Supporting Big Data Research at Temple University

Temple University

Libraries

Table of Contents

Authors (in alphabetical order by last name):

Will Dean, Research and data services librarian, Temple University Libraries
ORCID: https://orcid.org/0000-0001-7871-9611
Fred Rowland, Librarian, Temple University Libraries
ORCID: https://orcid.org/0000-0003-2281-8797
Adam Shambaugh, Business Librarian, Temple University Libraries
ORCID: https://orcid.org/0000-0002-6685-3327
Gretchen Sneff, Science and engineering librarian, Temple University Libraries
ORCID: https://orcid.org/0000-0001-7621-5997

The authors would like to acknowledge the contributions of Nancy Bartman Turner, Associate director for organizational research & strategy alignment, Temple University Libraries, and Bernadette Mulvey, Executive director Information Technology Services: Applications technology services, Temple University. They attended meetings, read drafts, provided advice, and shared insights. The authors would also like to acknowledge Annie Johnson, Assistant Director for Open Publishing Initiatives and Scholarly Communications, for providing feedback on an earlier draft of this report.

Introduction

A team composed of Temple University Libraries (TUL) and Information Technology Services (ITS) staff participated in the multi-institution Supporting Big Data Research project to learn about the practices and support needs of big data researchers at Temple University. The project was coordinated by Ithaka S+R, an organization that works with academic and cultural communities. Both big data research and research using data science methods were in scope for this study. Big data research, according to Ithaka's definition (personal communication, April 14, 2020) is "conducted with data that is high in volume, velocity, and variety – that is, with large, diverse datasets that are often collected in real time." Data science methods, as described by Ithaka (personal communication, April 14, 2020) are methods "used by scholars across STEM, the social sciences, and the digital humanities, and may include machine learning/artificial intelligence, data/text mining, modeling, analytics, informatics, and other data science techniques." We recruited 14 researchers to participate in 45 to 60-minute semi-structured interviews, conducted in fall and winter, 2020-2021. (See Appendix A for sample invitation email and Appendix B for interview questions). Candidates could be faculty, postdoctoral or full-time researchers whose primary duties include research. Using the same questions, we interviewed researchers from computer and information science, mathematics, biology, chemistry, physics, neuroscience and public health.

> *We went from like 30,000 users to like over a million users and a great part of the spring and going into the summer was just trying to keep up with demand and really switching all of our research efforts towards trying to simulate a lot of the viral proteins and trying to understand how they function and trying to actually use [the distributed computing platform] to do virtual screening of potential drugs for one of the proteins, the main protease.*

Working with big data at Temple University frequently involves research teams with a variety of specializations and skill sets that span multiple departments, colleges, and institutions. Initially and throughout the research process, crucial decision points are encountered regarding such things as technology, staffing, and research design. While these challenges are not unique to big data research, the rapid development of hardware, software, and algorithms means that researchers may need to negotiate between their own initial research design and evolving tools, technologies, and methods. What we discovered bears little resemblance to what one might find at companies whose profit-driven business model revolves around big data, such as Google, Amazon, or Alibaba. Instead, we found passionate researchers who are working under challenging organizational and financial constraints to incorporate the explosive growth of data and computing technologies into scientific research.

> *To me, essentially, data is data and the problem becomes big data essentially when you – when you actually try to extract structure from the data that you see in an automated fashion. So, you use computer techniques to do that, rather than a human looking at it or something. That's usually what I think as when you say machine learning, or big data or something. It's not the amount of data that really matters. It can be a small data set and still use techniques to see its equal [underlying?] structure.*

With the aim of understanding the support needs of big data researchers at Temple, this report analyzes and evaluates the 14 interview transcripts for insights on the conduct of big data research. Our interviews revealed more insight on how researchers perform their work than on the kind of support they might need. As productive as this study was, it should be considered a preliminary assessment of the current support needs of big data researchers at Temple. Due to the distributed nature of support across campus, further study building on this report would be necessary for a comprehensive understanding of both the support offered and the support needed.

> *You have this massive amount of data stored. For each what we call an entry this is basically the equivalent to a line, a row, in an Excel sheet. But the columns are hundreds, thousands of variables. Thousands of variables: A, B, C, and now you have maybe several thousand entries in there. And the rows are your events. And you have billions of these events.*

The first and longest section, The Life Cycle of Big Data Research, describes the approach of our interviewees to the different stages of big data research. We examine their research practices for accessing, cleaning, and curating, analyzing, and sharing data, as well as the prevailing norms for publishing and communicating their findings. The following three sections – Collaboration, Infrastructure, and Training – highlight general issues that we found most big data research projects encounter. Our interviewees discuss the ways that these are dealt with. Finally, we end with Discussion and Recommendations sections. We found, among other things, that our interviewees often struggle to keep up with the rapid pace of change, work in teams to assemble the necessary complement of project skills, value access to open data and embrace data sharing, and rely on both internal university and external resources and data.

