Supporting Big Data Research at New York University

Local Report Prepared by Members of NYU Division of Libraries in Conjunction with Ithaka S+R

October 18, 2021

http://hdl.handle.net/2451/63363

Prepared by:
Katie Wissel, Data Librarian
Margaret Smith, Head of Science Services
Nicholas Wolf, Research Data Management Librarian
Vicky Rampin, Librarian for Research Data Management and Reproducibility

I. Executive Summary

Research reliant on big data is being conducted at New York University in a variety of disciplines, including those associated with the social sciences, sciences, humanities, and professional fields. Characteristics of big-data research include but are not limited to a unique dependence on secondary data licensing and acquisition (often sensor- or instrument-based captures, or else arising from social-media or web-based sources), a need for deeper computational training by researchers than has typically been the case in the past, and higher expenses for conducting research because of data acquisition costs and the requirement for advanced personnel expertise. The landscape of sources for big datasets is marked by a lack of findability and documentation for reuse, a feature that data-archiving institutions can mitigate through preservation and data-advocacy efforts. However, such institutions will need to build more expansive infrastructure to store and distribute that data. The field sees the need for responsive and deep but not software-dependent training, particularly skills acquisition in data visualization, data and database management, and high-performance computing. Researchers see a need for greater support in navigating data licensing and permissions surrounding machine-assisted access to previously licensed data-provider platforms.

II. Introduction

New York University Division of Libraries participated in the Ithaka S+R "Supporting Big Data Research" initiative in fall 2020 and spring 2021, one of ten universities to contribute to this project. For the initiative, each local university team, consisting of library and technology professionals, conducted interviews and collaboratively assessed the needs of researchers engaged in work with big data. The NYU team completed interviews with eight researchers from a range of professional appointments, including faculty members and research engineers, and both junior and senior researchers. Interviews were all conducted remotely, recorded to enable transcription, transcribed, and coded using the qualitative analysis software Taguette, and the findings used to build the report that follows.

A. Institutional Overview

There are a number of characteristics of New York University that have the potential to shape the place of big-data projects among its researchers. It is among the largest private universities in the United States, with over 60,000 students enrolled and employing over 5,000 faculty and researchers. This size and funding status situates it among the country's community of R1 research institutions while also conferring on it the financial independence ($1 billion in research per year) as well as the challenges that come with being a private university in today's higher-education environment.

Its growth, however, has been most pronounced in recent decades, meaning that many of its features—including its unusually large number of distinct majors and tendency to be organized around independent schools and research centers—bear the imprints of higher-education characteristics of the past few decades (and especially the 1990s and after). In addition to the prevalence of schools and research centers, which bring with it the tendency for a single discipline (including big-data-oriented research) to be practiced independently in multiple departments, NYU bears the features of a trend in recent decades toward global campuses. This latter feature is unusually marked at NYU, which has three "portal" campuses (New York City, Shanghai, Abu Dhabi) designed to provide a seamless student experience conducted across all locations, as well as twelve global sites across Europe, Australia, South America, North America, and Asia geared more toward the study-abroad experience. Its emphasis on enabling all NYU students to study at these global locations at some point in their degree pursuit without impacting core degree requirements means that instruction—and with it, the researchers and their university-wide collaborative connections—is also conducted across multiple countries as a matter of course.

The use of big data in research is not confined to any single discipline, and if anything, it is pushing strongly into new disciplines; this makes it difficult to connect it with any particular department or school. NYU's tendency toward a more decentralized school and research-center format reinforces this diffuse concentration of big-data work. Big data, in the end, thus has its practitioners at NYU in potentially any corner of the university. At the same time, NYU has programs and centers that have historically had closer connections to big-data projects. These include the Courant Institute of Mathematical Science (f. 1935), the Center for Data Science (f. 2013), and the Tandon School of Engineering (f. In 1853 as Brooklyn Polytechnic but aligned and later incorporated into NYU in 2008), all of which have active degree programs and/or research being conducted in data science. Other relevant centers for big data work include the Center for Social Media and Politics, Center for Neural Science, and the Institute for the Study of the Ancient World, which has research initiatives around 3D imaging and linked data.

NYU also supports a centralized High Performance Computing cluster (newly refreshed with the acquisition of a new Lenovo-built supercomputer, Greene, launched in 2020), run by its central IT's Research Technology group. Not unexpectedly, users of such systems tend to form a community of practice around big-data research, and the presence of such a system on campus (as opposed to the use of a non-NYU system by NYU researchers) contributes to the potential for cross-disciplinary discussions among users of that system. The central HPC team is not the only place for big-data computation to occur, however, as separate schools and colleges (for example, the NYU School of Medicine) maintain their own high performance computing systems of various sizes; NYU researchers also work with nearby centers such as the Simons Foundation that maintain their own computing infrastructure that can support big-data work, and some faculty use the computational facilities at collaborators' institutions.

