "a lot of effort in first understanding what we are dealing with, processing, and doing some kind of data exploration." The possibilities for error are innumerable and very difficult to detect, making careful and laborious cleaning mandatory.

The scale of big data also exacerbates problems of discerning signals from noise. This was particularly challenging for researchers who generated their own data, much of which, as one researcher described it, was "junk because it's not the data that we need, but they get generated

anyway." Researchers in fields ranging from pharmacology to biology, computer science, and engineering described the difficulty of finding useful information in the sea of data their work generated. The growing use of AI may exacerbate this trend. As one researcher observed, "deep learning methods generate a lot of data, but that data is not intuitive to interpret." Another, whose institutional home is in engineering and applied sciences, noted that as AI algorithms become more automated and easier to use, they can produce more outputs, but those outputs become increasingly difficult to evaluate and still require arduous cleaning and interpretation before they can produce useful knowledge.

While some research projects draw entirely on the basis of a single large data source, for example data gathered via Twitter's API, most big data researchers use "heterogeneous datasets" that must be painstakingly stitched together. One researcher described their workflow as concentrated on "aligning different data sets," finding ways to connect data sets to allow for more complex comparisons and to create new, more nuanced variables and metrics. Much big data research depends on the resulting synergies, correlations, and insights made possible by cleaning, wrangling, and reformatting data in ways that make it possible to "reconcile[e] multiple disparate sources into a computable format" before the more celebrated work of analysis can begin.

Data Management

The data that researchers generate, collect, and process requires constant reorganization. Data management was another major challenge faced by big data researchers, though it seemed most daunting to those who generated their own data. "Organizing the data that we collect," a mathematician remarked, is the "primary challenge." The size of datasets was an important contributing factor to the challenges of data management. One researcher aptly described facing a "data avalanche." A computer scientist spoke about the challenge of establishing version controls for 50 terabytes of HD video. It "doesn't seem feasible," they said, to actually keep track of such a large amount of data: instead, they resorted to creating multiple backups and hoping for the best. Likewise, a biologist struggled to keep track of terabytes worth of molecular simulations created during a project that spanned several years. "It's difficult to organize, in a rational way, backups, long-term storage, and [ensure that you] have a certain ability to retrieve data" for later use.

Researchers frequently expressed concerns about their ability to effectively manage version control, naming conventions, and metadata, all tasks made both more urgent and more difficult by the collaborative nature of their work. Version control could be particularly challenging when data is stored centrally but used locally. These issues were especially challenging for researchers working in emerging fields, niche topics, or with novel data types (such as 3D data), who often had little guidance about standards and procedures for data management and metadata. Left to figure it out on their own, they struggled to label data artifacts and archive them in ways that allowed them to retrieve it for their own use or for sharing with other researchers.

Researchers understood the importance of developing solid data management plans (DMPs). An engineer described DMPs as things they used to you dashed off quickly, but now felt ready to

"actually spend a day or two" on. Libraries and other university units have invested heavily in programming and staff to support data management plans: interviews for Supporting Big Data suggest uneven awareness and modest use of those resources. Nevertheless, anxieties about the repercussions of faulty data management, and the challenges of managing data across long-term, team-based, projects were at the front of many researcher's minds.

Managing Complex Workflows

Big data research operates at scales that ensure it is almost always a collective endeavor involving close collaboration between students, faculty, and staff. Labs are often the core unit, but collaborative networks extend across universities and include institutions outside higher education. Personnel and project management skills are important aspects of big data research, as is storing data in ways that facilitate cross-institutional cooperation.

Interdisciplinary and Inter-institutional Collaboration

Collaboration is a ubiquitous part of big data research, even in humanities fields where scholarship is normally a solitary pursuit. Collaborations extend across disciplines, institutions, and continents, and take many forms, ranging from informal one-off conversations to long-term, employee-employer relationships. However, in many fields the lab—organized around a tenure-track faculty and a cohort of undergraduate and graduate students—is the core unit of big data research, sometimes with additional labor provided by postdocs, and, less frequently, in-house research or administrative staff. This human labor is too often relegated to the background of studies on big data and AI, but as the project interviews make clear, human networks are as essential to big data research as computing networks.[17] Indeed, project management emerged as among the major challenges confronting big data researchers, especially when their collaborative networks cross disciplinary, institutional, or national borders, or extend beyond academic circles.

For many researchers, collaborating with colleagues from different disciplinary backgrounds is a routine part of their work. These collaborations usually involve faculty members from traditional liberal arts disciplines who shape the research agenda and experts in computer science, statistics, or data sciences who are tasked with conducting the technical analysis of data. Said one professor with an appointment in public health, but whose PhD is in data sciences, "I always find that most data science projects are team projects," requiring coordination between a content expert and data scientists who provide methodological expertise. Likewise, an engineer frequently collaborated with programmers and statisticians: "we are," they noted, "sort of consumers of that kind of expertise." For their part, data and computer scientists sometimes described similar divisions of labor. "I work with doctors," said a research staff person with a

[17] Sarah T. Roberts, "Your AI Is Human," in *Your Computer Is on Fire*, ed. Thomas S. Mullaney et al. (Cambridge, Mass: MIT Press, 2021), 51–71; Kate Crawford, *Atlas of AI* (New Haven: Yale University Press, 2021); Paola Tubaro, Antonio A Casilli, and Marion Coville, "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence," *Big Data & Society* 7, no. 1 (January 1, 2020), https://doi.org/10.1177/2053951720919776.

computing background, "but I'm the only person that wants to look at the code and analyze the data. So, they're all, I guess, happy for me to do that on my own." A computer scientist concurred, remarking that they often work with "people who don't have experience with data processing/analysis," who come to them for help. "There are a lot of people who don't have experience," the professor said, "I'm the one with experience."

Collaborative networks frequently extend across institutions, and can be quite large: a plant biologist, for example, mentioned that over 100 researchers were involved in their project. Cross-institutional teams tend to revolve around faculty with similar research interests but may also take advantage of archival collections held by different institutions, specialized instruments available only at a few institutions, or serve to build redundancy and reproducibility into a project. Said one researcher, "it's good to have two or three other groups doing the same thing, even better if it is different software, same data, different software." Research projects frequently cross national borders, which can complicate project management and data sharing, especially if the research involves human subjects and/or politically sensitive topics.

Most researchers we interviewed worked primarily with other academics. However, collaborations with researchers or funders from the private sector or government agencies were not uncommon. Projects with industry often involved an exchange of funding and proprietary data for early or privileged access to research findings. Partnerships with government agencies—especially those relating to national defense—can include restrictions on publication of data and findings. One particularly challenging issue surrounding collaborating with government entities is increased geopolitical tension between the US and China.[18] Two researchers in this study, both working with NASA, noted that they were required to report whether Chinese nationals or those with ties to China worked in their lab, and there were limitations on what data they were allowed to handle.

