Assessing the Research Practices of Big Data and Data Science Researchers at Texas A&M University: An Ithaka S+R Local Report

John Watts

Director of Research Data Management Services

Laura Sare

Government Information and Data Librarian

Paria Tajallipour

Health and Life Sciences Librarian

Carolyn Jackson

Agriculture and Life Sciences Librarian

Table of Contents

Executive Summary

Project Aims

- Deeply explore the ways in which researchers work with big data
- Understand researchers' support needs
- Develop actionable recommendations for stakeholders
- Build relationships within the institution and across institutions

Background

Big data research and data science methods have been adopted by researchers with a broad array of disciplinary backgrounds. Rapidly changing technologies both challenge and advance scientific methods. Researchers engaged in big data research navigate these challenges in diverse ways to answer research questions that were unfathomed only a decade ago. The risks and rewards of big data research at Texas A&M University are undocumented and the conveyance of the current practices and needs of researchers across disciplines is difficult to codify. The University Libraries and other campus partners can collaborate with big data researchers to navigate this rising research trend and provide support as new big data methods and technologies emerge.

Methods

This needs assessment is part of a multi-institutional study coordinated Ithaka S+R, the non-profit research and consulting arm of Ithaka. Twelve full-time faculty members were recruited through convenience sampling. The semi-structured interview instrument was developed by Ithaka

S+R staff. Interviews were conducted from September 2020 – December 2020 via Zoom. Qualitative analysis was performed based on the principles of grounded theory.

Findings

Making Big Data Work: Community, Tools, and Support

Researchers working in big data noted that they had to look outside of their traditional disciplinary community to perform their research. They also looked beyond academia for collaborators, often working with local, state, federal, or international government agencies; public utilities; and industry. Researchers also sought collaboration with campus partners such as the Institute for Data Science and the High-Performance Research Computing center to support their efforts in the storage and analysis of big data. Many interviewees shared that collaboration was necessary to understand diverse datasets, but it can be difficult for collaborators to identify a shared epistemology and ontology for their work.

Applied Learning: Current Practices and Needs

Many researchers expressed that the nascent nature of data science practices and the use of big data in their respective fields required them to learn new skills throughout their career. Indeed, the emergence of these practices invites researchers to ask new research questions that, in due course, dictated the skills and knowledge they will need to acquire. Interviews revealed that most big data skills were learned informally through self-paced inquiry using inexpensive, online forums such as YouTube and Stack Exchange. Researchers also sought short courses offered at conferences or at the University, but the data revealed that there was not a primary, vetted source for learning how to work with big data.

Acquiring and Utilizing Data: Challenges to Access and Quality

Finding big data that is accessible and accurate was a consistent theme throughout the interviews. Subscriptions to some proprietary data are inevitable, but researchers are unable to sustain long-term costs due to the temporary nature of external grant funding. When using secondary data, many researchers noted that the context of the data is critical to understanding the data's veracity and bias. However, information regarding collection or processing is often unavailable, leaving many unanswered questions. Proprietary data may be generated and held by private industry protecting their investment or public entities that need to protect it from malicious parties.

Sharing Big Data: Current Practices and Challenges

When speaking about sharing big data, topics such as collaboration, storage, security, and dissemination were discussed by researchers. To share their research, some of the interviewees create their own webpage to upload and publish their datasets, codes, or models. These webpages are sometimes personal websites but often maintained by the University, such as laboratory websites. Publishing datasets in an open data repository does not count towards the tenure process of the faculty, and therefore, the incentive is low for researchers to publish that way.

It is also unclear for some faculty whether they can show the impact of their data, as they were unsure about tracking citations to see if their data has been used in other papers or policies. Other challenges with sharing big data among group members or when publishing the data are related to security or licensing issues. Working with high-security data makes it difficult to collaborate and publish the data openly.

Recommendations

Provide Targeted Training

The University Libraries currently offer a suite of data management workshops for all University students, faculty, and staff, but researchers expressed a need for additional learning opportunities on big data topics in the form of short courses or workshops.

- These courses should be free of charge and offered regularly.
- Training should teach both introductory concepts and theories as well as specific skills for using software to clean, manipulate, analyze, and publish data.
- Teach Carpentries workshops in several key software applications for big data that were noted by researchers, including R and Python (The Carpentries, 2021).
- Seek opportunities to plan and host workshops in partnership with other campus units such as the TAMIDS and the HPRC center.
- Offer more resources and instruction on best practice for de-identifying data in partnership with members of the Division of Research and the Institutional Review Board.

Incentivize Data Sharing

Incentive and privacy were noted as significant barriers to sharing data among the interviewees. Many indicated that their disciplines and the University did not provide significant incentives for sharing data as a product of research.

- Libraries should serve as a leader in incentivization by offering a suite of services and resources.
- Provide greater support by librarians for data deposit and curation in open access repositories.

- Create an open access data award for graduate students who deposit data in the data repository.
- Feature big data researchers who currently deposit data in newsletters and on the Libraries website.

Collect Big Data

There are opportunities for the Libraries to support the purchase of big datasets that have potential value to multiple researchers across campus.

- Create an annual call for proposals for a one-time purchase of big datasets that can be stored and distributed by the Libraries.
- organize, clean, and store regional GIS data currently held by the Libraries for immediate use by researchers.

Collaborate with Campus Partners to Provide Support

General Support

- Create a university-wide committee to identify, define, and communicate learning opportunities, storage, analysis, and policy for data so that information can be codified at a single point of contact.
- Work with the Institutional Review Board (IRB) as a place to house guidance for sharing big data as part of the IRB application.

Proposal Development

- Provide guidance on drafting grant proposals to include a budget for subscriptions to proprietary, big data collections.