B. Methodology

The team recruited participants by distributing a call for interviewees to a list of potential candidates that were known to be working with big data (as a result of their interaction with NYU Libraries or IT services) or were referred to the team through another researcher. The call was distributed in multiple rounds until sufficient numbers of participants could be secured. The resulting pool included eight researchers, five men and three women, from the fields of data and computer science, environmental science, archaeology, physics, engineering, and chemistry. Of these participants, six had faculty appointments consisting of a mix of ranks (assistant through full professors), and two had appointments as full-time researchers.

The interviews were conducted based on a standardized script created by Ithaka S+R and done via Zoom. Audio recordings of the sessions were transcribed and anonymized. The team used an iterative, open-coding method in which the transcripts were first read through by the team, preliminary codes separately created by each team member, and a final set of codes assembled by collating and merging the separate code schemes. The final markup was conducted such that each transcript was encoded by two team members through a first and then second pass.

C. Study Limitations

Recruitment was by far the most significant hurdle facing the project. The pandemic brought with it drastically reduced time for researchers to participate in interviews; many expressed concern at their ability to conduct their core research during this time, let alone take on additional time to participate in an interview. Fatigue and burnout in the late fall of 2020 further curtailed recruitment, though it rebounded enough in the spring to secure a minimal number of interviewees.

This study therefore most closely reflects the concerns and interests of researchers willing to respond to the call, rather than a representational cross-section of the big-data community (a limitation anticipated by Ithaka S+R). While this does not reduce the insight given by researchers into the topic, it does signal that the report's findings must consider that the most consistent or urgent needs of the big-data community may not be fully represented.

Lastly, as would be expected, the pool of researchers at NYU, as any section of the research community, reflects the ethnic, racial, and demographic characteristics of the academy on the whole. This issue was itself raised by some of the interviewees regarding the state of their fields, which they acknowledged face problems similar to other parts of the academy. The subject of this study should be read in light of these concerns that also should be foregrounded in any consideration of research findings focused on how university work is conducted.

III. Results

A. Data Collection

Because handling of big data necessarily increases the complexity of data acquisition, it is useful to consider two main avenues by which a researcher would secure the data needed for analysis: primary data acquisition, in which a researcher collects the data directly, and secondary data acquisition, in which previously collected data is obtained or licensed for use.

i. Primary Data Collecting Strategies and Challenges

Researchers who were engaged in primary collection of big data were in the minority among this group, which was not surprising, given the scale at which a data collection or generation initiative must happen in order to create big-data levels of information. While the data collection tools vary across disciplines, the collection of large data sets involves ever-evolving and often highly sensitive instruments and hardware. Researchers are operating in an environment that requires the acquisition of new data collection skills to take advantage of evolving technologies and collection methods. One theme that was evident across disciplines is that the volume of collected and collectable data continues to grow, and that many of the challenges of big-data collection involve keeping up with a similarly growing volume of required skills, human resources, and technical resources. And funding, or lack of funding, plays a role in all of those elements.

The two most common challenges noted by researchers for primary data collection involve time and resources, and in particular the need to secure significant funding. In fields as different as astronomy and archeology, technological advances are allowing researchers to ask new questions by gathering data on a massive scale. The costs of these sophisticated data collection technologies, and the difficulty of securing that funding, are a challenge for several of our interviewees.

One means of reducing this cost is through shared use of instruments and hardware needed for data collection. This is common in the field of astronomy, for example, where telescope time for an instrument located outside of the university can be shared among multiple research teams, allowing some reduction in cost. A respondent who had obtained LIDAR data (laser-based measurements of the earth's surface from an aircraft) had similarly needed to mobilize shared resources to commission the aircraft that was used to collect that data.

Investments needed for data collection beyond equipment, tools and hardware include funding for research assistants and staff. Researchers from across the disciplines referenced shortages of research team members as a challenge to their work. As one respondent noted, it would be nice to secure "a Ph.D.