> *It's really trendy right now to say let's forget about all [hypothesis-based science] of this and just do deep learning, but it's – I'm not sure that that's really a better way to go. We tend to be more focused on testing mechanistic hypotheses and doing hypothesis-based science. So that's a distinction from a lot of the way that most people who would fall into big data categories do science vs. how we do science.*

The Life Cycle of Big Data Research

In the life cycle of big data, researchers access data, clean and curate their data, perform analysis, publish their findings, and share their data with the larger research community. These sections illustrate how this process works for our Temple interviewees.

Accessing Data

The researchers we interviewed accessed and used data in a variety of ways. Many emphasized that having access to open and comprehensive data was important to doing their work. Some generated their own data and some used data combined from different sources. All were aware of security and privacy issues relating to data.

We found that researchers are aware of good data sources in their field, having absorbed this knowledge ("I just know them") through professional networks and previous research projects. Increasingly they find datasets published along with research papers or learn about datasets and find they can download them from research group websites. "…you see papers or see a talk or something hey, there's this new data set and so on, and usually, there's a website that's associated with the project."

A lot of data found online is open data, available for anyone to use without restrictions. Most of the researchers we spoke with expressed enthusiasm for, "the incredible data sets that are public," that facilitate scientific research today. Researchers from computer science and mathematics who develop new algorithms and methods were not tied to any datasets and reported that they simply require some good data. As one explained, "we just go to internet and try to find something." They often use Google or browse software repositories like GitHub. A computer scientist noted that having access to online data means that "the majority of people working in machine learning, data mining, big data, computer science, in these areas, they don't have to generate their own data," since, "they can just reach out to the internet." A mathematician was emphatic about using open data because it's "important to have publicly available data out there. So, I tend to sometimes even select research projects based on what data I use. So, what are good data that are out there? I want to support that." Some researchers consider it a violation of community norms to create unnecessary barriers to access to data that should be open. One researcher pointed out that public data opens research to those without traditional institutional support and disapproved of paying for data:

> There are some data sets where the – it was produced on Federal funds, but then you have to pay money to get the data, and I usually say no, because I have a fundamental objection to such approaches... because I think it's just wrong; if somebody has a project that is funded by NSF or something, and then they force people to pay…that's just fundamentally wrong.

Though researchers were enthusiastic about open data, we also heard about areas where researchers found barriers to accessing data. The data stores of some research areas are less organized than others. A biologist explains that, in contrast to genomics, if you want "the phenotypic ecological behavioral traits of a particular species" you are likely to find "that those have not been organized into a central database," a situation they attribute to the distinctive characteristics of the ecology research community. Privacy considerations also limit access to data in such areas as health and education. A biologist familiar with this area detailed how homomorphic encryption, a technique for analyzing encrypted data, may enable use of this protected data. Private companies hold vast stores of data that are proprietary, and the motivations of profit-seeking companies differ in important ways from those of scientific researchers. A biologist reflects on privately-held genomic data, "You know 23andme data I would love to have, but is it available? No. No way. Because that's their business model. So, there's a lot of stuff that's just not available.

The researchers who preferred reusing existing data cited cost or convenience, for their preference. A biologist engaged in a wide range of research projects explained that "I used to generate my own data; then I just found out it was easier to look at other people's data." Data reuse does not only apply to reusing open data. A neuroscientist described a "huge push" to increase the use of the electronic health records system, EPIC, in their department's research.

Reusing existing data—whether open or restricted data--can be contrasted with generating data. Generating data can be very costly in terms of both time and money. A physicist, for instance, collaborates on experiments using colossal energy-intensive particle accelerators. While generating and capturing enormous volumes of data, these experimental apparatuses are funded by national governments because they are so expensive to build, operate, and maintain.

In some cases, researchers combined data from different sources. Researchers in neuroscience, biology, and public health described combining genomic with clinical and demographic data to find correlations between genetic expressions and disease. A public health researcher described an iterative process of collaborating with laboratory scientists performing basic research, demonstrating the cross-referencing of public and lab-generated data:

> So, my data basically have two parts. One is publicly available data…we have the GEO datasets which has massive expression data available. Then the other part is my collaborators' in-house data…we combine the public data and their private data to [perform an] interrogative analysis. So, the result that we present to our collaborator… [is that] a few biomarkers are really interesting. So they go back to the lab and do more – more data generation…to validate our result.