Team Structures, Roles, Workflows

In most big data projects, students play important roles as inexpensive, skilled labor, taking on much of the grunt work of data gathering, cleaning, and coding, with faculty serving as supervisors and analyzing, or at least writing up, the data. As one computer scientist put it, "I have a team of, I don't know, usually 10 to 20 students who work with me. . . I'm getting a lot of help from them. Actually, I'm not doing any work on my own." Graduate students, in particular, are essential workers in many big data projects, with significant roles in data gathering and data analysis. In some cases, research outputs are basically dependent on the quality of the work graduate students do. An economist, for example, noted that students write most of their code, very little of which is ever checked for accuracy by the principal investigator (PI). "You have to have some level of trust with your graduate students," the economist continued, since they actually do much of the wrangling, coding, and analysis on which the PI's reputation will ultimately depend. Colleagues in other fields concurred, including a biologist who declared that

[18] For an overview, see, Roger C. Schonfeld, "Global Science and the China Split," *Ithaka S+R*, August 27, 2021, https://doi.org/10.18665/sr.314295.

"you know, all the PIs," they said, "we never look at the data; we never touch it." Faculty, especially in lab fields, also occasionally reported depending on their graduate students to keep up with secondary literature in the field. Undergraduates were part of many research teams, often working on data entry or coding. Indeed, a few faculty preferred undergraduate computer science students to grad students as coders, finding them more enthusiastic and skilled, and less hindered by content knowledge. Given their centrality to the research process, the educational and training needs of students (discussed in more detail below) are important considerations for supporting big data research.

Within labs, student labor is frequently supplemented by postdocs and other research support staff, sometimes employed directly by the PI, and other times by departments or other university units. Staff support is usually highly specialized and may involve handling security and ethics protocols, working with specialized data formats or serving as software engineers and/or hardware technicians. In some instances, researchers mentioned having access to statistical support staff or IT/network administrative staff. This was comparatively rare, but where available statistical support and IT staff were highly valued assets. Often, these individuals were employed by departments or research centers and provided support for multiple projects rather than being in the direct employ of a specific PI.[19]

> "I love libraries, but also my brain still imagines like '80s libraries or the way libraries were when I was growing up. So, it honestly hadn't really occurred to me to think of the library as a resource for big data tools and a resource in this space."

In recent years, university libraries (and other campus units) have made substantial investments in a range of data services designed to support big data research.[20] Project interviews suggest mixed reviews of these offerings. On a few campuses, researchers reported seeing evidence of a "concerted effort" between IT and the libraries to build infrastructures of big data. However, many faculty members remain only dimly aware of the resources libraries offer. "I am very pro library," said one researcher. "I love libraries, but also my brain still imagines like '80s libraries or the way libraries were when I was growing up. So, it honestly hadn't really occurred to me to think of the library as a resource for big data tools and a resource in this space." The perception that the library was little more than a place to "get access to research, to papers and to books," was common among faculty. However, on occasion, researchers described having established collaborative relationships with librarians. An agricultural economist recalled a "really great guy at the library that helped me pull together a really complicated dataset," while a computer

[19] Driven in large part by security concerns and efforts to consolidate costs, university trends seem to point in the direction of increasingly centralized IT, something that can be difficult to accomplish in what remain relatively decentralized institutions. This issue did not often come up explicitly in our project interviews, but on the whole researchers - who often made use of centralized resources (including around basic security) exhibited an overall preference for decentralized resources. This may ultimately point towards the need for what Jim Davis has described as a "layered approach" to IT. See Jim Davis, "Beyond the False Dichotomy of Centralized and Decentralized IT Deployment," https://www.educause.edu/research-and-publications/books/tower-and-cloud/beyond-false-dichotomy-centralized-and-decentralized-it-deployment.

[20] Radecki and Springer, "Research Data Services in US Higher Education: A Web-Based Inventory."

scientist had consulted with librarians about tools for running data analytics, and the PI of a digital humanities project "frequently" consulted with their libraries' digitization expert about issues involving formatting and archiving digitized cultural heritage objects.

Storing and Using Data

Active storage issues were among the most common choke points described by big data researchers, who largely agreed that existing centralized university resources were inadequate. Institutional repositories, in particular, were often regarded as ill-equipped for the volume of data that researchers generated in the course of their work. An engineering professor, for example, noted that their libraries' repository had a capacity of three to four petabytes. Yet, their lab alone stored six petabytes of data. "We already have more storage needs than . . . the library provides to the entire campus." How, the professor wondered, can the library position itself as a leader in data stewardship, if it lacks the capacity to store the contents of even a single lab? The volume of data accumulating in individual labs was a consistent problem for researchers, and the question of how to store the staggering volumes of data being generated across a research university is a major strategic challenge research university leaders face.

> "We already have more storage needs than . . . the library provides to the entire campus."

One common strategy that universities have adopted to meet this challenge is to encourage the use of commercial cloud-based services such as AWS, which effectively outsource the computing resources required to interpret and store data. A number of interviewees reported feeling pressured to use these resources. However, researchers were generally skeptical of cloud services, with most preferring to invest in local storage options. Some researchers favored local storage because it allowed them to keep closer tabs on their data or because they worried about privacy issues associated with cloud storage. However, regulations around the use of grant funds drove much of the opposition to cloud storage. Speaking for many of their colleagues, one researcher noted that "cloud computing is very powerful, but it requires monthly payments. And the problem is that I don't have a budget that allows me to make monthly payments for computing costs." Because ongoing, subscription-based charges cannot normally be charged to research grants, faculty purchased their own hardware, in forms ranging from external drives to dedicated servers. Due to funding structures, researchers tended to hoard storage space in their labs, a model that many felt was adequate for storing data, but that creates continual missed opportunities for economies of scale, hinders data sharing efforts, and contributes to the sense, endemic in many of the project interviews, of disconnection that many researchers reported feeling towards their universities.

Many campuses also host their own high-performance computing centers (HPCs) for use in conducting large-scale analysis and experimentation. Several researchers interviewed for this project reported using these resources, but the disciplinary background of researchers significantly impacted their perception of the value of local HPCs to their work. Some reported that access to HPCs at their institution was largely restricted to a few fields, or that the specific

builds of their HPC were of limited use for their research. An environmental scientist, for instance, reported that "the centralized computing facilities at my institution are developed by engineers for engineers and bioinformatics and life sciences computing is an afterthought." The "architecture, the system and the accessibility of large amounts of storage from compute nodes" didn't fit the needs of their research project. Similarly, another researcher noted that their university's HPCs "seems like they're not designed for our kind of work." One common work-around was to use computing clusters at other institutions that were better suited to their needs, provided, of course, one had developed the collaborative relationships necessary to make this possible.

Given the diverse technical requirements of big data research, it is difficult to imagine any individual institution being able to adequately support the computing requirements of all its researchers. Cloud computing is one possible solution to this problem, but financial obstacles deter many faculty members from using cloud resources, since they charge based on usage, leaving researchers exposed to surprise bills. One researcher, sharing a story echoed by others, described having had students who accidentally left GPUs running, creating bills that the researcher then had to pay. For financial and technical reasons, individual labs will likely continue to depend heavily on their own computing infrastructures for analysis and storage. Most researchers seemed able to patch together computing power sufficient for their needs. However, this reliance on decentralized infrastructure can pose challenges to universities. In perhaps the most extreme example of the heavy technical demands big data research can put on campus resources, professors of physics and engineering at a major research university noted that their lab's computing demands regularly run up against the limits of the university's electrical grid, which struggles to cool their machinery. This was not (yet) a common worry that researchers expressed, but it is illustrative of the costs of supporting big data research.