[researcher] in Computer Science [whose] goal is to help other people do their science really, really well and that attitude is so great to have in the group that I would love it, but I don't see any way I could possibly fund it." As tools allow researchers to collect ever more data, the core skill sets required are evolving beyond simply domain knowledge to more computational disciplines. Lack of funding for postdoctoral fellows or Ph.D. students challenges researchers who rely on a diverse set of skills in their lab or on their team. Moreover, the skills required for many large, long-term data projects such as data science, computer science, and research engineering are in demand in private industry as well as higher education. Even researchers with funding must compete for this talent with relatively constrained resources.

ii. Secondary Data Acquisition Strategies and Challenges

Much of the data used by the researchers in our interviews was secondary data. This tendency arises out of the resource mobilization required to collect data on such a scale, whether it be via sensor, instrument, a mass-public-participation platform (such as a social-media network), or by way of longitudinally collected data. The scale of these data collection efforts, or simply the impossibility of replicating them, necessitates that researchers draw data from elsewhere rather than collect it themselves.

The main challenges for acquiring secondary data include finding it, licensing it, and funding its acquisition. Finding secondary data can be a challenge on many fronts and involves discipline-specific idiosyncrasies. In some disciplines, there are obvious leading repositories where researchers can begin their search for data, or a search can be conducted by combing through published academic articles for referenced data. But where such lead repositories have not been established, finding and reusing secondary data can prove challenging, as much of the data is housed in local repositories, or stored in other local servers. The permanence and continued preservation of such data, whether for citation or for enabling reproducibility, is often unreliable. The increasingly common assignment of DOIs to datasets, and the development of standardization around metadata, both have the potential to mitigate some of the challenges noted for both data discovery and persistence.

Big data poses an added complication to the problem of finding secondary data, however, because such data is not always produced within an academic setting, and therefore it is not destined for a stable repository; it may never make the leap, absent the big-data researcher's intervention, from source to stable repository. In fact, in some cases such as social media data, for a data creator to make that data available would be to undermine its own stated business model. As one researcher noted, Facebook is the prime example of this: "Facebook has a genuine ethical reason for not wanting to make data too accessible, but they also have a very pragmatic business, legal, and PR reason for being cautious about sharing those data. This is true of other companies as well, but Facebook is sort of like the quintessential example of this." Such data not only comes with restrictions, but by its very nature is not meant to be findable.

Licensing challenges were another potential obstacle noted by several of the interviewed researchers. Academic publishers often have separate licensing terms for access to data that is computationally ready (text-as-data typically falls into this category). Researchers hoping to consume this type of data may need financial support from the university or department, as well as some level of licensing expertise to navigate this access. In some cases, this level of access may be entirely disallowed by the vendor. Another noted challenge is understanding the licensing terms and permissions for proprietary or sensitive data. One researcher noted the difficulty in understanding the details of a complex license for using social-media data and was uncertain about what was actually allowable under the terms. Such complexities are even more magnified when—as many big-data projects require—collaboration across multiple institutions is essential. As one researcher stated, “I do a lot of collaborations and the data piece is definitely a challenge, both because of the logistics of actually sharing the data, but also because of the ethics and the contracts.” Large cross-institutional research projects are, in short, inherently difficult because the use of standard contract terms by companies can interfere with collaboration opportunities.

Lastly, there is the issue of cost associated with such big datasets, owing to their creation and compilation by commercial data generators who do not operate within the financial world that academics (even those in well-resourced departments) occupy. As one researcher noted in regard to Twitter data, “It’s almost like the data doesn’t even exist or it’s like not an option for me” because of the cost. For others, it is not so much a matter of the impossibility of funding to access the needed data, but that big-data projects may fall in between the cracks of two main tiers of funding access. For example, costs of acquiring the data are too high for many NSF grants to adequately fund, but not at the scale of, say, a national initiative (e.g. a space exploration or pandemic response call for funding). As one researcher described it, “These projects, at this scale where you’re doing a project for years, involves a lot of people and millions of objects, it’s above the financial scale at which you can just write an NSF grant, but it’s below the financial scale where it becomes like a key project of NASA or something. So it’s really in a sore spot for funding and so that’s why a lot of our issue is fundraising.”

Aside from the aforementioned issues surrounding licensing, acquisition, and funding, secondary-data challenges arise out of the need to rely on the organization and documentation decisions made by big-data creators. One researcher described the experience of participating in a shared community of pooled big data, saying the landscape was akin to the”Wild West.” Problems include nonuniformity in data structuring, the use of bespoke data structures or query languages, or worse, shifting data structures as data providers abruptly change their data model, requiring research teams to regularly rewrite the code that is used to automatedly access and filter data: “Sometimes your query will fail, sometimes it will get different results, sometimes the data have changed under you, sometimes the interface has changed under you.”