Our interviewees expressed concern about privacy and security but were confident in the security of their systems. Following proper procedures was deemed sufficient for handling most challenges. Handling confidential data and controlling access to computer systems were two areas of concern. The protection of confidential and personal health information (PHI) data is addressed by laws and regulations and there are established compliance procedures. Researchers mentioned HIPAA guidelines and working with the IT department to "make sure that what we are doing is according to their protocols." Security measures are necessary to prevent malicious actors from damaging research computer systems, but one researcher explained that there are limits to how far they will go to secure their data:

> We don't take any such precautions with security beyond what's forced upon us by the national laboratory. We don't want our data destroyed but we could care less if someone thinks they want to take our hundred terabytes of neutrino oscillation data and try and make sense of it.

Security issues are handled through user access controls and standard IT security measures: "you have to have the user ID at the national lab to have access to that. But once you have that, then it's all public, you know."

Cleaning and Curating

After data has been captured, it must be cleaned and curated before it can be analyzed. This can be the most labor-intensive stage, as the following researcher explains:

> We can spend, I suppose, more than half of our time facing those challenges, dealing with the data. Converting, getting the data, processing it, figuring out what you're looking at, what are the issues, cleaning it. So I suppose yeah, 50 percent, maybe even more of our time is really spent wrestling with the data.

Our interviewees often referred to this as turning raw data into processed data, which can involve converting measurements from one standard to another, removing extraneous or unnecessary characters, and transforming data from an unstructured series of numbers into a spreadsheet. A biologist talked about the "cost of thinking" about cleaning data as an expensive part of the research process, and another mentioned the care needed to handle large datasets "to single out only what we're interested in."

Curating data requires knowledge about the dataset as well as the context within which it was collected in order “to filter early on the good stuff from the bad stuff.” Filtering data is time-consuming and introduces a high degree of uncertainty since the importance of specific data elements may not emerge until a later stage of analysis. This makes data curation a highly iterative process whose burden falls directly on the research team. A researcher explained that a team could spend six months curating data “unless you discover you made some stupid mistake the first time around,” which might necessitate starting over again. Finally, if one wants to effectively share data with other researchers, metadata descriptions must be applied. A mathematician, who frequently uses datasets from open repositories, provided some insight on the time-consuming activity of cleaning and curating by reflecting on data quality:

> Big data isn’t really what we are after. We’re after good data…I think the next generation initiative might be good data, where - slightly smaller, but good data is going to make much more difference than huge amount of you know, semi-good data. And I think that’s where there is going to be some tug of war: what is better? And I personally have much less good data, than huge amount of bad data.

Analyzing

After capturing, cleaning, and curating huge volumes of data, the purpose of a project is to resolve the data into a form that allows data science methodologies to address an original research problem, whether this involves answering a question, validating a model, or illustrating correspondences. The particle physicists we interviewed provided explanations of this phase that differed significantly from the life scientists. The physicists spoke clearly about the collaborative process between team members that resulted in the analysis of data. We learned less about the technical details of analysis. The life scientists, on the other hand, focused more heavily on the data science methods employed to analyze the data. We learned less about the collaboration between team members. It appears that the capital-intensive nature of particle physics and the far-flung collaborations involved have generated a highly organized research area with shared norms, tools, and equipment. The life scientists described a far more decentralized environment in which new tools, methods, and approaches are rapidly emerging. Researchers in computer science and mathematics also analyzed data but mainly for the purpose of developing new methods and tools. A computer scientist described themself as a "research enabler," which aptly describes one role these specialists play, creating mathematical models and efficient algorithms that other scientists deploy.

The particle physicists called themselves "experimentalists" and described the intricate detection equipment they collaboratively design to capture particle data, in one example, "…250 or so individual detector components are read out at gigahertz rates, and we form this train of data." This raw data is continually filtered to converge on results of computer simulations performed before the experiments are run. The raw data is processed over what might be one or two years before it is ready to be analyzed. Shared tools are developed for analysis. C++ and Python were mentioned as programming languages used in this late-stage analysis, which is usually carried out by graduate students and post-docs under the direction of the PI or advisor, "either on a small desktop terabyte drive or maybe something on a computer storage facility nearby." The students and post-docs appear to have a wide latitude as to how to perform their analysis, but it is scrutinized by the research team. Despite its complexity, the physicists expressed very little frustration with the tools and infrastructure that supports their work.