Introduction

Ithaka S+R Research Study

To study the research practices of scholars engaged with big data, Ithaka S+R, the non-profit research and consulting arm of Ithaka, partnered with 18 university libraries across the United States. Texas A&M University Libraries was invited to be one of the participating libraries in this study, which aims to identify the support needs of faculty using big data and data science methodologies. This local report seeks to describe the experiences and perceptions of these scholars within the context of Texas A&M University.

The University Libraries chose to engage in this study to identify potential pathways to collaborative support for big data researchers. Texas A&M University is home to a large and diverse community of scholars who advance discovery and create new knowledge through innovative, interdisciplinary research. In recent years, innovation has taken the form of big data research and data science methodology. To that end, the Libraries seeks novel approaches to research support this nascent and evolving area. Through this project, the Libraries will make informed decisions regarding the development and sustainability of novel resources and services devoted to big data and data science research.

Texas A&M University

Founded in 1876, Texas A&M University (TAMU) is a Carnegie R1 institution and one of the first universities to be designated a land, sea, and space grant institution. TAMU has a heavy science and engineering focus and provides 133 undergraduate degree programs, 175 master's degree programs, 92 doctoral degree programs and five professional degrees (Texas A&M University, n.d.). The faculty includes over 3500 instructors and researchers in 19 colleges and schools (Texas A&M University, 2021). TAMU is one of two flagship universities in the state of

Texas and in Fall 2020 had the largest student body in the nation with an enrollment of 71,109 students (Texas A&M University Accountability, n.d.).

One primary data services provider at TAMU is the Texas A&M Institute of Data Science (TAMIDS), which promotes research, education, service, operations, and outreach in data science across the University. TAMIDS provides data science education, training, and research support (Texas A&M University Institute of Data Science, n.d.). Another predominant data services provider for the campus is the Texas A&M High Performance Research Computing (HPRC) group, which operates advanced computing and data resources to enable computational and data-enabled research activities. HPRC also provides consulting, technical documentation, and training to support users of these resources (Texas A&M University High Performance Computing, n.d.).

Interviews were conducted with full-time researchers from the following TAMU colleges and institutes:

- The College of Engineering
- The College of Geosciences
- The College of Agriculture and Life Sciences
- The School of Public Health

Methods

To assess the needs, practices, and challenges faced by researchers working with big data, a qualitative research approach was taken. To identify the researchers appropriate for the study, the authors met with four stakeholders across campus. These individuals were selected due to their knowledge of research practices across disciplines. Meetings were conducted with individual stakeholders via Zoom. Authors met with each stakeholder to discuss current research in big data at the University, identify candidates for the interviews, and explore areas where this project could be impactful.

In the second stage of the research project, a list of researchers who work with big data at TAMU was created through purposive sampling. The TAMIDS webpage was used to identify potential interviewees. The authors also listed researchers they worked with, or those researchers suggested in the stakeholder meetings. The authors also consulted the Scholars@TAMU database to search for potential candidates. Interviewees were selected solely based on their research interests and work with big data.

At first, 38 faculty were identified and 28 were contacted. Most of the participants came from the School of Public Health, College of Agriculture and Life Sciences, College of Engineering, and Department of Geography. A recruitment email template, created by Ithaka S+R, was used to adhere to IRB (Institutional Review Board) protocol and keep all emails uniform (see Appendix A). If no response was received, a follow-up email was sent approximately two weeks after the first email (see Appendix B). Twelve faculty and staff accepted the invitation (30% response rate).

Prior to the interview, a second email (see Appendix C) was sent to the big data researchers with a calendar invitation and a copy of the consent script (See Appendix D), which were both provided by Ithaka S+R. Participants were interviewed individually through Zoom by one member from the project team. During the interview, webcams were turned off and Zoom settings were established to record only audio. Audio recordings were later transcribed using a third-party transcription service.

Using a 6-step grounded theory approach, the authors first independently engaged in the open coding of three de-identified transcripts to find recurring themes. In step 2, the authors independently grouped and ordered their open codes to create hierarchies and eliminate duplicate or irrelevant codes. In Step 3, the authors met to compare their individual hierarchies of codes to identify four core themes among the three sample transcripts. In step 4, each author was assigned a single theme and individually coded all 12 transcripts based on their selected theme. Finally, each author reviewed the codes for their theme to summarize findings.

Limitations

Through purposive sampling, twelve faculty members volunteered to take part in this project. Although every effort was placed in recruiting a diverse group from various colleges, participants represented only four colleges. Due to the sample size, the results cannot be generalized to include all needs and concerns of faculty at Texas A&M University.

Findings

Making Big Data Work: Community, Tools, and Support

Community and Tools

While researchers had their disciplinary affiliations, there did not appear to be a community for big data. They also had to look outside of academia for collaborators, often working with local, state, federal, or international government agencies; public utilities; and industry. Some of the researchers had issues publishing in peer-reviewed journals, as reviewers did not understand what they were doing, especially when creating new models and tools to work with big data. Several researchers explained that their fields were so new that people within were still trying to define common vocabularies.

Keeping up with technology is a major issue for researchers. This effort to maintain currency is a struggle due to time and the frequency of new software programs and new features added to existing software. Often a main reason for collaboration is that a researcher needs to collect or analyze data through software they are unfamiliar with, and so they seek a collaborator to do that portion of the work. Most of the time, these collaborations are with other researchers, mostly at a local level, but some collaborations are with vendors of the software. Some researchers seem to rely on their graduate students to keep up with the latest software developments. Student projects can also drive research. Students often handle more of the tedious tasks related to big data. One researcher summed this up:

"I don't know how to manage GIS, but I know more on the decision making, risk approach. And so, partnering with other collaborations, collaborators and students that come with these backgrounds, that has helped us to create new fields that now are supporting our research and expanding our research base."