Sharing Knowledge

Peer-reviewed articles remain the most highly incentivized form of scholarly communication. However, data and code are increasingly considered important research outputs and the formal and informal sharing of both are now expected in many academic fields. Researchers must often juggle their ethical commitment to open data with the reality of working with data that is sensitive, proprietary, or otherwise difficult to make openly available.

Communicating Research Outputs

Scholarly communication norms have changed more slowly than research methods, thanks to incentive structures that continue to privilege traditional scholarly outputs for promotion and tenure and professional prestige. Except for a few outlier disciplines—including "book fields" in the humanities, and fields like computer science, where conference and conference proceedings were valued for their greater immediacy—the peer-reviewed article remained the *sine qua non* of scholarly outputs. Researchers prioritized peer-reviewed scholarship over communicating with nonacademic audiences and sharing code and data. Their reasons for doing so were straightforward: most worked in departments where promotion and tenure standards reinforced

conservative approaches to scholarly communication. Few researchers felt that their institutions or departments had created incentives to encourage data or code sharing or communication with public audiences. Indeed, most believed that existing infrastructures discouraged these practices, except insofar as they were required conditions for publishing research or could be shown to lead to increased numbers of citations.

Nevertheless, the interviews demonstrate that efforts to promote open research cultures, particularly in the sciences, are bearing fruit. One clear example is the growing use of preprints in STEM-fields, something that many researchers noted was rapidly becoming a standard means of disseminating research, confirming a trend explored by analysts at Ithaka S+R and others who follow practices in scholarly communications.[21] (Our interviews turned up little evidence that preprints were being used outside of STEM fields.) Researchers valued preprints for their speed and efficacy, something particularly important in fast-moving fields like computer science, where researchers faced pressure to stake early claims to developments as quickly as possible. Others found the opportunity that preprints provided for informal peer review to be very useful. Preprints were not universally accepted: some found them cumbersome, questioned their value, or worried that preprints increased their exposure to research theft, since they lacked copyright protection. Scholars seem more likely to use preprints than to publish in open access journals, which in many fields were regarded as relatively low prestige venues. Data and code sharing practices are explored in more detail below.

Big data researchers, like their colleagues using traditional methods, expressed mixed feelings about the value of communicating their work to nonacademic communities, though many had made at least sporadic efforts to do so. Some remained skeptical of the value of public outreach, noting for instance that "I don't see my role as sharing information out with the general masses outside of academia." Another indicated that while they would like to do more public-facing work, "it's not actually something I do regularly, and it's not something I feel well situated to do in terms of my time or capacity." However, efforts to communicate with the public audiences were relatively common and ranged widely in genre and media, from personal blogs or lab websites, presentations to local audiences, including K-12 classes, TED talks, white papers, volunteering with community organizations, and social media. Said one researcher, "I try personally very hard when publishing something, even if it's somewhat esoteric, to do the TLDR [too long didn't read] version of it on social media, as well as I have my own personal blog where I'll write a short synopsis." Among social media platforms, Twitter and Facebook were predominant, though Instagram, and LinkedIn were also mentioned as venues for disseminating research findings to both academic and nonacademic audiences. On occasion, researchers' work had been featured on the news or in popular media publications. A few

[21] Oya Y. Rieger, "Preprints in the Spotlight: Establishing Best Practices, Building Trust," *Ithaka S+R*, May 27, 2020, https://doi.org/10.18665/sr.313288; Andrea Chiarelli et al., "Preprints and Scholarly Communication: An Exploratory Qualitative Study of Adoption, Practices, Drivers and Barriers" (F1000Research, November 25, 2019), https://doi.org/10.12688/f1000research.19619.2; "Rise of the Preprints," *Nature Cancer* 1, no. 11 (November 2020): 1025–26, https://doi.org/10.1038/s43018-020-00151-y; François-Xavier Coudert, "The Rise of Preprints in Chemistry," *Nature Chemistry* 12, no. 6 (June 2020): 499–502, https://doi.org/10.1038/s41557-020-0477-5; Rob Johnson and Andrea Chiarelli, "The Second Wave of Preprint Servers: How Can Publishers Keep Afloat?" *The Scholarly Kitchen*, October 16, 2019, https://scholarlykitchen.sspnet.org/2019/10/16/the-second-wave-of-preprint-servers-how-can-publishers-keep-afloat/.

specifically mentioned either drafting press releases or coordinating with university communications departments to encourage media attention. Researchers who had received industry funding often reported their findings to those sponsors. Overall, many researchers do seem to make at least some efforts to communicate their findings with audiences beyond their immediate academic communities.

Nevertheless, there was widespread agreement that reaching public audiences was difficult, and less important than other forms of scholarly communication. Beside the ubiquitous pressures of time and the imperative of consistently producing peer-reviewed publications, researchers frequently felt that the complexities of data-driven research made communicating with non-specialist audiences especially difficult. One interviewee, for example, described the challenges of describing both the intricacies of data science and the nature of drug interactions without relying heavily on jargon. Data science, in particular, this individual felt, had not yet developed a vocabulary that was "publicly understandable." Researchers in the "big data space" frequently worked with ideas that "are barely understood by the experts," and had not yet "figured out the right ways to talk about it in a more understandable way." A computer scientist who worked in public health research concurred. Especially for assistant professors trying to earn tenure, the task of translating material written for specialist audiences into a "layman's version of the research" was difficult to justify.

Communicating with other scholars was not necessarily any easier. Several interviewees described the interdisciplinary nature of their work as ill-suited to an academic publishing industry still organized largely around disciplines. "Finding an academic home," said one public health researcher, can be difficult when one's work "straddle[s] between many different fields." Often, the researcher ended up feeling caught in the middle, uncertain where to publish and frequently told by editors that their work wasn't a good fit for their journal. Another scholar, with an appointment in mathematics and statistics, worried that the disciplinary orientation of journals distorted their research by forcing them to split their publications into discrete, discipline-specific components. They felt pressured to "avoid talking about interesting biological things and focus on statistics," when trying to publish in statistics journals, and in turn to minimize the statistical aspects of their work when publishing in biology journals. Likewise, peer-reviewers were difficult to find and tended to push revisions towards one discipline or another. The researcher expressed concern that the full importance of their work ended up lost in the publishing process as it was broken down into disconnected pieces. Several others described difficulties finding audiences for their research. For those seeking tenure, this could pose acute problems. One faculty member, hired with a joint appointment in computer science, history, and English, described competing promotion and tenure standards that made it difficult to understand how they would be rewarded for interdisciplinary work, since truly interdisciplinary research wasn't necessarily the research that would "easily get me promotion or chits on my yearly review." Other researchers worried that their CVs ended up looking very fragmented, as their publications were spread across disciplines and topics.