Of the biologists we interviewed, one, who self-describes as a biological physicist, uses data mining methods to study the relationship between nucleotide sequences and protein structures, with the goal of developing a predictive model of protein structure. They referred to this as "the problem of folding" and were eager to share that Google's DeepMind subsidiary had recently demonstrated the AlphaFold algorithm, which "reached an accuracy and reliability that makes many people now say that the problem is solved." This same researcher also develops rule-based biochemical models to simulate molecular configurations, which has little to do with data mining. Another biologist described a project using electronic health records to discover disease comorbidities that suggests data mining and machine learning methods, though neither of those terms was used. A few more biologists do not use data mining or machine learning (or the associated technique of deep learning). Instead, one describes their work as comparative genomics and is "more focused on testing mechanistic hypotheses and doing hypothesis-based science. So that's a distinction from a lot of the way that most people who would fall into big data categories do science vs. how we do science." One more biologist develops computational tools and methods for studying viruses and isolating drug targets.

The methods of the remaining life scientists were similar to those of the biologists. A chemist described their work as computational chemistry and, focused “mostly [on] molecular simulation of proteins and other types of biomolecules.” This researcher’s group is doing basic research with possible future applications such as “virtual screening of potential drugs for one of the proteins, the main protease.” A neuroscientist and a public health researcher described what appeared to be projects using data mining or machine learning to discover relationships between various diseases or injuries – HIV/AIDS, concussions, diabetes, Alzheimer’s – and genetic variations.

Programming languages used by life science researchers, physicists, and those engaged in computer science research varied based on the disciplinary differences and specific research projects. There were, however, some recurring tools used for data analysis. A biologist noted, “I tell students that they have to learn...the holy trilogy, which is R, Python, and Linux.” And while our small group of interviewees were not all as prescriptive about programming languages, they did indicate an outsized popularity of these three tools, particularly Python, which was mentioned by all but one respondent as a language that they or their collaborators use.

As open languages like Python and R have become more popular, multiple researchers described wanting to move away from older languages such as FORTRAN and MATLAB. However, adoption of new languages and tools does present a challenge as a biologist remarked, “I've been trying to get my group to – just because of public availability issues, to do more things in C++ and less in MATLAB. But some of my students just really like working in MATLAB, and we also have legacy issues of some import code that has now been written by a former post doc in MATLAB that people don't want to rewrite.” As preference for and prevalence of certain tools changes, legacy issues are likely to persist in disciplines that make heavy use of data science.

While respondents from across disciplines mentioned many of the same tools and technologies for writing code, sharing data, and analyzing findings, several individuals mentioned the use of highly specialized tools that are specific to their own disciplines or research needs. A neuroscientist who studies disease and neurodegeneration mentioned the use of PANTHER, a gene and protein database that is part of the larger Gene Ontology initiative. A biologist mentioned data from the National Center for Biotechnology Information (NCBI) and the Protein Data Bank. Elsewhere, a computer scientist noted coding platforms that are central to their work in graph neural networks, specifically PyTorch and TensorFlow, which are tools developed to aid deep learning research. Although these discipline-specific tools may not have the same far-reaching implications for support to big data research, it is important for subject librarians and other individuals in research support roles to be aware of these tools as the need to support big data initiatives increases.

Publishing

Despite the evolving landscape of scholarly communication and research dissemination, interviewees' responses to questions about scholarship demonstrated the enduring nature of traditional approaches to academic research. Researchers mentioned publication in academic journals and conference presentations as the primary venues for sharing their research: "We publish. Every time we finish a study, we publish" and "Yeah, we publish papers. So either conference or journal papers." However, several researchers did acknowledge the need for new approaches to scholarly communication, which accelerated during the COVID-19 pandemic. Life science researchers referred to the emerging scholarly practices of depositing a draft of a work-in-progress to preprint servers. A biologist acknowledged that peer-reviewed journal publication is "still the gold standard for communicating science" but noted that the unique circumstances surrounding COVID-19 have spotlighted the need for new approaches:

> As far as specifically in relation to Covid and SARS-CoV-2, preprints and preprint servers have taken on a, you know, a much more prominent role simply because – I mean we have a typical review cycle for a journal might be three to four months, right? And if you are trying to communicate science that is relevant right then, you cannot wait for the process to go through.

While research on the Coronavirus pandemic is an extraordinary example, there are other areas in which the slow pace of the peer-review process stands at odds with time-sensitive research. Preprint servers also provide a unique opportunity to receive feedback from other scholars before research undergoes a formal peer review process. A biologist noted the appeal of the preprint server bioRxiv:

> You got something out good, throw it to bioRxiv and people will get to look at it, they get to comment on it, which is fantastic, right? So, if people see something that they're really interested in they'll look at it and go oh that's cool, but you should have done this. Okay, we'll do that.

While such platforms ameliorate the slow pace of scholarly journal publication, they also raise concerns about quality and scholarly rigor. A biologist noted, “I don’t think anybody knows what the right balance is between, you know, speed with which important findings are disseminated, versus quality control. I don’t think the system moderates itself very well, not yet.”