University Support

University level support is a significant need. The HPRC was the most noted example of support both a tool (computer hardware/software) as well as training (workshops and consultations relating to HPRC services). One idea from a researcher was to use the IRB as a place to house help or guidance beyond just getting a research project approved. If the researcher must go through IRB anyway, data suggestions or tools could become part of the IRB protocol. Also, with the newness of the field, some researchers create their own tools to gather and analyze the data they need. Because of the limits on their time, while they understand that it is necessary to share data, they want tools to make this easier for them to create metadata and deposit their data. Most of our researchers were willing to share data with other researchers if asked, but the concept of making data FAIR (Findable, Accessible, Interoperable, and Reusable) and posting it publicly for anyone to use was a bigger hurdle and not seen as part of the job (Go FAIR, n.d.). University support was identified as one way to overcome this hurdle. One researcher shared that it would be helpful to have University training on best practices for preparing big data to share it openly.

It did not appear that researchers saw data management as something to be included in their project overhead. Again, the newness of the field and the new types of data being used can mean that researchers must spend money to purchase data for their projects. This is potentially expensive – requiring the purchase of hardware and server space to handle the large amount of data they were working with in addition to the more traditional costs, such as funding to hire students or support graduate assistants. Some researchers wanted the University to provide support for hardware and purchasing data, especially data that could be used by multiple researchers and disciplines. Most researchers did not collaborate or allow for assistance in analyzing, storing, or managing data. Only one researcher recruited a librarian in their grant to post their research findings online and in repositories.

Applied Learning: Current Practices and Needs

Many researchers expressed that the nascent nature of data science practices and the use of big data in their respective fields required them to learn new skills throughout their career. Indeed, the emergence of these practices invites researchers to ask new research questions that, in due course, dictate the skills and knowledge they will need to acquire. Many researchers shared that they did not receive training in these practices in their PhD programs. While data analysis was a part of their foundational learning, the availability of data and the methods and tools to process or analyze big data were uncommon. However, researchers expressed that the fundamental principles of their field were still essential to their work. One researcher shared that disciplinary knowledge, principles of scientific inquiry, and research integrity are core competencies for data science professionals. Academic programs offer disciplinary foundations for research, but many of those interviewed sought additional learning opportunities outside of their programs and well into their careers to learn how to leverage big data in their research.

The Rapid Pace of Development

While the rapidity of technological advancement has enabled researchers to ask and answer new questions, the pace of these advances creates a discernibly significant and ongoing awareness gap. For some researchers, the integration of data science and big data felt like a departure from their original training as a researcher. One researcher specifically shared that this integration felt like a balancing act of working within two disciplines - the original discipline for which they were trained and the data science discipline. According to the interviewees, these skills are developed at the point of need; since big data is so new, researchers educate themselves on topics or tools as their project demands it.

A Dynamic Data Landscape

While data science and big data were noted as complementary to researchers' primary disciplines, there are disciplinary differences that place the researchers in a liminal learning space. Understanding what they need to know to answer a research question can be a difficult beginning. Moreover, it can be challenging to stay abreast of contemporary trends in data science because it is not their primary field of study. Information about emerging big data trends is abundant, but researchers shared that they often choose only a handful of information sources to stay current due to lack of time and the eventual information overload. Another cause of liminality is a lack of stability, which is an outcome of data science and big data's dynamic landscape. As technology advances, core competencies of this relatively new area shift. One researcher noted that data analysis software that was essential to big data methods only five years ago is now antiquated, and he must learn new ways to approach this work on a regular basis.

Working across disciplinary boundaries, several researchers shared that the highly technical nature of big data practices pushed them to learn to be both a scientist and an information technology manager at once. Many researchers grapple with advancing their scientific agenda while learning to use cloud computing and high-performance core processors. At TAMU, researchers rely on the Division of Information Technology to set-up these services, but many indicated that they learned to leverage the capabilities of these systems through trial and error or with training from the software and hardware vendors.

In addition to the dynamism of technology, researchers flag the interdisciplinarity of data science as both an advantage and a disadvantage. Big data is, by definition, an assemblage of data from multiple sources, some of which are not germane to an individual researcher's primary area of study. For example, a researcher working with electrical data may integrate weather data into their models to predict power outages. While this diversity of data fosters collaboration among disparate domains, a lack of a shared epistemology and vocabulary was noted as a unique challenge when working with big data.

Modes of Learning

To overcome the barriers of learning described above, researchers shared several approaches to advancing their own understanding of big data and data science practices. While these approaches follow some traditional methods, such as academic conferences and journal articles, one primary theme among interviewees was the notion that learning is often self-paced and sought at point of need. One researcher shared that it was difficult to identify one course or learning experience that would sufficiently meet their needs. Rather, they sought an assortment of online and in-person learning experiences - online tutorials and short courses - for both initiatory learning and ongoing development. This could be described as just-in-time learning for a specific outcome to achieve in an established timeframe.

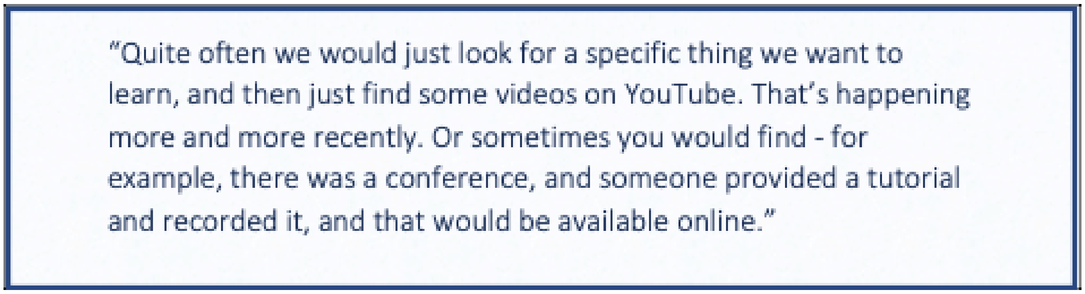
Low Cost

A subtheme from the interviews indicated that most of the data science and big data learning experiences were either low-cost or free-of-charge. Two researchers shared that learning can be an expensive investment of both time and financial resources. Those who serve as primary investigators also stated that choosing these learning experiences for members of their labs required some strategy. If a primary investigator invests time and resources in the professional development of lab staff, it is difficult to justify the investment in a graduate student employee who will eventually graduate and move on. Therefore, low-cost training was identified as a primary need when engaging a team of researchers in professional development.