The lack or incompleteness of data documentation is not limited to data created in academic research projects, but also in the commercially generated data that researchers often source for their projects. It was noted that, at times, documentation may not exist at all because the data being accessed is not viewed as a "research" dataset, or may have been built by a provider with little or no engagement with research communities. This was particularly true of researchers working with web resources, which may include both public web sources as well as "dark web" resources, and involve advertisement data, public-forum interactions, and webpages themselves. While these data are structured and obtainable—sometimes even with API facilitation—they are nevertheless rarely provided with provenance information. One researcher indicated that secondary data available from commercial generators has proven not useful because of its structure, requiring direct primary data collection. As an example, the researcher mentioned data that had not been provided in a format ready for distributed file systems (systems that split up data over multiple computing nodes to maximize capacity on less expensive hardware). Such data also contained further problems in its format:

> We felt was that the existing commercial systems don't provide the necessary support, either because they don't work in distributed computing and the datasets are so big—you know when you've got terabytes of data—or because they work in distributed computing but they're predominantly either a 2D system or they're text-based and they don't provide the functionality, or maybe they don't even support all the datatypes.

iii. Shared Primary and Secondary Data Challenge: Storage

Closely related to the high cost of obtaining large amounts of data is the challenge of data storage. The problem is not just a matter of capacity, as one researcher noted in discussing a related challenge, that of conducting data integrity checks. Performing checks to ensure that a dataset moved from one storage location to another is complete and its contents unchanged becomes much more time-intensive and involved at large data scales. This adds to the effort and outlay needed to manage storage, and encourages researchers to try to move data as few times as possible. In other words, setting up big data on a new device is rarely as easy as compressing a set of data files and moving them over; network speeds, ability to perform integrity checks, compressed file options, and other logistical challenges come into play, and along with them, cost and effort. As the researcher described a well-tracked, expertly managed big-data storage system, these are "beautiful systems, but that is a full-time job for a very good person on a survey [i.e. a big-data collection project], and most surveys don't want to hire a full-time person just to check MD5 [a commonly used hash function for ensuring two copies of a dataset are the same]."

Some storage needs in this realm are truly short term, yet still essential to the progress of a project. For instance, a researcher noted that it is common to need only a subset of a dataset, but that creating that smaller slice requires an initial processing on the full dataset (for example, to perform a merge of two input raw datasets). The researcher comments:

> When you have that huge dataset and you are like, OK, but I just need this one merge to work and then I can subset it down from there because you can't load it in memory. You need to build the whole infrastructure to be able to do that initial merge, just to then be able to bring it into Python later or whatever, so I think to me that's one of the biggest challenges I face in terms of the actual analysis of data.

The eventual analysis would thus be manageable, through storage on commonly available systems, but the initial handling of the data requires temporary use of a much more capacious system.

The problem of secure-data storage (notably, when there is data with personally identifying information) can potentially affect any researcher, whether working in big data or not. But in the case of large-scale data, having access to a system that is both secure and capable of holding big data is more difficult—even more so if the data cannot leave that storage environment for the purpose of analysis. The need for secure data storage among the researchers interviewed was strongest with regard to social-media data and data collected from the web. In these cases, the data was not necessarily IRB-protected because it was open data available on the web. However, there were ethical concerns that led the researchers to comment that a more secure environment was desired, especially when it was not always clear whether there might be personal information in the data (as in a public Tweet that contained personal information and just happened to be within a researcher's dataset).

The interviews conducted with NYU researchers revealed a patchwork approach to solving data storage. Some researchers store their data on the NYU-provided clusters, and noted that NYU's investments in high-capacity storage and HPC have significantly aided their research efforts. Other researchers contract with private data storage providers outside of the university. One noted reason for offsite storage was the need to be able to collaborate with partners from other institutions. Researchers also indicated that it can be difficult to understand the appropriate level of storage security needed for different types of data, and to understand what storage options are available at the university for these different levels.

B. Training

i. Current Practices

The characteristic of big-data-related training most frequently mentioned by researchers is its informal nature. Researchers sought training on an ad-hoc, as-needed basis, though several perceived a need to make this training more formal and widespread. Training in skills needed for big-data work is conducted through self-learning (often through online resources), peer-to-peer teaching (a colleague is sought out to impart data skills), and by way of mentors teaching mentees (and vice versa) within a field. Podcasts, YouTube, Stack Overflow, and other online or media-based resources were cited often as sources for self-learning. Researchers also recommended workshops and data bootcamps as strong options. Many mentioned that this training is not pre-planned, but takes place as needed when a particular skill is required to proceed with research. One factor that likely plays a role here is the ever-changing nature of needed data skills. It makes little sense to expect a standard set of skills of all big-data researchers when those needs change with the rise of new computing languages, software, or techniques. This in itself prompts the need to continually train in response to new techniques in the field, as an ongoing learning process, rather than one entrenched training curriculum that could be completed and never revisited.