Sharing

Sharing data and computer code are key components of information dissemination in data science fields. All our interviewees engaged in this behavior, although one biologist expressed reluctance to share code. This is likely due to the central role of writing code for this researcher, who expressed a willingness to share code after a project is completed. Researchers talked about different platforms for sharing data and code and expressed a variety of reasons for sharing.

When discussing data sharing, it was often difficult to discern whether our interviewees were referring to sharing data or code. They primarily shared code on GitHub but several researchers also shared data on this platform, though there are size limits that restrict its utility for this function. Many of our interviewees expressed uncertainty about the best ways of sharing, as this chemist notes, “code is different than data, right? Code is smaller and you can share it like on GitHub. That’s something I'm still learning about too, like the best ways of doing that.” Depending on the stage of research, the volume of data being shared can range from the genuinely huge to the relatively small. This same chemist also shares data on Zenodo, often breaking up datasets into multiple deposits because of size constraints. Similarly, an interviewee from the biology department said the challenge of sharing large volumes of data meant that instead of posting a download link, they make a declaration of data availability, which is stored, “in some server and can be given…access to upon request.” Another respondent explains how his research group utilizes Zenodo to organize and share its work:

> We utilize a facility at CERN called Zenodo. That allows us to automatically, whenever we make a release, we can flag it and it will be picked up on CERN and then made available with the DOI. One for the project as a whole and one for each release, so that it's easier citable. And of course, they keep a copy of those snapshots.

Researchers mentioned using the Figshare and Dryad repositories, often in the context of meeting journal publication requirements.

Reasons for sharing included the requirements of journals, funders, and professional associations and more philosophical motivations relating to ethics, a desire for transparency, and a commitment to open science. A chemist remarked, “It’s now becoming a requirement and something that we need to do,” even though they are not given specific training or resources to facilitate sharing. This ‘unfunded mandate’ aspect of data sharing was echoed by other researchers but did not dim researchers’ enthusiasm for sharing data. The same chemist described projects where code and data are openly shared as “something we aspire to,” because it allows for more collaboration and a chance for others to fix problems and build new solutions. While many researchers shared a similarly pragmatic view of the value of sharing data, others saw data sharing as having broader philosophical and societal benefits, speaking of a “social contract” or the responsibility of professors to “share with the world.”

Collaboration

A neuroscientist described their typical process of building a research team, starting with thinking about the research question, identifying “gaps in knowledge” and calculating “if we could this…maybe we could get somebody who could do that kind of thing." Commenting on their projects, they explain that “the bottom line is it’s not hard if you have the right team…” The mathematician and computer scientists we interviewed collaborate extensively with laboratory and field scientists for whom they develop fast and efficient algorithms and numeric models. A computer scientist who participates in brain research using machine learning emphasizes that “you have to [have] very close and deep interactions, and iterated interactions between say doctor studying cancer and computer scientist.” This same researcher explains "it is really important for the domain expert to collect data, annotate the data. And then present it to their machine learning collaborators to work on.” A public health researcher working on COVID-19 data stated, “I found out one of the problem[s] is coordination of all different parties. There’s repetitive information, or they’re lacking information...”

The particle physicists we spoke with described teams that can vary in size, “by small I mean have less than typically five to ten, or maybe twenty collaborators. Some are much larger. It might involve hundreds of collaborators." But they described less interdisciplinarity on their teams because the tools that particle physicists use are very specialized. Even most computer scientists at the national laboratories where they conduct research, one physicist said, will have PhDs in physics.

Students on the research team

Research groups at Temple often include graduate and undergraduate students. A computer science professor explains that they have "a team of, I don't know, usually 10 to 20 students who work with me." Students may lack pertinent training when they join a research group, and their departure may result in a loss of information or skills for the team. Student turnover is one challenge of doing research while training the next generation of researchers and necessitates a unique kind of coordination. A chemist described how it works when the challenge is interdisciplinarity:

> …our work is so cross disciplinary. It's biology, it's physics, it's chemistry, it's computer science, it's applied math in a lot of cases; it's everything, so you know, students that come into my lab come from all different places and they kind of just pick up what they need to know.

A mathematician shared that since they joined Temple "there have been huge strides--and I think it's one of the success stories--to getting really good students. And also producing really good students." Here they explain how this works:

> it's very easy to tie them in, because there's a lot of tasks that we don't necessarily need to understand all the complicated mathematics, but you can say okay, this is the bigger problem; you'll pick up more on group meetings on this, but for now, this is your problem.