Online Learning

One of the most popular educational opportunities for interviewees is online courses. Massive open online courses (MOOCs) were flagged as a predominant avenue for introducing researchers to big data and data science. Coursera was mentioned specifically by two researchers as a quality source of introduction to big data and data science methods. Coursera partners with instructors from over 200 universities to offer a suite of MOOCs on data science (Coursera, 2021). Learners from universities can gain access to these MOOCs free of charge, but Coursera does charge learners seeking certificates in data science or big data topics.

In addition to MOOCs most researchers sought learning opportunities freely available on YouTube and pointed their students to this site to learn more about big data. YouTube videos of lectures recorded by other researchers in their field, specific troubleshooting instruction, and software guides were identified as the dominant purpose for using YouTube as a learning tool. YouTube is an ideal platform for point of need learning as the content is sometimes developed and published by other experts. The content is often brief and quickly fulfills a learning need that researcher may not be able to find in formal academic literature, which takes more time to write and publish. Therefore, YouTube content was noted as a contemporary source of information and more desirable than published books or articles when seeking learning opportunities at a point of need. Much like MOOCs, these videos are free to learners and easy to find using a simple Google search.



"Quite often we would just look for a specific thing we want to learn, and then just find some videos on YouTube. That's happening more and more recently. Or sometimes you would find - for example, there was a conference, and someone provided a tutorial and recorded it, and that would be available online."

A related platform for point-of-need learning shared by researchers is discussion forums. These forums are often community driven and led by other researchers or experts working on similar issues. Stack Overflow was named as an example of these online forums. This site offers a network for individuals to ask and answer questions on technical topics where users crowd-source questions and answers for specific software. Researchers indicated that Stack Overflow and comparable sites are vital to troubleshooting specific issues related to software packages and computational coding (Stack Overflow, 2021).

Short Courses

In addition to online opportunities, synchronous short courses provided at Texas A&M University or other institutions were shared by multiple researchers as a forum for learning new skills. Short courses are typically one-to-three days of training on a specific technology related to data science and big data practices. The HPRC and TAMIDS short courses were named specifically by multiple researchers as helpful for learning about high-performance computing, cloud computing, and data manipulation using software packages such as R and python.

Short courses offered by industry partners or software companies were also a source of learning opportunities. Researchers who leverage Amazon Web Services to store big data will often attend free training offered by Amazon both virtually and in person. If researchers collaborate regularly with government agencies to perform their research, they will also take advantage of short courses or lectures offered by these groups. Esri ArcGIS was also mentioned as a source by four interviewees who use the software. Companies like Esri who create the systems and tools used by researchers are often the best instructors as they employ a host of trainers to provide general and individualized instruction to research groups using their products. Three researchers also noted that these training sessions are critical because the companies continuously upgrade software with new features. These upgrades can be difficult to understand without direct instruction from product trainers.

While most researchers pointed to the advantages of the fast, low-cost learning solutions, several shared that these sources of information are often ephemeral. A website, discussion thread, or YouTube video will be removed without warning causing researchers to seek alternatives. Moreover, general online learning does not always apply directly to their specific needs. One researcher shared that he would locate online content that addresses his need in a general way, but he must make adjustments to apply this new information to his research context. This is especially true with software instruction, which is often broad or presented through a disciplinary lens other than his own.

Online learning and in-person short courses are often core to learning about big data and data science. However, two researchers shared that they frequently perform a review of academic literature to craft research questions. The majority of those interviewed also shared that they learn about innovative applications of big data and data science at academic conferences in their primary discipline. Researchers found that information shared by their contemporaries at conferences was more recent and easily applicable to their own research questions.

Credit Courses

While brief and inexpensive learning opportunities are the dominant method expressed by researchers, several noted that they have attended credit courses in other disciplines. For example, engineers may take courses in the College of Geosciences to learn about geographic information systems (GIS) in their own research. This was especially true for those who recently graduated from a degree program. Researchers also shared that they often rely on graduate students engaged in credit courses to teach members of their labs to employ new methods. However, researchers shared that credit-bearing courses are often too expensive, too in depth, or unable to address the specific needs of a current research project.

Learning Needs

While researchers felt that their learning needs could be met in modes described above, several mentioned a need for more training on how to prepare data to be shared with other researchers. Removing and replacing sensitive data was highlighted specifically. One researcher also shared that if she needed to share her data, she would need considerable time to learn how to do so. She shared that training in this area would be a valuable support service for researchers across the University.

Another researcher shared that graduate students would need to acquire data science and big data skills, regardless of their discipline, to be marketable upon graduation. Learning at least one programming language, such as R or Python, and the ability to leverage cloud computing are now essential skills expected of graduates working in academia or the private sector. Learning to acquire and apply these skills throughout their program is the best way to form expertise, but the ability to direct students to a high-quality, low-cost, and efficient learning experience for an introduction to big data and data science would be helpful.