Formal big-data training via a university course is not unknown, but researchers identified several caveats about this approach. Eligibility for courses is typically most easily obtained by those still having student status, and the availability of such training was only discussed here in the context of graduate students who were collaborating with senior researchers. The most common examples were students taking courses in statistics or data science outside of their home program, usually on an as-needed basis to fulfill specific needs for research. This type of training was viewed (regretfully) as happening outside of the norm—it was more difficult than it should be (leading one to recommend that universities should better facilitate cross-departmental class enrollment) or happened only when a student broke the mold by seeking additional training. Formal training as part of a graduate degree program—particularly for those senior researchers who obtained training for their Ph.D. in past decades—was described as nearly nonexistent, especially when it came to the day-to-day of working with data. Instead, graduate training was described as largely methodological or theory focused, with little time spent on how to do the work of building, manipulating, or managing data.

ii. Training Challenges

Challenges to training those new to a field have not disappeared, even for those individuals who are working with senior researchers dedicated to facilitating that training. One problem is the wide range of computing skill levels that new researchers bring to the field. As one respondent put it, one finds a very wide range of backgrounds among researchers who are brought together around a single research field: those with solid but mostly functional programming skills; others with a full-fledged computer science degree; as well as a third category, having very advanced, software-engineer-level ability in computing. Given this diversity within a single lab or research group, further training cannot happen effectively with a one-size-fits-all approach. Another challenge lies in the need for such training to be non- or cross-disciplinary in nature; researchers expressed concern that training in big-data skills specific to a particular department would involve discipline-based features, rather than remaining a neutral place for data-skills training. A few mentioned the library as a place where this more neutral training might occur, though the central idea behind this idea was the institution's neutrality from a disciplinary point of view, not necessarily its data-rich features and expertise.

iii. Researcher Recommendations

Researchers offered some recommendations regarding the training content needed for big-data work. One surprising need, because it is not confined to big-data projects only, is for better training in data visualization and scientific communication. As with training in big-data techniques in general, one respondent noted the prevalence of self-taught or community-led learning around visualization techniques. Another often-cited topic for training was high-performance computing use, as well as the related topic of distributed computing (e.g. computing across a multiple-node cluster and/or computing using files distributed across multiple file systems—essentially, the setup needed for big-data computing). One respondent noted that the size of data in the researcher's field had become so unwieldy that it had been seven years since a newcomer to the field could manage performing calculations needed on a standard desktop system. High performance computing, in other words, has become necessary to do all work in their field. Other researchers pointed to similar trends—high performance computing knowledge either conferred a tremendous advantage to the work, or else it was becoming a liability for disciplines not to have knowledge of working in that environment.

Less-often cited, but still notable subjects for training include specific languages (Python, R—especially approaches that teach one language through an approach that assumes familiarity in the others, basically cross-training on both languages using commonalities), machine learning (both as a methodology and how to deploy it on a computing system), and data-science skills generally. Managing data from a computational perspective, especially data hosted in a formal database, also arose as a need, database skills often being omitted from data training because they are seen as part of a more strictly computer-science knowledgebase, and largely the purview of a system administrator. Generally speaking, there is a perception that more advanced training ought to be offered, particularly within the context of respondents who knew about classes offered at NYU Libraries but found them insufficient for the needs of those moving beyond introductory skills. One respondent, however, also expressed skepticism about whether any training could be advanced enough to meet the needs of today's research, noting that even some computer science Ph.D.s struggle with the computing required for big data, meaning that those coming from a more casual background in computing would have a long road to travel to master advanced training at a level necessary for the type of research being done.

As with nearly every aspect of research, lack of time was cited as the primary barrier to receiving training. This manifested itself not only in concern that the training itself would take up too much time, but that even the effort required to find the proper training resource was a time cost for which no provision could be made. More than one researcher indicated that just putting in enough time to maintain disciplinary expertise in a field was barely possible, and that adding skills training on top of that was incompatible with those time requirements.

C. Researcher Needs and Areas of Support Big-Data Research

i. Perceptions of Current Support

When asked about the kind of support they receive from their departments and university, researchers noted several areas where they are receiving support, as well as those areas they would like to see more. NYU's significant investment in technology infrastructure was noted as an asset to researchers working with big data. Some researchers have also taken advantage of some of the training and consultation available from the NYU Libraries' Data Services group as well as university IT's Research Technology group. The Data Services group offers guidance on data management planning, the selection of a repository, and the creation of DOIs. In conjunction with Research Technology, they are able to facilitate access to large storage workspace allocations and high-performance computing.

ii. Expressions of Need

Researchers also noted areas where they would like to see more support from internal partners. Training was brought up as a need that could be met by university departments. Researchers would like to have additional training for computationally intensive projects—both for themselves and for members of their teams and departments. Some of the challenges of meeting this need, given the diversity of necessary skills and wide array of tools, are noted in the training section above.