However, a biologist expressed concern that so much analysis is done by postdocs and graduate students (and sometimes undergraduates) who "are all like wet behind the ear, green to the stuff," and do not have a lot of experience, but also remarked that, "and you know, yes, you could easily train these students... but at the end of the day…it's basically data that's like, very expensive that's being managed by undergraduates."

Support from outside the research team

Researchers may seek support from sources outside the team. Vendors, including cloud computing providers, provide training and support. National laboratories provide training on cybersecurity. Basic data analysis can be obtained from a sequencing company if a researcher does not have a computational background. Data center and the Owl's Nest—Temple's shared high-performance computing cluster—staff provide help. A neuroscientist who studies neurodegeneration works closely with a biostatistician to analyze data. They describe a research project looking at six different genes and hundreds of different patients, whose records contain both clinical and demographic data, "it's just unending, really, how many different variables have to come into play." Working with a biostatistician is necessary. "[Our biostatistician] is absolutely amazing and he helps me – well, he does all the work. All the hard work." However, while the neuroscientist is focused on specific genes and mechanisms, the biostatistician's approach is different, and it can be challenging to communicate across disciplines:

> ...sometimes it's difficult to bridge the gap between what they do every day and what I do every day and say, 'No, this is what I need. No, this is what I need.' And then when I try to explain it, they speak a different language...

Another researcher echoed the divide that exists when they want to use data science methods in their work "since there's not typically a need in academia right, for big data, and often it's kind of this corporate thing that requires software and data engineers to know very specialized technologies, it's hard to interface those two groups." A biologist opined that "some groups probably have a problem where their colleges don't have sufficient research informatic support." They provided this comparison:

> This chilled water coming through the taps; gas going through the burners; and if you think of all of those things, none of that kind of infrastructure support exists for informatics research, in traditional research environments... We still pay the same overhead of 50-some percent, but you had zero support. And nobody understands that…

Some of these issues may be growing pains as the university learns how to support big data research. The issues raised also present opportunities to forge new connections and develop needed services.

Infrastructure

A robust infrastructure of computing and storage facilities is necessary to support big data. Our interviewees found a wide range of local, national, and international solutions to their infrastructure needs. The Owl's Nest is Temple University's shared high performance and scientific computing cluster. In addition to this, there are numerous smaller high-performance computing (HPC) resources operated by individual research groups on campus. The extent to which respondents use the Owl's Nest varies considerably. A mathematician remarked that, "oftentimes many of the projects we have, we make use of Temple's HPC cluster, so…that definitely plays an important role... it couldn't be done without it." And a chemist describes their laboratory as "extensive users of the Owl's Nest cluster." Conversely, a computer scientist mentioned the Owl's Nest, but only as an afterthought: "then there's Owl's Nest, which we rarely use, because it's just like – it's a hassle to program it, and installing programs, because it is a very different kind of system." Of the 14 researchers interviewed for the Big Data project, only half mentioned the Owl's Nest, and only three indicated that it played an important role in their research. However, a computer scientist closely associated with the Owl's Nest explained that "typical HPC facilities on campus are basically always fully used, or nearly fully used," so it was difficult to gain an accurate understanding of the Owl's Nest.

Several researchers said that they rely heavily on external academic and government-funded organizations. Others take advantage of private sector cloud computing. The physicists discussed their use of infrastructure at Oak Ridge, Brookhaven, Jefferson Labs, and CERN (in Geneva, Switzerland). One physicist explained that "we never store the data here at Temple. Don't need to. National laboratories have those resources, and we make use of them." A chemist who reported extensive use of the Owl's Nest is affiliated with high performance computing centers at other institutions, and one computer scientist associated with the Owl's Nest meets annually with representatives from other universities. A biologist explains that he has "scientific developers" who write code but that all the tedious stuff is off-loaded on private corporations, "You basically shift all the development of these tedious things to big corporations like, you know, Google, Apple, and consortia that make sure that all your web browsers run superfast, and support everything, so you just piggyback on this technology." This researcher also uses Amazon's cold storage offering, Glacier, as a cheaper option to other cloud storage solutions, but notes that this increases the technical burden when accessing the data and that the company could stop offering discounts for public service research in the future. A chemist intensively involved with COVID-19 research explained that Amazon "has really helped a lot in hosting cloud infrastructure for this data" but worries about potential future costs and the learning curve for leveraging this cloud infrastructure.