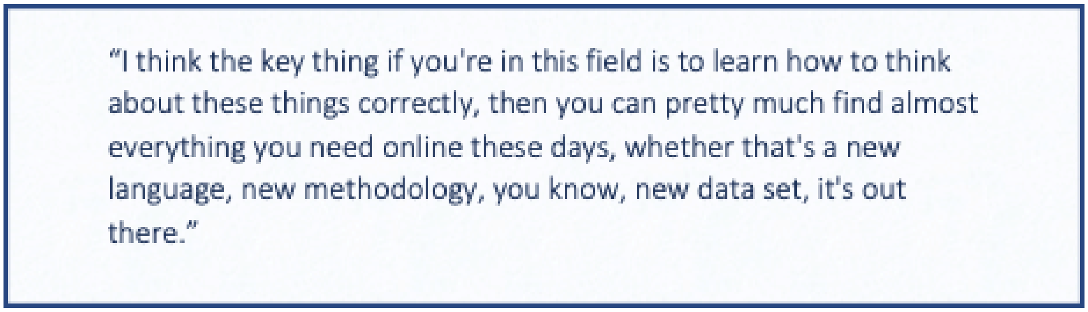
Acquiring and Utilizing Data: Challenges to Access and Quality

Six predominant themes were identified across the responses as challenges in acquiring and utilizing data—finding data, collecting data, costs and payment models, special needs for proprietary and sensitive data, data quality, and a need for wrangling.

Finding Data

While data is abundant, finding useful data is a significant challenge. Several researchers focus on collecting their own data, augmenting it with secondary data, as available. Another researcher crowdsources data collections through social media, and some researchers exclusively use secondary data. Secondary data expands research capacity, but it requires time and expertise to find and access exactly the right information needed to address a specific research question. Flexibility in establishing a research question is often required, and creative thinking is essential to finding the information needed. Quickly distinguishing useful information from what is not relevant to them is important, and should the desired data be unavailable, developing other ways to answer the same question can become necessary. One interviewee stated that they often find it necessary to reverse engineer the research question - asking what questions could be answered with the data available. Instead of asking a research question and then finding the data to answer that question, several researchers look at the data they have and then determine the questions they can answer with that data.

Finding useful data is time-consuming and researchers are likely to assign this task to their PhD students, considering it part of the education experience. At first, they might point them in the right direction, letting students learn from the experience, and build their familiarity and expertise with what is available, or they might suggest experts to approach to ask for or about data, "look at these sites, ask this person."



"I think the key thing if you're in this field is to learn how to think about these things correctly, then you can pretty much find almost everything you need online these days, whether that's a new language, new methodology, you know, new data set, it's out there."

Identifying what is needed to answer a question is a challenge. Often the exact dataset that is needed is not available, so researchers need to think creatively about how to fill in the gaps. However, researchers indicated they were eventually able to find what they needed once they were able to consider their research question within the context of the big data available.

Collecting Data

For respondents who compile their own original data, keeping track of the data (documenting and tracking though the research life cycle) was a major challenge. With graduate students moving in and out of laboratories on a regular basis, their expertise and project knowledge goes with them. For another respondent, who is using an app to crowd-source data, data collection is dependent upon engaging the users within the social media community, which is essential for adoption of the practice.

Gathering data differs for those using secondary data. For many researchers there is little issue in finding data, but they must reformat the data to make it usable. Publicly available datasets often have challenging interfaces and access is provided in a number of ways, requiring more time to learn how to work with these datasets. Several researchers seek cross-disciplinary training first. For example, one researcher sought training for collecting publicly available GIS data.

Collaboration across TAMU and the availability of University resources is a critical piece of the data collection process. This is particularly true when addressing the ability to download and store data. Researchers rely heavily on University resources for bandwidth and capacity as well as computing capacity, preferably located close to storage.

Cost of Data

Data can be expensive, and researchers, hampered by legal restrictions, tight budgets, and irregular funding, find it hard to compete with private industry to gain access to data. Additionally, data skills are in high demand and individuals with data skills can command large salaries at private companies. Several researchers expressed frustration and concern that data is available to businesses using it for profit but not researchers working toward the public good.

Predicting and budgeting for the cost of big data is problematic since costs are numerous and often unpredictable. The numerous visible costs include purchase price, downloading and storing, infrastructure (software and hardware), and time of researchers/staff. These costs are further complicated by hidden costs, which create a drain on resources – messy data, network access, employee training, complexity of data, privacy, and the rapidly changing technology and developing nature of the science (Alharthi et al., 2017). Several researchers noted that tangential or hidden costs make budgeting difficult. For example, researchers do not always know how to estimate ongoing costs for cloud computing or subscriptions to data sources.

Proprietary and sensitive data are particularly expensive, both in terms of time and resources. Time commitments include the rigorous requirements for IRB approval, setting up and practicing secure data use, compliance assurance for meeting the Health Insurance Portability and Accountability (HIPAA) regulations, private data agreements, and seeking legal advice. Other resources involve creating a secure computing and processing environment and a skilled workforce.

Data price models are a challenge for several researchers. Data is often available for purchase through a subscription model with on-going costs, while research funding, whether through start-up funds or grants, is generally episodic and unpredictable. Researchers try to minimize ongoing costs, since they are harder to fund through grants.

Researchers identified several ways in which they are addressing these challenges. Many indicated they make the most of University and departmental resources, such as utilizing the HPRC for processing and storing data and as well as using research support services to purchase data subscriptions and floating software licenses, which can then be shared across disciplines and labs. One researcher noted if the most recent data, which comes at a premium, was not necessary, they could save money purchasing an older set of data.