Two researchers expressed a general need for hands-on data managers, specifically as managers of database-structured and database-stored data (e.g. as a SQL-based system, though the need seemed to be more for a queryable and stabilized data environment than a SQL system exactly). This was expressed specifically as a desire for personnel to handle this, something not often available within the confines of projects as currently designed. Comments one researcher, "We know where to look [for high-performance computing resources] but for this data storage, database management [support] . . . I do not know actually."

While support from the libraries was noted in several of our interviews, there are indications that researchers would welcome more assistance from partners in the libraries. Clear guidance regarding accessing large storage allocations and high performance computing was a request from one researcher. And while the NYU Libraries offers robust research data-management support, not all researchers know of its existence, or understand how to avail themselves of this support. One researcher's comment regarding data management was: "it's like wading into a stream and not realizing how deep it is and then all of a sudden you're in trouble and you don't even know how you got there."

Lastly, although most researchers indicated that the data they tended to use were open data, and thus licensing was not usually an issue, there was some expressed need for assistance with licensing data (especially, large, proprietary data sets such as Twitter).

D. Sharing the Results of Research

While the researchers we spoke to were largely supportive of the idea of sharing data in theory, their motivations for sharing were varied, and the burden of sharing was noted. While journal and funder mandates were explicitly mentioned as reasons for sharing, researchers also made reference to the classic Mertonian norm of “communism” (i.e., sharing and collaboration are essential to the larger scientific enterprise). At the same time, they expressed different fundamental philosophies for determining what they share. At one extreme, there is the kind of sharing that is considered an integral part of a messy, communal process. This attitude is to share as many research products as possible, as soon and as widely as possible, and not to worry about the “messiness” of what was shared. This attitude considers it the active role of the community to use and correct each other’s data and code. Some also mentioned that sharing messy data can create a feedback loop to encourage even more sharing because it lowers the stakes and removes the stigma: “I think it’s actually valuable for peoples’ code to be visible on the web even if it’s a mess. Maybe even especially if it’s a mess because you get other eyes on it, other people feel more comfortable about sharing their code.”

At the other extreme end of the spectrum is the more “cautious” kind of sharing that is done to comply with journal and funder mandates. This attitude is much more conservative, with researchers only sharing final stages of their data or code and making sure that everything is functionally reusable before sharing. What explains a more conservative attitude to the “messiness” of what is shared? In some cases, researchers consider sharing one’s “messy” data to be costly to one’s reputation: “You do not want to just kind of make it public, and then [have someone] download it and find it does not work. It hurts your reputation and your credibility, so . . . you need to do those checks before you go public.” In other cases, they anticipate that releasing “messy” data will require extra time/effort to answer questions about it after the fact and to provide “hand-holding” for its reuse by the community: “We tend not to release the code, mostly because I don’t want to get into a situation where I'm trying to help somebody re-implement it, because I'm not a computer scientist and, you know, there is a reason for that.”

In between these two extremes lies a range of attitudes and approaches. Also, sometimes researchers elect to share or not share specific datasets depending on the circumstances of their creation and their perceived unrealized promise: “We spent a lot of money and headache collecting it, and it provides us a strategic advantage for going after certain grants, and we’ve not had the funding to support the students to really look into that.” Potential patents and profitability were also mentioned as reasons that specific data may not be shared, even when sharing was otherwise generally described by the researcher in positive terms.

Some participants noted that the sheer size of their data presented technical and legal barriers to sharing it. Or, even when size technically may not have been a barrier to sharing, its size, combined with the way the data had been gathered, made it impossible to know whether it would be okay to share it: "We don't know exactly the source. If it is copyrighted material, or I don't know. In this case, for example, we are crawling data from the web."

Naturally, privacy was another reason that researchers did not share their data, though in some cases, even when sensitive data potentially could be shared, researchers still chose not to. One reason mentioned for this is that the algorithms and processes for deidentification of data have become so complex that it is hard to be sure you have been completely successful in producing a releasable dataset.

E. Ethics/PII/Secure Data

i. Prevalence of Secure Data Issues

Ethical concerns and the presence of personally identifiable information (PII) in data are a factor in the ways researchers approached their work, but the issue was not an overbearing aspect of their research. There are a couple of reasons why PII is not relevant to this particular group of big-data researchers. First, the realm of data available from sources that can generate big-data scale datasets is often either not identifiable, or does not involve human subjects at all. Astronomical calculations, mass 3-D scans of archaeological objects, and repositories of chemical compounds, for example, are not by their nature relevant to PII or human-subject concerns. Even medical research, it should be kept in mind, does not always involve identifiable data—a large-scale dataset of neurological scans or cell imaging, for instance, where that data has been divorced entirely from their source, exist beyond the realm of identifiable data.