Emerging Ethics of Openness

Open access to data, code, and publications has emerged as a major issue in scientific research and communication over the past decade, driven by concerns about reproducibility and equity of access. Big data researchers, who frequently rely on secondary data, benefit greatly from open access resources, and our interviews suggest that the principles of open science are embraced by the majority of those interviewed for this project. "I'm a professor," said one faculty member, "so my job really is to do new work and share with the world." Many interviewees felt an ethical obligation to share data in the interests of advancing science and the public good. In addition to these abstract commitments, many researchers identified tangible benefits of professional reciprocity within research communities. "I definitely benefit a lot from the open data ecosystem," said a physicist, "and I'd like to contribute back to that." A computer scientist regarded the open sharing of code and data as a "major driving force behind the rapid advance in computer science," while an engineer remarked that "the open-source data and software community," which they drew on daily for their research, is essentially "a worldwide thing at this point." For many, this practice is normalized to the point of being taken for granted: "I've never known, really, a mode of scientific practice that didn't involve sharing your data and code."

These sentiments were widely shared, but it is also clear that many scholars require additional incentives to consistently undertake the substantial labor of documenting and preparing code and data for sharing. At many institutions, this labor was not rewarded by promotion and tenure committees. Indeed, except in schools of data science and certain medical fields, most scholars agreed that sharing data "doesn't carry the same weight" as publishing articles and that sharing data is "just a service"—morally upright behavior, but not something likely to result in professional rewards or accolades. The competitive and individualistic culture of academic research stoked concerns that sharing data or code could lead to being scooped or otherwise having one's work stolen by other researchers.[22] "I don't want people publishing off my work," said a political scientist who refused to share data until they felt they had exhausted its value. "It gets a little selfish," they continued, but "if I put forth all this effort to put together this data set, I'm definitely not going to just give it away to somebody to build their career before I've had a chance to get as much out of it" as possible. Some researchers specifically described these concerns as rooted in cultural norms of fields without strong traditions of data sharing. A law professor, for example, explained that their field did not have a robust "culture of attribution," while a faculty member in a business school remarked that "in business we're about competition, it's just our nature." Meteorologists, physicists, and engineers also feared data sharing could lead to being scooped, so it was not just professional school faculty who had these concerns.

Requirements from funders and journal editors are clearly the most powerful incentives pushing research cultures towards open science. As one engineer put it succinctly, people share data "because there's no choice." "Anything that's tied to funding is a strong incentive," said a researcher, while another agreed that the "main incentive, really, is that NSF has their data

[22] On the deleterious impacts of individualism in research cultures, see Kathleen Fitzpatrick, *Generous Thinking: A Radical Approach to Saving the University* (Baltimore: Johns Hopkins University Press, 2019), 26-33.

management requirements." Researchers in fields spanning from demography, political science, biology, engineering, economics, to oceanography indicated that journals in their field often required data or code sharing.[23] The imperative to publish is strong enough to ensure that scholars will comply, but without those incentives, relatively few will make the effort to formally deposit data or code in open repositories, though they might still share informally with colleagues.

Formal and Informal Sharing of Code and Data

The gold standard for dating and code sharing are those that meet FAIR standards, which usually entails permanent archiving in a public repository. This is the most equitable and durable form of data sharing, central to the goals of open science, and is what many researchers seemed to have in mind when they described data sharing as an ethical practice. Preparing data and code to meet FAIR standards is a labor-intensive proposition, something many researchers described as a serious barrier to participating in formal data sharing.[24] Concerns about the labor involved in formal sharing were often compounded by a sense that much of the data researchers generated was of limited interest outside tight-knit research communities who already informally shared data with each other. As an agricultural economist noted, for any given project they work on, there's "probably an audience of zero to three" who might be interested in the raw data or underlying code. An oceanographer described writing highly-specialized code to measure gravity using radar altimetry, something of interest to "probably only five groups in the whole world." Sharing directly with those labs was relatively straightforward because they already had personal relationships with interested colleagues, and could share data with minimal documentation, because it was being shared with researchers who already understood its purpose and context. Sharing with wider audiences, these researchers believed, was rarely worth the effort. In short, the interviews suggest that sharing should be conceptualized as a spectrum of activities shaped by communities, disciplines, cultures, and incentive structures. Many of the researchers interviewed for this project engage in less formal sharing of data (and, to a lesser extent, code) as either a supplement to or replacement for more codified forms of sharing.

Another common practice, a middle ground of sorts between the relative insularity of sharing only with colleagues and fully open sharing, is the practice of making data available on request. Several researchers preferred this means of sharing data because it allowed them to vet potential users and identify potential collaborators while minimizing the labor required to preparing their data for formal archiving. However, making data available on request is a somewhat controversial practice – one interviewee called it "B.S.," noting that researchers are likely to lose

[23] For an overview of disciplinary trends in data sharing policies in academic journals, see Michal Tal-Socher and Adrian Ziderman, "Data Sharing Policies in Scholarly Publications: Interdisciplinary Comparisons," *Prometheus* 36 no 2 (2020): 116-134 and Antti Rousi and Mikael Laakso, "Journal Research Data Sharing Policies: A Study of Highly-Cited Journals in Neuroscience, Physics, and Operations Research," Scientometrics 124, no. 1 (2020): 131-152.

[24] As David Crotty has pointed out, data alone will rarely reproduce complex experiments absent proper documentation. See David Crotty, "Reproducible Research, Just Not Reproducible By You," *The Scholarly Kitchen*, May 24, 2017, https://scholarlykitchen.sspnet.org/2017/05/24/reproducible-research-just-not-reproducible/.

track of their data if it isn't responsibly and publicly stored. Sharing on demand can also be seen as inadvertently reinforcing social and academic hierarchies. However, it is a practice that some researchers, motivated by concerns about time and energy, or a desire to retain control over the circulation of their research outputs, consider it sufficient to meet their ethical obligations around sharing.

Barriers to Sharing

Researchers often depend on data generated by others for their own research, and most agreed with the principles of open access. However, many voiced concerns that the ideology of open access did not account for the complexities involved in big data research, pointing to important barriers to sharing that limited their ability to participate in fully open science and raised concerns about across-the-board, one-size-fits-all implementation of policies requiring disclosure of code or data.[25] One particularly significant limitation occurred when researchers used ethically sensitive data, proprietary data, and secret or restricted data.

Several researchers described ethical issues that restricted their ability to freely share data. This arose frequently when researchers conducted human subject research, especially among those working in medicine and public health. IRBs, institutional policies, and federal legislation have implemented rules to protect individual privacy and guide human subject research ethics, and frequently limit the release of data and encourage conservative approaches to data release. Sometimes the solution to these problems is to ensure that data is properly deidentified, but the issue of privacy points towards a key limit on open data: some data may best be left uncirculated. This issue came up in various contexts. The PI of a digital humanities project involving historic photographs, for instance, was reticent about sharing photographs (even though copyright was not an issue), because the individuals depicted could not have envisioned the future of use of their images and were sometimes shown in culturally sensitive contexts. Another researcher, who worked with real estate data, worried about commercial exploitation of their data. A genomics researcher described excluding data from publication that had revealed a pattern about an underrepresented community that "could have [negative] implications for communities that we study." Others grappled with the potential implications of publishing data that might endanger members of religious minorities in China, or of political speech harvested by scraping social media. Collectively, these examples serve as necessary reminders that the ethical dimensions of big data research remain contentious and that the tensions between openness and privacy are on some level unresolvable.