Several interviewees indicated that building relationships with data providers helps them negotiate the price for data or even allows them access to data free of charge. As members of the

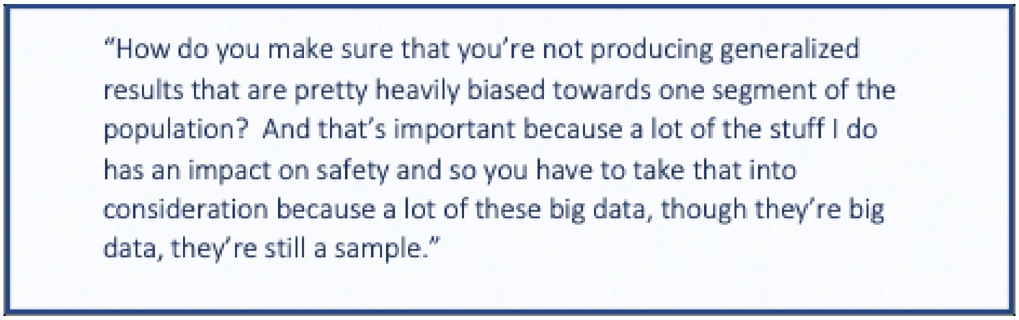
TAMU research community, they can establish relationships with data providers paying only for the cost of transferring data. Another researcher utilizes open-source software, rather than paying for a proprietary software subscription.

Data Quality

Several researchers reported that, while data is increasingly available, they see a lot of “bad” data and data quality is their biggest challenge. Poor-quality data makes the handling of large datasets even more difficult - inaccurate data introduces uncertainties, while inconsistent or messy data prolongs the processing time.

When using secondary data, many researchers noted that the context of the data is critical to understanding the data’s veracity and bias. However, information regarding collection or processing is often unavailable, leaving many unanswered questions. How is it collected? Who is collecting it? Who are the subjects? What is their age/gender/race distribution? What are the uncontrolled variables? Is there inherent bias? Difficulty in understanding the context of secondary data was a common issue experienced across most researchers.

For one interviewee, using social media data highlights the risk of bias and misrepresentation. Another researcher finds most of their data online, but it may not have metadata, forcing them to make their own determinations. For another researcher working with industry, the changes in technology happen so fast that they are difficult to keep up with, so maintaining a conversation with industry partners is critical to understanding the changes both in data collection and in the information needed by industry. Complicating all of this is the nature of big data as a new field of study; researchers are building new tools to answer new questions, and thus, there is often no base line of historical data or results against which results can be compared.



"How do you make sure that you're not producing generalized results that are pretty heavily biased towards one segment of the population? And that's important because a lot of the stuff I do has an impact on safety and so you have to take that into consideration because a lot of these big data, though they're big data, they're still a sample."

Good quality data is clean and standardized. Unfortunately, data does not come out of the box ready to use, particularly when coming from multiple sources. Data may come in many formats, use inconsistent terminology, and have metadata with varying levels of granularity. Many of the researchers spent considerable time, 30- 50%, wrangling data - manipulating, preprocessing, cleaning, organizing, integrating their data - to prepare it for analysis.

Sensitive Data

Sensitive data is any data that needs to be protected. It includes data that can be used to identify an individual, financial data, or intellectual property. Proprietary data may be generated and held by private industry protecting their investment or public entities that need to protect it from malicious parties. It is often quite hard to acquire and comes with strict data-sharing agreements, stipulating that the data will not be shared. Researchers reported that they take abundant precautions to protect this data and expressed concerns about mandates to share data.

Several researchers identified the special accommodations necessary when working with sensitive or proprietary data as a challenge. Sensitive data often must be anonymized or deidentified to be made publicly available. This can make the data unusable, or researchers are left scrambling to find the missing details by tracking down authors and requesting the data directly. Laws governing sensitive medical data are increasingly stringent and access to data is highly restricted. Researchers expressed frustration around the different standards for private companies who collect personal data via apps, cookies, or surveillance to use for profit, while academic researchers, using the same information for public good, must follow IRB rules and are sometimes not allowed access.

Security needs make proprietary and sensitive data much more time consuming to maintain. One researcher indicated they spend approximately half their time on the infrastructure (hardware, software) and data governance side of their research. Researchers rely on information technology specialists to meet the security standards and protect proprietary or sensitive data and secure storage and processing systems.

Some researchers identified synthetic data (an artificial copy of original data that represents the authentic data and ideally shares the same and modeling) as one way around the issues surrounding sensitive datasets, although others did not find them useful. For one researcher, synthetic data are increasingly available as awareness of what they are, and their use is more commonly understood. Having publicly available dataset will help with replicability and reproducibility.

Sharing and Publishing Big Data: Current Practices and Challenges

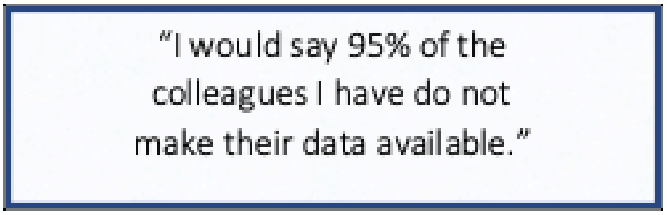
When speaking about sharing big data, topics such as collaboration, storage, security, and dissemination were discussed by researchers. Almost all researchers (9 out of twelve) use some type of service to share their data internally among their research team or externally, when collaborating with teams outside of the University. GitHub and Google Drive are used by several researchers while others use cloud-based systems such as the one offered by Texas A&M’s HPRC and Amazon Web, or in-house servers to facilitate collaboration. Cloud servers allow some teams to work together in a way that one team member can finish their portion of the work, then pass it off to the next team member. Local servers have the benefit of keeping the information inside the building, which provides more security for the data. However, one researcher mentioned that they do not share their data with their research team since each member has a different task.