This situation is further impacted by the strong trend toward secondary-data use by researchers. This means that researchers are working with a dataset that has already been de-identified by the data creator, or that PII can be removed in the course of generating the secondary data needed for the project. This aspect also may simplify the IRB aspects of data use, in that researchers can follow the original stipulations that guided the original data collection, rather than assembling a policy from scratch.

ii. Social-Media Data Special Case

One area that is heavily implicated in larger PII and ethical concerns in the world of big data, however, is social-media data. As several researchers observed, social-media sources are in many ways at the heart of big data because they are one of the few places operating at a scale sufficient to generate mass data. In some ways, big-data research is inseparable from the rise of social media. And the use of social media involves ethical concerns regarding use of its data. While web-published content has typically been treated as outside of IRB considerations, since it is public and does not fit the format of a consentable data collection activity, its users are considered by researchers to have a reasonable expectation that their data will not be used, or will be used in a limited fashion. Moreover, social-media data sources are inevitably subject to the gatekeeping of the companies behind them, and that has implications for licensing of that data, the transparency of how companies have assembled data explicitly for research use, and prosaic concerns like costs and citability. Finally, at least one researcher raised additional concerns over the ethical ramifications of vulnerable populations recorded in social-media data, referencing data produced by Black Lives Matter participants, for example, as a case where the risk to those persons' data would be higher than the average social-media user.

iii. Support Needs for Secure/PII/Ethics Related Data

The researchers indicated that, where ethical and PII concerns are present, there is a need for better support and guidance from the university. One researcher noted that, with regard to social-media ethical concerns, IRBs are "very much behind the times and haven't really caught up to sort of where things are now." In that absence, the community of researchers often substitute as a sounding board for how to approach these ethical concerns, but there are limitations to how effective this can be—the research community desires the ability to use data to answer questions, so it is hard for that community to also be the means of setting boundaries for when that data should be used.

Another need related to PII data is technical, as a few researchers mentioned that big data can exhaust the limitations of a local computer that would be suitable for PII (presumably because access is so restricted). This requires shifting to a PII-appropriate high-performance computing environment, something that is not always easy to obtain.

One unexpected concern raised by a researcher concerned the ethical concerns—or really, intellectual property questions—surrounding data generated by cultural-heritage objects. This most visibly involves 3D scans of objects, but could also be extended to various types of data generated around archaeological, artistic, or other cultural objects. Because these objects have traditionally been treated as the cultural representations of national entities, and therefore restricted in where and how they can be reproduced, referenced, or described, there is more potential for restriction on researchers' use of data. To date, that has not played out overtly, but there is strong potential, similar to the gatekeeping attempted by social-media companies, for those in possession of the source of big data to put limits on dataset use that are not typically seen in the sciences.

F. Trends

Many participants noted the interdisciplinarity within their fields generally, as well as disciplinary "fuzziness" of their specific research topics, expressing that this makes it hard to keep abreast of developments, and to communicate their work meaningfully to potentially interested funders and audiences. Also, because research involving big data necessitates deep computational expertise, the distinction between domain knowledge and computational knowledge has become increasingly blurred, making it difficult to recruit and train collaborators and students with the right skill-sets. Furthermore, once workers with appropriate skill-sets are identified, they are in high demand across industry, so it can be increasingly expensive to pay them.

Researchers noted a trend toward using data-science methods (especially machine learning) for their own sake, rather than considering them a new way to answer domain-specific questions. One result of this shift is a reduction in one's options for future specialization and focus: "Anybody at this point can go out . . . and collect fairly good X kind of data or Y kind of data. What I tell my Ph.D. students is, you know, what you really want is to think about an area of expertise to develop a line of research that very few other people can do." Along those same lines, they note a tendency toward indiscriminate (and potentially inappropriate) use of existing data-science tools and methodologies, rather than developing new ones that may work better for specific domains, purposes, or types of data. Researchers may also feel significant pressure to push themselves "across the data-science line" in order to secure funding.

The new need for, and creation of, specialized databases (e.g., for management of numerical data, for large-scale datasets, or for distributed databases beyond SQL) was another trend mentioned by multiple participants. Teams and labs are also increasingly in need of specialized databases to help them find and manage internally generated data and be aware of its quality and the other ways it's already been reused.