Sharing raw, and in some cases even heavily processed, data is impossible for many researchers who work with restricted, secret, or proprietary data. A civil and environmental engineer who works on sensitive infrastructure issues described choosing not to release certain data because of fears they might "end up doing more harm than good by giving information to people you don't want to give it to." The same researcher, who also did contract work with the Department

[25] Natasha Susan Mauhnter and Odette Parry, "Open Access Digital Data Sharing: Principles, Policies and Practices," *Social Epistemology* 27, no 1 (2013): 47-67, http://dx.doi.org/10.1080/02691728.2012.760663. For a detailed look at informal data sharing and barriers to sharing in a specific field, see Danielle Cooper et al., "Supporting the Changing Research Practices of Civil and Environmental Engineering Scholars, *Ithaka S+R*, July 16, 2019, https://doi.org/10.18665/sr.310885.

of Defense, noted that important projects sometimes went unstudied because researchers declined to work on projects that they couldn't talk about with their peers or publish papers about. Some researchers have adopted workarounds, by publishing heavily processed data or directing inquiries to the company who owns the data they have used. Finally, though issues around proprietary data came up primarily when researchers collaborated with industry, a few researchers were hesitant to release code because their university hoped to patent and monetize it. As an engineer at a large public university noted, their institution is invested in supporting open science, but "the open sourcing kind of thing goes out the window" when it seems possible to profit from a researcher's code.

Preparing data for sharing was another major area of concern for researchers. As one researcher noted, journals and funders often required indefinite storage of data, but it was ultimately up to individual researchers to make their data accessible and uncorrupted. Speaking for many of their colleagues, an assistant professor of mechanical engineering described their struggles with preparing code for open-source release, noting that "thus far, I haven't identified specific funding mechanisms that will pay for that." Barriers relating to cost are sometimes framed as disciplinary matters: Ruth Ahnert et al., for instance, have recently argued that digital humanists should be held to lower standards of transparency and reproducibility than STEM researchers because of the "huge and unrealistic burden they place on individual researchers."[26] While it is true that some disciplines—humanities and social science fields in particular—are more likely to pursue big data research on shoestring budgets, the better-staffed and funded projects pursued in STEM fields often create disproportionately large amounts of data, ultimately presenting similar challenges when it comes to documenting data.

Finally, some projects have generated so much data that sharing it is difficult. One computer scientist described having 56 terabytes of data associated with a project, only some of which would be made publicly available because it was too large to download. A biologist reported similar problems, noting that they sometimes made data available on request simply because this was the only practical way to share it was via sneakerware. Given the escalating size of datasets, this challenge is likely to become more acute, as will the monetary costs associated with sharing data and code.

End-to-End and Point-to-Point: Support and Training Practices

Researchers tend to favor informal training methods, such as internet tutorials, over formal training in big data methods. While these DIY methods work well for solving immediate problems, they are less well suited to acquiring foundational knowledge, leaving the potential for blind spots in academic research.

[26] Ruth Ahnert et al., *The Network Turn: Changing Perspectives in the Humanities* (New York: Cambridge University Press, 2020), 97.

Problem-Solving

When asked how they learned new skills, researchers overwhelmingly favored DIY approaches focused on learning "learning just enough" about software and programming languages to move projects forward, with an emphasis on finding immediate answers to specific problems. The tools of choice were usually online resources. An assistant professor of mechanical engineering, sharing a sentiment echoed by many of their colleagues, said that "If I encounter something that I don't know how to do, I just Google how to do it." Online tutorials, MOOCs, YouTube videos, and a veritable host of similar resources were all valued for their efficiency at providing quick, actionable, help. As a full professor in a law school described it, these types of resources were in some ways "better than any training because if you get something you can move on, right? Or if there's something specific you want to focus on, you can just go straight to it, instead of learning things that have no relevance to you like how to do a math problem in Python. I don't really care, right, I'm not going to do that." The flexibility of online resources offers the ability to "get to the part that I want to get to and not have to suffer through everything that's somewhat extraneous, even though I'm sure it's useful for someone."

In contrast, researchers often portrayed the workshops in specific software or programming languages offered by libraries, computing labs, and other university units as inflexible and overly generic. Perhaps just as importantly, they required a commitment to learn at a specific time, rather than offering the promise of having questions answered in something approaching real time. One researcher described having decided not to attend future library trainings because they had decided that "there's nothing you can learn there that you wouldn't learn in the same hour reading something online." Overall, faculty seldom attended university sponsored workshops and training events. Those who had frequently described them as too basic, better suited to novices than to intermediate users. Several described a "donut-hole" problem, with help available for beginners and for very advanced users, but relatively few offerings for those in between.

Perhaps as importantly, researchers simply preferred training tailored to their specific needs. When available, researchers felt they benefited more from one-on-one consultations, where they could receive expert advice on pressing issues. Interest in just-in-time, highly individualized support—of a type that might be better described as collaboration or consultation than training— was widespread. It is also a request that universities will be hard-pressed to fulfill. The general lack of enthusiasm for introductory training among faculty suggests that libraries and other units may be wise to avoid adding more of these services aimed at this segment of the research community. Faculty did see value in those offerings for students, who might be targeted as audiences. Libraries might also consider creating curated resources for common programming languages and software packages, and—where possible—individually tailored consultation services. Existing offerings relating to data management plans, discussed below, might offer models for such services.

Data Management

Faculty often expressed concerns about their ability to create data management practices including version control, metadata, and other logistical aspects of organizing data for use, publication, and preservation. This was a particularly daunting issue as the size of the datasets they worked with grew. A researcher in the information sciences described the difficulty of "knowing when has your project outgrown a spreadsheet, when has it outgrown trying to manage a bunch of files on a server," and of "knowing how to plan to future-proof your data management projects so that you have a plan" for adapting when the growth of a dataset requires adopting new systems and software. "Putting the data and code in the right place, and keeping track of how they run with each other," they said, is a constant challenge. An agricultural economist described a similar feeling of bewilderment as their research scaled up to include millions of data points. "I've never received any training," they remarked, about handling the influx involved in having your dataset increase by several orders of magnitude. A mathematician at another institution concurred: "the biggest challenge is organizing the data that I collect." "I don't know how to do it," they continued, and their efforts to find models had borne little fruit, hindering their ability to make an impact with their research.

The DIY ethic that pervades much faculty training is ill-suited towards systematic issues like developing data management plans (DMPs). Researchers frequently felt overwhelmed by the demands of developing file naming conventions, version control for code and data used and updated by multiple team members, and creating metadata. Libraries and other university units have recognized this need and often offer data management support services.[27] Our interviews suggest that few faculty have utilized those services. Those who had done found them useful—indeed, though mentioned only a few times, assistance with data management was one of areas where researchers were most likely to have turned to librarians for help, and those who had done so uniformly reported that they had benefited from the advice.