Although most researchers primarily discussed sharing data between members in the University or other teams in the United States, one interviewee mentioned the difficulty of sharing data between two countries under the official data sharing agreement. For them, the easiest way was to publish the data so that the other team could easily find the publication and the data associated with it. Disseminating the results of their data and research was also brought up by seven other researchers. Since several of the researchers are sponsored by non-academic institutes, they either share the results of their research in the form of reports or do not publish them at all. Other participants disseminate their research through conference presentations, published articles, webinars, or arXiv.

To share their research, some of the interviewees create their own webpage to upload and publish their datasets, codes, or models. These webpages are sometimes personal websites but often maintained by the University, such as laboratory websites. Marketing and communication units of the departments or colleges help with presenting the research through social media. Making the datasets or models open source or including the DOI in the published articles are other ways of sharing the data with other researchers in the field.

However, sharing data through the mediums shared above comes with various concerns for the researchers. Publishing papers or datasets on arXiv does not count towards the tenure process of the faculty, and therefore, the incentive is low for researchers to publish that way. It is also unclear for some faculty whether they can show the impact of their data, as they were unsure about tracking citations to see if their data has been used in other papers or policies.

Other challenges with sharing big data among group members or when publishing the data are related to security or licensing issues. Working with high-security data such as critical electrical infrastructure information makes it difficult to collaborate with different members as some team members have access to the data while others do not. Due to their contract with the institute that provides the data, some research teams are only able to share the derived data externally which makes collaboration difficult with members outside of the University. Not having common terminology when sharing data for re-use among larger teams (across the country) was also identified as a challenge.



"I would say 95% of the colleagues I have do not make their data available."

Some departments at the University still lack necessary tools such as enough storage space for big data. Maintaining websites for sharing data is often cumbersome and they too lack adequate space and capacity for big data. To address some of the issues, several researchers ask their collaborators to store the data. Others only provide data upon request. Only one participant requested help from the University Libraries, and several did not ask for help at all when sharing their data. As research funding agencies and academic journals place emphasis on sharing research data, researchers note that sharing data will become more normalized.

Conclusion

It was evident from these interviews that the Libraries can play a strategic role in supporting big data research. Researchers expressed that there is a sharp learning curve for those who wish to leverage data science and big data in their research. Specifically, they found it challenging to keep up to date with new methods and current research; locate collaborators with data science skills; identify learning opportunities that meet specific needs and criteria; find high-quality and affordable data; prepare disparate datasets for analysis; prepare data for sharing; store and share sensitive data. The following are potential areas where the Libraries can impact big data research.

The University provides several of the services identified in the interviews. However, it was clear that many researchers were unaware of these services. The Libraries can place additional emphasis on the visibility of these services. In the future, it will be a key priority to promote existing services and adapt these services to evolving needs of big data researchers. The following recommendations seek to mitigate the barriers addressed by the researchers and raise awareness of resources across the TAMU community.

Recommendations

Provide Targeted Training

The University can provide online or in-person courses free of charge to teach introductory concepts and theories regarding big data as well as specific software skills for cleaning, manipulating, analyzing, and publishing data. Education on how to store, archive, and share sensitive data was a determined need for researchers working with personally identifiable data.

The University Libraries currently offer a suite of data management workshops for all University students, faculty, and staff, but researchers expressed a need for additional learning opportunities on big data topics in the form of short courses or workshops. In step with the themes that emerged, these courses should be free of charge and teach both introductory concepts and theories regarding big data as well as specific skills for using software to clean, manipulate, analyze, and publish data. To that end, the Libraries have identified seven librarians to undergo instructor certification for The Carpentries.

The Carpentries “builds global capacity in essential data and computational skills for conducting efficient, open, and reproducible research” (The Carpentries, 2021). Once certified, instructors can teach Carpentries workshops covering a host of topics including data cleaning, visualization, and analysis. Workshops also include instruction in several key software applications for big data methods noted by researchers, including R and Python. The Libraries should seek opportunities to plan and host workshops in partnership with other campus units such as the TAMIDS and the HPRC center.

Incentivize Data Sharing

Incentive and privacy were noted as significant barriers to sharing data among the interviewees. Many indicated that their disciplines and the University did not provide significant incentives for sharing data as a product of research. The Libraries could serve as a leader in incentivization by offering a suite of services and resources. Incentives could take the form of funds provided by the Libraries to encourage data publication for faculty, an open access data award for graduate students who deposit data in the Libraries data repository. Additionally, the Libraries can feature big data researchers who currently deposit data in newsletters and on the Libraries website.

The Libraries and campus partners can also support researchers who use sensitive data by offering resources and instruction on best practice for de-identifying data. This should be offered in partnership with members of the Division of Research and the Institutional Review Board.

Collect Big Data

Finding big data that is accessible and accurate was a consistent theme throughout the interviews. Subscriptions to some proprietary data are inevitable, but researchers are unable to sustain long-term costs due to the temporary nature of external grant funding. There are opportunities for the Libraries to support the purchase of big datasets that have potential value to multiple researchers across campus. Moreover, the Libraries can identify high-use, open datasets and provide some preliminary cleaning and managing of the data to expedite analysis by researchers. For example, regional GIS data currently held by the Libraries can be organized, cleaned, and stored for immediate use by researchers in several disciplines. This service would be especially useful to graduate students seeking big data projects for their dissertations and theses.

Collaborate with Campus Partners to Provide Support

General Support

Resources to support big data are available across the University, but a central group to help identify, define, and communicate these resources would be an asset to those working with big data. Learning opportunities, storage, analysis, and policy for big data can be organized and codified at a single point of contact. Many institutions have formed a distinct group comprised of data support professionals who provide a single point of contact for researchers seeking guidance throughout the research lifecycle from planning to data publication. Harvard University has a model group that provides comprehensive data management support across their institution (Harvard University, n.d.). Some potential examples at TAMU are working with the IRB as a place to house help or guidance for sharing big data as part of the IRB protocol. Collaboration across the University is especially important when using proprietary and or sensitive data. Researchers would benefit from having a dedicated specialist in information technology to meet security standards and protect proprietary or sensitive data and secure storage and processing systems.