Many of our interviewees expressed anxiety and frustration that resources required for big data research would outstrip the ability of the university to keep up, either with other large, more well-endowed universities or the private sector. One spoke of the “haves and have nots.” Summarizing the storage choices that our interviewees make is difficult because so many appear to be ad hoc. As one researcher bluntly put it, “Where on earth can we put all this stuff?” adding that the choice of storage is critical before even starting a project. A computer scientist worries that Temple is “not geared up to be an institution where we can truly deal with the real data of the future.” A biologist explained that “molecular simulations tend to generate an enormous volume of data. So, it’s easy to fill up terabytes just with one project.” A computer scientist explained that when they began their career, “even ten megabytes was considered a large hard drive” and now “our biggest storage facility on the large HBC cluster has 1.5 Petabyte [~1,000 Terabytes] actual useable.” However, at the Owl’s Nest this same researcher explains that “there's no way that we can afford to have a suitable backup system to back up all the stuff that is written on multiple petabytes of storage…if our hard drives all fail on the same moment, all data is gone.”

Training

Few researchers we spoke with received any formal training in big data. According to one researcher, “You just kind of learn what you need to know, to do what you need to do.” Three other researchers explained that learning often happens while engaged in research projects “on demand,” “by the seat of my pants,” or by “having colleagues who were computer scientists or statisticians or had bits of understanding of how to do things that you didn't have at the time and just learning from them.” Another interviewee explained that “I don’t train in anything; I just read up.” Sources for learning included books, manuals or papers; seeking information on the internet; and learning from colleagues, collaborators, and experts in a field. Less commonly researchers said that they take advantage of workshops, conferences, or tutorials. “I’m not too proud to get my hands dirty, still, so I participate when there’s a good mini-tutorial or a conference or something.” Two others highlighted the importance of one’s professional community as a source for training, “It is about the community, research community. So that’s actually what I…consider a go-to place . . . They have their own activities, including conferences, workshops.”

Researchers found it challenging to stay current with big data methods and tools due to the field's explosive growth. Coping with the tools necessary for storing and sharing data can present a major burden. A chemist noted that "you can spend a year taking a course on just how to use all of Amazon's cloud computing and cloud data storage" and a biologist explained that "we try to limit the number of tools that we use because it just means you have to learn more." In addition, there is a fundamental need to understand developments in data science methodologies, as this biologist explains:

> I'm trying to keep up in general with sort of major developments in statistics and machine learning, even if it's not in my field, just to have a sense of how this is going to apply to what we do. But the volume of information is just overwhelming.

Another researcher summarized:

> Big data is a dynamic, changing field. What we are discussing today potentially will be new topics tomorrow. It's a learning process. We never say, 'I know big data." I'm learning big data every day.'

The complexity of the workflow and collaborative environment means that researchers are often training and advising students and colleagues. The specialized nature of working with big data can make staff departures particularly painful. Specialized knowledge can easily leave with students and post-docs as this chemist explains, "does that knowledge die once you leave the lab? Can you tell the rest of us how you learned this technique?" A computer scientist makes the same point, "Sometimes if a person leaves who used to work for us in a specific area, we just don't want to do the work in that area again, because that is too much. To have the next person come and get up to speed in six months – most NIH and NSF grants are three years." This same researcher, however, is satisfied with the training their group provides, "we have a good system of training young people or new people coming in, because…when you have a large enough group, you have dynamics that allows people to know things from others."

Faculty emphasized their personal role in training students by answering questions, directing and instructing students *in situ*, and providing pointers to resources, papers or experts to consult. Researchers refer students and colleagues to the same sources of self-learning that they themselves use such as books, workshops and tutorials. In addition, our interviewees recommended that students take advantage of the learning resources of third-party tool providers, vendors, online course platforms (Codeacademy) and question/answer forums (Stack Overflow). There was a special emphasis on professional communities:

> So, whatever is there within the professional organizations, there's usually online resources; video tutorials and so on that guide you through, and then regularly, you just get these announcements, emails, mini workshops on big data…and so on, and then students can sign up for this, for either nominal fee or free, and sometimes you also can organize it of course with colleagues on campus.

Students learn within research groups from more experienced students and post-doctoral fellows. There are also important interactions between research groups, as one researcher surmised, such as between graduate students in bioinformatics and computational biology groups. Another researcher described how colleagues work to encourage students to train each other, "can your student show mine how to use TensorFlow?"

At a more formal level, faculty recommend Temple courses on big data and data science, organize summer data workshops for graduate students, and steer students toward toward the HPC training that the Owl's Nest offers. There is also enthusiasm for a new data science course requirement being developed for College of Science & Technology (CST) students, based on the Data 8 curriculum at UC Berkeley. There was a general agreement on the broader need for a stronger Temple focus on data literacy and data skills training. Even at the graduate level, a physicist is dismayed over poor data skills, "I'm amazed by the lack of ability of a graduate student knowing how to at the very basic level dealing with data," and commented that students need training in, "very basic aspects of statistics, of statistical analysis," and producing visualizations. A second physicist is concerned that in the first two years of graduate work there is "not a single course on data handling [or] software development." Despite these challenges, one researcher expressed optimism, "Big data is the way to go, and I think at Temple, we're actually doing that right now, at least with the curriculum, we're definitely getting there."