Graduate Education

Graduate education is an important part of conversations about big data. Graduate students perform much of the actual labor involved in research projects, and graduate education is where faculty and researchers develop their research skills, making it a key venue for developing what one interviewee described as the "end-to-end" competencies necessary to mitigate the growing risk of black box problems. A professor of civil engineering described the challenges of preparing graduate students to be well-rounded data researchers, illustrating a difficulty that affects faculty as well. Students, the professor said, can't simply be prepared to make narrow technical contributions to ongoing projects, which is too often the case now. As future researchers, they need an integrated conceptual understanding of the range of disciplinary perspectives and specific knowledge that big data relies upon, an "end-to-end" perspective. Big data practitioners, the engineer continued, require "a diverse toolkit rather than an approach that you rely on over and over again." This ideally involves a familiarity with the philosophy of data and science,

[27] Radecki and Springer, "Research Data Services in US Higher Education: A Web-Based Inventory"; Blummer and Kenton, "Big Data and Libraries: Identifying Themes in the Literature."

understanding the strengths and weaknesses of multiple models or methods, and an awareness of the ethical and political implications of a given research project.

Researchers also believed that many of their students had serious gaps in their foundational knowledge of computer and data sciences. One researcher noted that doing data science research involved more than learning to code. Students also require a grounding in data and statistics and an understanding of "what to do, how to approach the question, how to structure the data, how to get the data": absent this conceptual foundation, "no amount of teaching them about R or Python is going to help." Another described spending substantial time explaining the need for rigorous cleaning, testing, and validating data as well as basic principles for data analysis. The lack of "high-level overview training," they believed, regularly required them to spend months getting students up to speed on basic data science techniques before they could usefully contribute to the research team. Many graduate students arrived with little experience working with large datasets and are unprepared for the challenges associated with cleaning and interpreting data at scale.

Students often had similar gaps in basic computing knowledge. Across disciplines, faculty described their graduate students as unfamiliar with installing and working with Unix systems, using command lines, and navigating GitHub. Said one neuroscientist, graduate students usually brought content, but not programming, backgrounds to the table. "The concept of telling a computer to do something that isn't a point and click interface," is something they have "just never done." A humanist colleague teaching at a different university concurred, noting that even students who wanted to learn basic coding and data science skills found it difficult to do so while keeping up with the demands of their disciplinary training. In humanities fields, students had to justify bothering to spend time learning to code at all. Even in more hospitable disciplines, the competing demands of developing coding and domain knowledge presented significant challenges, despite faculty who advocated for the need to weight disciplinary training towards a greater emphasis on data science. An environmental scientist declared that "all scientists need to know the basics of computational thinking," a sentiment echoed by an associate professor of biology, who believed that teaching students to work with data, to build a toolkit focused on coding and data science, would provide more versatile instruction than one exclusively focused on biology. This mentality extended even into professional programs. A physical therapist thought that even programs designed to train practitioners had to reckon with the fact that being a professional now required learning to "collect, organize, understand, and utilize data to help lead the footprint on the ground." "That's where our field is going," the instructor said, "whether we like it or not."

One possible venue for acquiring foundational knowledge is through coursework in other departments, such as computer science. However, faculty expressed some skepticism about technical training that occurred outside a student's home department. A neuroscientist, for example, often sent their students to the computer science department to learn Python. However, because those courses were taught from within the perspective of computer science, students interested in learning how to use the language to analyze behavioral experiments could easily become lost. Faculty were slightly more willing to recognize that workshops and trainings offered by libraries and other campus units provided learning opportunities for their students

than for themselves, but rarely directed students to those resources, which they also felt suffered from being divorced from the actual research projects students would work on.

In practice, this means that much training around coding and data science occurs within individual labs. The idea of integrating coding and data science into disciplinary training, especially into graduate education, was identified as a desirable pathway towards contextualized technical expertise, but given the demands on curricula, will likely continue to be left to the initiative of individual labs and students. Graduate education, like many structures of higher education, is not well-equipped to provide deep training in multiple disciplines. As a professor of civil and environmental engineering put it, building the knowledge required to work fluently in big data requires "deep" rather than "shallow" interdisciplinarity, "which means you get to do at least double the work of normal PhDs. But the reward, of course, is that most of the interesting problems these days happen to be at the intersection of fields." Gaining all these competencies is a challenge, and many researchers described struggling to balance the day-to-day demands of research with the need to cultivate high-level perspectives on the lifecycle of research projects. As the above quotes suggest, graduate curricula are where researchers obtain their deepest foundational knowledge and should be priorities for institutions interested in supporting big data research. This is no easy task, as graduate education is rooted in disciplinary perspectives and departmental structures, curricula are already overloaded, and universities are facing pressure to reduce than add time to degrees.

Nevertheless, the absence of deep, multi and interdisciplinary knowledge left many researchers worried that students (and by extension faculty) risked falling prey to black box issues that could distort their research. Concerns about "black boxes" came up repeatedly, particularly in relation to software and AI. A materials science and engineering professor worried that "as the software becomes more powerful and easier to use there is greater and greater temptation to use it as a black box," a risk especially true of machine learning algorithms, which could tempt students and researchers to accept outputs without understanding either the mathematical structure of the data, biases of training datasets, and inner workings of the code. Less often, researchers expressed concerns that a lack of fluency with data science techniques or statistical analysis could cause similar blind spots to those that software might create.

Big Data Ethics

Researchers who worked with human subjects struggled to balance competing ethical commitments and to understand best practices around ethical research. Despite the bureaucratic hassles involved, guidance from IRBs was welcome, but not always seen as sufficient. Recent research has raised questions about the suitability of extant IRB guidelines to handle big data methodologies, for example, in projects involving data scraped from social media sites, where practices of informed consent, foreseeable harm, and even the definition of a human subject are fraught with new implications.[28] Several researchers echoed these concerns.

[28] boyd and Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," 673; Maddalena Favaretto et al., "Working Through Ethics Review of Big Data Research Projects: An Investigation into the Experiences of Swiss and American Researchers," *Journal of Empirical Research on Human Research Ethics: An International Journal* 15, no. 4 (2020): 339–54; Agata Ferretti et al., "Big Data, Biomedical Research, and Ethics Review: New Challenges for IRBs," *Ethics &*

A political scientist described IRBs at their institution as "very much behind the times," when it came to big data research, especially regarding research data scraped from social media. One particular area of ethical ambiguity is that such practices often involved technical violations of terms of service. Within the field, they noted, researchers made personal decisions about when and how such violations might be ethically absent substantive guidance from IRBs. A computer scientist noted similar issues, feeling that their work—which also involved aggregating internet data, fell into a "grey area of research."