Proposal Guidance

Related to campus partnerships, it would benefit researchers if the Division of Research and the Libraries could provide additional guidance on drafting grant proposals that include a budget for proprietary, big datasets. For example, researchers do not always know how to estimate ongoing costs for cloud computing or subscriptions to data sources.

References

Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons*, *60*(3), 285–292. https://doi.org/10.1016/j.bushor.2017.01.002

The Carpentries. (2021) *About us*. https://perma.cc/QR23-YE2G

Coursera. (2021). *Our vision.* https://perma.cc/B87U-L5X8

Go FAIR. (n.d.) *Fair principles*. https://perma.cc/GJ56-4ELA

Harvard University. (n.d.). *Research data management @Harvard.* https://perma.cc/DR7L-4LUG

Texas A&M University. (2021). *Faculty and staff*. https://perma.cc/6RMM-VEJT

Texas A&M University. (n.d.). *About Texas A&M.* https://perma.cc/9EZ5-KL9P

Texas A&M University Accountability. (n.d.). *Recognitions*. https://perma.cc/QF67-56U8

Texas A&M University High Performance Research Computing. (n.d.). *About High Performance Research Computing*. https://perma.cc/U5QM-ET54

Texas A&M University Institute of Data Science. (n.d.). *Institute of Data Science*. https://perma.cc/MGP2-95UE

Stack Overflow. (2021). *Who we are*. https://perma.cc/J4FH-8T7K

Appendix A

Recruitment Email 1

Subject. Texas A&M University Libraries study on supporting big data research

Dear [*first name of researcher*],

Members of the University Libraries are conducting a study on the practices of researchers who use big data or data science methods to improve services for their work. Are you willing to participate in a one-hour interview to share your experiences and perspective?

Our local 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 Texas A&M will also be included in a report by Ithaka S+R and will be essential for Texas A&M to further understand how the support needs of big data researchers are evolving.

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

Follow-up Email

Subject. Texas A&M University Library study on supporting big data research

I hope your week is going well. I know this is a busy time, but I wanted to follow up and see if you are available to meet with me to discuss your research. We will need about 60 minutes to conduct the interview via Zoom. Let me know what you think and thank you for considering.

Best,

[name of investigator]

Appendix C

Recruitment Email 2

Dear [*first name of researcher*],

Thank you for expressing your interest in participating in this study. I would love to set up a time to interview you at your convenience. Please advise me of your availability in [*time frame*].

Before the interview begins, I will ask you to provide verbal consent to ensure that you understand the study and are willing to participate. I am attaching the verbal consent *protocol* to this email in case you'd like to look over before the interview.

Sincerely,

[*name of investigator listed on this protocol*]

Appendix D

Consent Form

Title of Research Study: Supporting Big Data Research

Investigator: John Watts, Laura Sare, Carolyn Jackson, Paria Tajalli pour

Why am I being asked to take part in this research study?

You are invited to participate in this study because we are trying to learn more about researchers working with big data. You were selected as a possible participant in this study because you conduct data-intensive research. You must be 18 years of age or older to participate.

Why is this research being done?

This study is an exploratory examination of the research practices of faculty and research staff in a variety of humanities, social science, and STEM fields who utilize data science or "big data" methodologies. The goal of the study is to understand researchers' processes in working with big data toward developing resources and services at Texas A&M University to support them in their work. The study at Texas A&M University is connected to a suite of parallel studies being developed locally at other higher education institutions. Ithaka S+R, a not-for-profit research and consulting organization that helps the academic, cultural, and publishing communities.

How long will the research last?

It will take about sixty minutes

What happens if I say "Yes, I want to be in this research"?

If you decide to participate, your participation in the study involves a sixty-minute, audio-recorded interview about your research practices via Zoom. Audio recordings will be transcribed by a third-party transcription vendor bound by a non-disclosure agreement. The investigators will not record video during this interview. Audio recording files will be destroyed immediately following transcription. Pseudonyms will be immediately applied to the interview transcripts and the metadata associated with the transcripts.

What happens if I do not want to be in this research?

Your participation in this study is voluntary. You can decide not to participate in this research and it will not be held against you. You can leave the study at any time.

Is there any way being in this study could harm me?

There are no sensitive questions in this survey that should cause discomfort. However, you can skip any question you do not wish to answer or stop the interview at any point.

What happens to the information collected for the research?

All identifiable information will be kept on a password protected computer and is only accessible by the research team. Compliance offices at Texas A&M may be given access to the study files upon request. Your information will be kept confidential to the extent allowed by law. The results of the research study may be published but your identity will remain confidential.

Who can I talk to?

Please feel free to ask questions regarding this study. You may the investigators later if you have additional questions or concerns at

John Watts: jwatts@tamu.edu, 979-458-6491

Laura Sare: lsare@tamu.edu, 979-458-2200

Carolyn Jackson: csj@library.tamu.edu, 979-458-0315

Paria Tajalli pour: paria@library.tamu.edu, 979-862-4321

You may also contact the Human Research Protection Program at Texas A&M University (which is a group of people who review the research to protect your rights) by phone at 1-979-458-4067, toll free at 1-855-795-8636, or by email at irb@tamu.edu for:

- additional help with any questions about the research
- voicing concerns or complaints about the research
- obtaining answers to questions about your rights as a research participant
- concerns in the event the research staff could not be reached
- the desire to talk to someone other than the research staff

Do you agree to this interview?

Do you agree to being recorded (audio only) for this interview?