G. Diversity

Interviewees were not directly asked about diversity in the field, but it would fruitful in future investigations related to trends in supporting big-data research to consider how access to training, incentive and promotion, and the distribution of university resources on a national level affect the ability of women, minoritized peoples, and members of other historically excluded communities to enter the field or conduct work. Moreover, issues of diversity were raised by a couple of interviewees even without prompting, and it is useful to report the concern expressed about the tendency of white males to dominate the field, in line with broader problems with diversity in the academy, and an observation that rising practitioners in the field (in particular, current students) represent a much more diverse cohort than in previous years..

IV. Summary of Recommendations

While the challenges of time and funding are somewhat inevitable in fields conducting large research projects, the interviews highlight ways in which both universities and libraries can assist big-data researchers with overcoming some of these challenges.

A. **University-Level Support**

   a. Training

      i. Universities may need to be wary of establishing static, set training resources; researchers indicated that skills needed are ever evolving. Skills training resources should emphasize flexibility in tools available for use, depth of training in concepts with proficiency in multiple computer languages, and discipline agnosticism

      ii. Facilitate ability of students to cross-enroll in relevant courses to big data work outside of their home department

      iii. Training should be one means of targeting advancement of the diversity of the field.

   b. Licensing and Data Ethics Support

      i. Design a standing advisory university body, akin to the Institutional Review Board model, that can advise researchers seeking guidance on the ethical use of microdata generated in a public context (e.g. on a public-facing social media platform or website)

      ii. Research groups will need coordination for licensing expensive big data in a way that can more efficiently take advantage of larger number of users (rather than expensively licensing a dataset for just a small number of users on a single project)

   c. Computational Infrastructure

      i. Providing storage and a high-performance computational environment for sensitive data—data with PII, for example—is a looming problem for the university community, albeit not one that will be universally needed by all big-data projects. This is solvable in a project-by-project way in the short term but as more researchers come to take on big-data projects with sensitive data this may need bigger scaling.

      ii. Big data poses unique challenges for moving data to and from systems, and researchers benefit from single-point storage systems that can be accessed from multiple computational environments.

   d. Costs

      i. Universities should be aware that personnel, particularly lab assistants working with big data may need longer and more in-depth training to manage the data needs of a project; whereas in the past salary may not have been a concern, training and retaining such high-skilled assistants against the pull of private industry is an issue and may drive up university costs of such personnel.

ii. As with personnel costs, universities should be aware that researchers are finding that the data acquisition and hardware costs associated with big data to exceed those of earlier projects, making even traditional large, first-tier federal grants insufficient to carry out work.

iii. Coordinating groups across multiple departments or schools to share data licensing and infrastructure costs will help mitigate expenses as well as bolster the naturally cross-disciplinary nature of big data work (especially projects centered on social media and 3D imaging).

B. Library Support

a. Training

i. Assist in targeting training in skills considered underserved in current training: visualization and data publishing, databases for managing research data, and high performance computing.

b. Licensing and Data Ethics Support

i. Build expertise and capacity for questions around licensing secondary data generated in the context of for-profit, often social-media or web-based entities.

ii. Take a leadership role in pressing vendors of content it licenses to allow computationally based access to that information, and provide the library and its users with clear contractional statements about what can and cannot be done when accessing its platform via automation.

iii. Licensing planning needs to take into account the often-required multi-institutional nature of big-data collaboration. Library should look for clarification from data providers it licenses from on access by research teams across multiple universities.

c. Archiving, Curation, and Data Publishing

i. Lead on researcher adoption of FAIR (Findable, Accessible, Interoperable, Reusable) data standards to mitigate big data researcher problems of interoperability, use and reuse.

ii. Increase capacity for long-term archiving, curation, and preservation of big-data project files, software, and documentation. Data publishing for big data outputs is an ongoing challenge because of the difficulties of both hosting and distributing it.

V. Conclusion

While the NYU portion of the larger research project collected data from a small sample of participants, the conversations unearthed some common themes relating to the opportunities and challenges in working with big data, as well as the remarkable pace at which this landscape is changing. While funding was often highlighted as one of the larger challenges, incentive structures, support-role alignment, and lack of technical training of graduate students were also noted. The data required for their research has many challenging aspects. Creating, finding, and re-using research data is time-consuming, expensive, and frustrating. Funders, university grants offices, and administrators, as well as leaders in library research departments, are all in a position to assist researchers overcome some of these challenges and foster this work. Our recommendations based upon these conversations, and informed by our roles in supporting academic research, are focused upon using the budgets, tools, expertise and administrative power of these roles to offer these pioneering big-data researchers the support they need to explore their disciplines and make new discoveries and breakthroughs.