Faculty raised a range of ethical concerns that they had to navigate, including how to properly de-identify data, and worries that their surveys or algorithms might perpetuate racial bias. These concerns were most often voiced by human subject researchers. Somewhat disturbingly, many researchers who did not work with human subjects and were thus exempt from IRB approval, described their work as ethically neutral. A biologist, for example, saw no ethical implications to their work because "we don't work with any animal or human-generated data." A chemist described their work as primarily computational and virtual, so "there's nothing to think about in terms of ethics." A postdoc in the geosciences concurred, seeing few ethical implications in their work "because our projects are mostly related to natural science not directly related to people." While some faculty believed that conversations about ethics and big data had become more widespread, project interviews suggest the need for more spaces for discussion about ethical dimensions of big data. Researchers voiced particular interest in opportunities to learn more about deidentification techniques, data security, and avoiding racial bias in big data research.

Conclusions

Big data research is resource intensive and here to stay. Even if we don't wish to hype it as a "disruptive" force, its effects on disciplinary perspectives and research methodologies are clear. As big data grows, the difficulty of supporting the research mission of universities—already a substantial challenge for administrators—will increase. Making big data sustainable, if that is possible (its carbon costs are daunting), will require coordinated action by universities, something that is difficult to accomplish at institutions with decentralized bureaucracies and cultures. Research labs are often more connected to labs at other universities than to their home departments and institutions and are protective of their agency and independence. Unsurprisingly, on many campuses, the big data infrastructure is highly decentralized, often based principally in individual labs, and the connective tissues between PI's and other university units are often weak and dependent on personal relationships.[29] The highly specialized nature of big data research creates additional barriers to centralized infrastructures, making it likely that the already pronounced trend towards inter-institutional collaboration will continue. The

Human Research 42, no. 5 (2020): 17–28; Alexandra Paxton, "The Belmont Report in the Age of Big Data:: Ethics at the Intersection of Psychological Science and Data Science," in *Big Data in Psychological Research*, ed. Sang Eun Woo, Louis Tay, and Robert W. Proctor (American Psychological Association, 2020), 347–72, https://www.jstor.org/stable/j.ctv1chs5jz.19.

[29] Brian Lavoie, "Working across Campus Is like Herding Flaming Cats," *Hanging Together*, June 2, 2021, https://hangingtogether.org/?p=9345.

decentralization of the research university is, in some respects, a source of strength, but it creates barriers to economies of scale and to the efficacy of support infrastructures provided by departments and research centers, libraries, computing centers, and IT and information professionals. These challenges are exacerbated by the fact that supporting big data research on campus will ultimately require coordination with actors beyond the university but central to the research ecosystem, including scholarly societies, publishers, industry, and funders.

Recommendations

University Research Offices

- Develop protocols for periodic systematic assessment of on-campus big data infrastructure, mapping resources and assembling working groups across IT, libraries, HPC, research offices, and other relevant units to coordinate support services, identify gaps, and reduce redundancies. Consider developing a formal catalog of data services and resources for circulation to researchers.
- Assess whether current IRB standards and university legal guidance adequately reflect ethical and privacy issues associated with big data research.
- Develop internal funding opportunities for under-resourced fields, including humanities disciplines, qualitative social sciences, and some professional disciplines. Consider prioritizing and marketing workshops and other training and support programming to researchers from these fields.
- Fund fellowship programs designed to provide opportunities for graduate students to gain expertise in data science and programming while contributing to ongoing big data research. Consider funneling fellowship recipients into fields with few opportunities for external funding of graduate student lines.
- Establish consortium relationships with other universities to build long-term data storage and computing capacity. Regularly assess on-campus active data storage needs and capabilities.
- Develop personnel and project management training for researchers in recognition of the essentially collaborative nature of big data research.

Departments

- Invest in strategic hires designed to further embed data science, data management, statistical, and computational staff to provide researchers with relevant expertise to assist in big data research.
- Doctoral programs, particularly in STEM fields, should seek opportunities to integrate foundational competencies in machine learning methodologies, data science, and programming into their doctoral curricula.

- Undergraduate majors that are likely to draw significant numbers of students into big data research teams should consider curricular interventions designed to build students' capacity to understand and contribute to big data research.
- Revise promotion and tenure standards to recognize well-organized data and code sharing as a significant research output.

Libraries

- Host forums, seminars, symposiums, and other opportunities for researchers involved in data-intensive research to share and network across disciplinary lines.
- Create and update curated guides to datasets of interest to specific research communities. To save labor, consider developing these in collaboration with other academic libraries.
- Allocate additional resources to purchasing subscription datasets to reduce costs to researchers.
- Develop staff expertise in metadata creation, data curation, and data management, as well as data analytics and data visualization.
- Increase marketing of extant data research management services, which are in high demand among researchers.
- When feasible, expand one-on-one consultation services or offer on-demand workshops tailored to the needs of specific research groups.
- Increase storage capacities of institutional repositories and increase marketing of them to researchers.
- Tailor workshops to students working in big data focused labs and researchers from fields that are least acclimated to technical, programming, or quantitative skills.

Funders

- Assess whether current legal and ethical guidance and standards required of grantees adequately reflect the emerging ethical and privacy issues associated with big data research.
- Develop mechanisms to provide funding for maintenance costs associated with long-term data storage, such as those incurred by users of cloud storage.
- Continue to support the robust development of data repositories.
- Assess how well existing code and data sharing regulations provide specific guidance to researchers about the proper handling of proprietary, secret, sensitive, low quality, and secondary data.
- Encourage departments to adopt promotion and tenure standards that recognize well-organized data and code sharing as a significant contribution to scholarship. To support this goal, fund efforts to ensure and evaluate the quality of data and code shared through repositories and other open forums.

Scholarly Societies

- Articulate discipline-informed perspectives on research ethics.
- Use meetings and publications to encourage nuanced discussion about the value of open sharing and the complexity of decision-making around data sharing. Coordinate with researchers, funders, publishers, and other stakeholders, to articulate data disposal policies and storage standards.
- Encourage departments to adopt promotion and tenure standards that recognize well-organized data and code sharing as a significant contribution to scholarship.

Vendors

- Enhance metadata of subscription databases.
- Coordinate with libraries rather than individual researchers to license datasets. Provide educational license and packaging to make datasets accessible and affordable to university communities.
- Develop fixed-price cloud storage options developed in consultation with university offices and research communities.
- Develop individualized consultation services to assist researchers with specific coding and data management challenges.

Appendix 1: Teams and Local Reports

Atlanta University Center Consortium

Bryan Briones, Justin De La Cruz, Rosaline Odom, "ITHAKA S+R Supporting Big Data Research Project."

Boston University

Paula Carey, Kate Silfen, "Supporting Big Data Research at Boston University Report: A Study Conducted in Partnership with Ithaka S+R."

Carnegie Mellon University

Neelam Bharti, Patrick Campbell, Hannah Gunderman, Huajin Wang, "Understanding the Research Practice and Service Needs of Big Data Researchers at Carnegie Mellon University Report," https://doi.org/10.1184/R1/16701958.v1.

Case Western Reserve University

E.M. Dragowsky, Ben Gorham, Jen Green, Roger Zender, Lee Zickel, "Supporting Big Data Research at Case Western Reserve University: An Ithaka S+R Local Report," https://digital.case.edu/islandora/object/ksl:2006079601.