Discussion

Big data requires a host of services and shared community commitments to sustain it, as researchers rely on a wide range of data, tools, methods, and infrastructure. Institutional support comes from Temple as well as national and international consortiums. Most of our interviewees worked with colleagues within and outside their disciplines and while these collaborations were generally positive, communication between specialists in a discipline and computer scientists or biostatisticians with whom they collaborate can be challenging. Staying current in big data research is a challenge for faculty. Undergraduates and graduate students just entering the field face a stiff learning curve. Researchers spoke about the need for student training in formal courses or modules associated with degree programs and through collaborations between departments and colleges.

All the researchers in our study cited storage as an important need and many described problems they have securing enough storage space for their data. On the computing side, many researchers did not use Temple's shared HPC system, the Owl's Nest, and some who did, found it difficult to use. Researchers expressed concern when a discipline lacked a central data repository, when public data was insufficiently cleaned or lacked the necessary metadata descriptions. Data preparation and curation are time- and labor-intensive activities. Properly curated data are essential for research and data that are not properly curated cannot be reused by others. Additional funding for tools that process large datasets, preprocessing datasets in repositories, and improved training might reduce this time commitment. Publishing papers in respected journals is still the dominant way that researchers share work and establish their careers.

Recommendations

Based on our conversations with researchers, we recommend the following actions to support big data and data science research at Temple. Generally, these recommendations would require cooperation across multiple campus units. We have noted those areas where the library can potentially offer support.

Use the library's unique position as a center for engagement and collaboration at the university to facilitate communication between Temple researchers doing big data and data science research. Researchers reported difficulty keeping up with new developments and shared that they often learned informally from colleagues and others in their research communities. Provide opportunities for big data and data science researchers across the University to share ideas by sponsoring lectures, seminars, and other activities at the library. To keep abreast of the changing needs of these researchers, create a liaison position to the big data/data science community that is filled by a library staff member with a background in research and data science.

Identify a campus research solutions consultant. Researchers reported that the technical and administrative needs of this kind of research were becoming more complex, and finding optimal solutions were burdensome. We recommend identifying a person, and a coordinating team, tasked with understanding the research infrastructure across all the schools and colleges at Temple who will provide advice and support to big data researchers. The consultant and team can recommend appropriate university or external services and technologies, identify the need for infrastructure upgrades, and develop best practices on the use of cloud services (internal or external), computer code and data management. This team could maintain a centralized information hub of university informatics and statistical services available to researchers.

Promote campus solutions for high performance computing and data storage. Some researchers reported finding solutions for computing and storage issues burdensome. We recommend promoting existing resources and processes that are Information Technology Services (ITS) and Office of the Vice President for Research (OVPR) sanctioned, such as the Owl's Nest and cloud storage options, to increase awareness and appropriate use for a cost effective and secure environment.

Offer more data skills training. While initiatives such as required data courses in CST for undergraduates, other courses within academic programs, and summer workshops offered to graduate students were positively received, some researchers noted that students need more training on topics including data literacy, statistical analysis and computing technology. Some suggested academic departments should expand course offerings on these topics for graduate students. Where possible, the Library should expand its offerings to address unmet training needs, building on coding and data visualization workshops in the Duckworth Scholars Studio and data management workshops by the Research Data Services. Providing training or support for statistical and data analysis may require that the library add staff with those skills. Researchers we spoke with valued open data and shared their data and/or code. The Library should continue to promote and teach open practices to ensure all researchers, including undergraduate and graduate students, have ample opportunities to learn about best practices and tools for data sharing.

Appendix A. Sample recruitment email

Subject. Temple University Library's study on supporting big data research

Dear [*name of researcher*],

Temple University Library is conducting a study on the practices of researchers who use big data or data science methods in order to improve support services for their work. Would you be willing to participate in a one-hour interview to share your unique experiences and perspective?

Our local Temple University study is part of a suite of parallel studies at 20 other institutions of higher education in the US, coordinated by Ithaka S+R, a not-for-profit research and consulting service. The information gathered at Temple University will also be included in a landmark capstone report by Ithaka S+R and will be essential for Temple to further understand how the support needs of big data/data science researchers are evolving more broadly.

If you have any questions about the study, please don't hesitate to reach out. Thank you so much for your consideration.

Sincerely,

[*name of investigator listed on this protocol*]

Appendix B. Interview questions

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 at your institution or in your 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?