Georgia State University

Kelsey Jordan, Bryan Sinclair, Mandy Swygart-Hobaugh, Jeremy Walker, "Supporting 'Big Data' Research at Georgia State University," https://scholarworks.gsu.edu/univ_lib_facpub/141/.

New York University

Vicky Rampin, Margaret Smith, Katie Wissel, Nicholas Wolf, "Supporting Big Data Research at New York University," https://archive.nyu.edu/handle/2451/63363.

North Carolina A&T State University

Tracie Lewis, David Rachlin, Iyanna Sims, "Supporting Big Data Research at North Carolina Agricultural and Technical State University: An Ithaka S+R Local Report."

North Carolina State University

Karen Ciccone, Susan Ivey, John Vickery, "Big Data Research Practices and Needs at North Carolina State University: An Ithaka S+R Local Report," https://repository.lib.ncsu.edu/handle/1840.20/39112.

Northeastern University

Jen Ferguson, Kate Kryder, James Macalino, Julia Unis, "Big Data and Data Science Research at Northeastern University – Final Report," https://repository.library.northeastern.edu/files/neu:ww72bs68t.

Pennsylvania State University

Seth Erickson, Lana Munip, Cynthia Vitale, Cindy Xuying Xin, "Big Data Research Support at the Penn State University," https://scholarsphere.psu.edu/resources/e3e74ea2-1b9b-4194-abca-1f71fff3c8de.

Temple University

Will Dean, Fred Rowland, Adam Shambaugh, Gretchen Sneff, "Supporting Big Data Research at Temple University," https://scholarshare.temple.edu/handle/20.500.12613/7068.

Texas A&M University, College Station

Carolyn Jackson, Laura Sare, Paria Tajallipour, John Watts, "Assessing the Research Practices of Big Data and Data Science Researchers at Texas A&M: An Ithaka S+R Local Report," https://oaktrust.library.tamu.edu/handle/1969.1/194888.

University of California, Berkeley

Erin D. Foster, Ann Glusker, Brian Quigley, "Supporting Big Data Research at the University of California, Berkeley: An Ithaka S+R Local Report," https://escholarship.org/uc/item/4403c0f4.

University of California, San Diego

Stephanie Labou, David Minor, Reid Otsuji, "UC San Diego Ithaka S+R Research Study: Supporting Big Data Research," https://escholarship.org/uc/item/8kr7p2c0.

University of Colorado Boulder

Emily Dommermuth, Cindy Edgar, Nickoal Eichmann-Kalwara, Rebecca Kuglitsch, Andy Monaghan. Report forthcoming 2022.

University of Illinois, Urbana-Champaign

Carissa Phillips, Chris Wiley, Jen-Chien Yu, "Ithaka S+R Supporting Big Data Research University of Illinois at Champaign-Urbana Report."

University of Massachusetts, Amherst

Thea P. Atwood, Melanie Radik, Rebecca M. Seifried, "Supporting Big Data Research at the University of Massachusetts Amherst," https://scholarworks.umass.edu/libraries_working_papers/2/.

University of Oklahoma

Claire Curry, Zenobie S. Garrett, Mark Laufersweiler, Tyler Pearson, "OU Libraries Support of Big Data."

University of Rochester

Daniel Castillo, Moriana Garcia, Sara Pugachev, Sarah Siddiqui, "Supporting Big Data Research at the University of Rochester: An Ithaka S+R Local Report," https://urresearch.rochester.edu/institutionalPublicationPublicView.action?institutionalItemId=36200&versionNumber=1.

University of Virginia

Jacalyn Huband, Jennifer Huck, "Assessing the Research Practices of Big Data and Data Science Researchers at the University of Virginia: An Ithaka S+R Local Report," https://libraopen.lib.virginia.edu/public_view/mg74qm17c.

University of Wisconsin, Madison

Cameron Cook, Tom Durkin, Tobin Magle, Jennifer Patiño, "ITHAKA: Supporting Big Data Research, Data and Analysis from UW-Madison Researchers," https://minds.wisconsin.edu/handle/1793/82384.

Appendix 2: Supporting Big Data Research
Semi-Structured Interview Guide

Note regarding COVID-19 disruption. I want to start by acknowledging that research has been significantly disrupted in the past year due to the coronavirus pandemic. For any of the questions I'm about to ask, please feel free to answer with reference to your normal research practices, your research practices as adapted for the crisis situation, or both.

Introduction

Briefly describe the research project(s) you are currently working on.

- How does this research relate to the work typically done in your discipline?
- Give me a brief overview of the role that "big data" or data science methods play in your research.

Working with Data

Do you collect or generate your own data, or analyze secondary datasets?

If they collect or generate their own data Describe the process you go through to collect or generate data for your research.

- What challenges do you face in collecting or generating data for your research?

If they analyze secondary datasets How do you find and access data to use in your research? Examples: scraping the web, using APIs, using subscription databases

- What challenges do you face in finding data to use in your research?
- Once you've identified data you'd like to use, do you encounter any challenges in getting access to this data? *Examples: cost, format, terms of use, security restrictions*
- Does anyone help you find or access datasets? *Examples: librarian, research office staff, graduate student*

How do you analyze or model data in the course of your research?

- What software or computing infrastructure do you use? Examples: programming languages, high-performance computing, cloud computing
- What challenges do you face in analyzing or modeling data?
    - If you work with a research group or collaborators, how do you organize your data and/or code for collaboration?

- Do you take any security issues into consideration when deciding how to store and manage data and/or code in the course of your research?
    - Does anyone other than your research group members or collaborators help you analyze, model, store, or manage data? *Examples: statistics consulting service, research computing staff*

Are there any ethical concerns you or your colleagues face when working with data?

Research Communication

How do you disseminate your research findings and stay abreast of developments in your field? *Examples: articles, preprints, conferences, social media*

- Do you keep abreast of technological developments outside academia in order to inform your research? If so, how?
- Do you communicate your research findings to audiences outside academia? If so, how?
- What challenges do you face in disseminating your research and keeping up with your field?

Do you make your data or code available to other researchers (besides your collaborators or research group) after a project is completed? *Examples: uploading to a repository, publishing data papers, providing data upon request*

- What factors influenced your decision to make/not to make your data or code available?
- Have you received help or support from anyone in preparing your data or code to be shared with others? Why or why not?
- What, if any, incentives exist in your department or field for sharing data and/or code with others? *Examples: tenure evaluation, grant requirements, credit for data publications*

Training and Support

Have you received any training in working with big data? Examples: workshops, online tutorials, drop-in consultations

- What factors have influenced your decision to receive/not to receive training?
- If a colleague or graduate student needed to learn a new method or solve a difficult problem, where would you advise them to go for training or support?

Looking toward the future and considering evolving trends in your field, what types of training or support will be most beneficial to scholars in working with big data?

Wrapping Up

Is there anything else from your experiences or perspectives as a researcher, or on the topic of big data research more broadly, that I should know?