Supporting Big Data Research at the University of Massachusetts Amherst

Item Type	article;article
Authors	Atwood, Thea P.;Radik, Melanie;Seifried, Rebecca M.
DOI	https://doi.org/10.7275/jsr0-2n21
Download date	2024-08-15 09:08:43
Item License	http://creativecommons.org/licenses/by/4.0/
Link to Item	https://hdl.handle.net/20.500.14394/32361

Supporting Big Data Research at the University of Massachusetts Amherst

August 24, 2021

Thea Atwood, Data Services Librarian
Melanie Radik, Science and Engineering Librarian
Rebecca M. Seifried, Geospatial Information Librarian

Table of Contents

Background

The ITHAKA S+R Research Project

This project aimed to examine the research support needs of faculty who employ "big data" and data science methodologies at the University of Massachusetts Amherst. The study was conducted by the University Libraries and was part of a larger suite of parallel studies of big data researchers at institutions of higher education across the U.S. The study was coordinated by Ithaka S+R whose goal is to "help academic and cultural communities know what is coming next, learn from rigorous and well-designed research studies, and … improve their performance and further their missions." Under the guidance of project coordinators from Ithaka S+R, librarians at 21 participating institutions—including Boston University, Northeastern, Texas A&M, and several University of California campuses—interviewed researchers across our campuses and compiled independent research results and recommendations for creating or enhancing local services and supports. In addition, participating institutions contributed their findings to a final capstone report by Ithaka S+R. The Ithaka capstone report provides a cumulative view of the evolving needs of big data researchers and includes recommendations for how the Libraries and campus research support structures can most effectively and strategically grow our support for this rapidly expanding area of research needs.

Big Data Research and Support at UMass Amherst

While it might be unexpected, researchers in a majority of schools and colleges across campus employ big data and data science methodologies. Most people would accurately identify the College of Information and Computer Sciences as a hub of big data research, and many could name the College of Engineering or think of departments in the College of Natural Sciences—such as Astronomy or Biochemistry and Molecular Biology—as active in big data research. But our study also identified researchers in the School of Public Health and Health Sciences, the Isenberg School of Management, and the College of Social and Behavioral Sciences, as well as their students in the Graduate School, who are active in big data research. We further noted some activity among researchers who are not currently involved in big data research but may become so in the near future, such as the Digital Humanities Initiative.

Big data research is an area that the campus has been interested in developing for several years. As a 2015 President's Office Report states, UMass Amherst has:

> *world-leading strengths in three key complementary areas: (1) building **infrastructure** to access and process Big Data efficiently and effectively; (2) developing machine learning **analytics** that construct statistical models and make accurate, meaningful predictions; and (3) creating **applications** in a wide range of fields, from fighting terrorism to finding cures for diseases.*[1]

[1] UMass Data Science Faculty Working Group, "UMass Big Data: Data Science for the Commonwealth Powered by the University of Massachusetts," 2015, p. 3, https://archives.lib.state.ma.us/handle/2452/430344 (accessed 24 August 2021).

Departmental initiatives to foster big data research include the Center for Data Science, which was established by the College of Information and Computer Science in 2016 to "be an international leader and destination-of-choice for education, research and industrial collaboration in data science,"[2] and which provides consultation, education, and access to computational tools that researchers in our community might not otherwise have. The Department of Mathematics and Statistics' Statistical Consulting & Collaboration Services "provides statistical consulting and collaboration to researchers at UMass, the Five Colleges, and elsewhere, through individual meetings and support for interdisciplinary research projects. SCCS can help with all phases of statistical analysis." And the research cores of the Institute for Applied Life Sciences explicitly strive to meet the challenges of big data and innovative research through such facilities as the Genomics Resource Laboratory. Each of these resources is available to all researchers on campus, but they are all also fee-based services.

Campus-level support is provided primarily by the dedicated work of Research Computing in the campus' IT department. Research Computing mediates the use of the Massachusetts Green High Performance Computing Center (MGHPCC), and has recently spun up a cluster of servers called Unity that is available to all researchers at UMass Amherst. They provide consultation and troubleshooting services to researchers and have previously provided or facilitated training, although offerings in recent years have been few. Some individual departments employ departmental IT support personnel, but their level of knowledge and expertise in software and hardware solutions for big data research varies widely. Individual departments or schools may also maintain servers and software subscriptions that serve big data research needs. The Libraries employ a Data Services Librarian whose purview is more in data management training and consultation than big data or data science methodologies setup or support. ScholarWorks, the University's institutional data repository, has recently faced its first challenge of a multi-terabyte dataset, and while a solution was found, it required extensive workarounds of the current infrastructure.

Findings

Overview

Our interviews demonstrate that big data researchers at UMass Amherst utilize a wide variety of workflows and resources to conduct their work (Figure 1). For example, some begin by testing analyses on smaller subsets of data: they scale their work down before they scale it up to larger, high performance computing infrastructure that has a higher overhead. Others leverage existing pipelines, such as those provided by the Illumina genomics sequencing system, which allows for a selected series of sequencing and analysis steps to run on a dataset. Some are juggling the complexity of Health Insurance

[2] University of Massachusetts Amherst Faculty Senate, "Special Report of the Academic Priorities, Program and Budget and Research Councils concerning the Establishment of a Center for Data Science (CDS) in the College of Information and Computer Sciences (Sen. Doc. No. 16-075)," 2016, p. 2, https://www.umass.edu/senate/sites/default/files/APC%20PBC%20RESEARCH%20CENTER%20FOR%20DATA%20SCIENCE%20-%20Sen.%20Doc.%20No.%2016-075.pdf (accessed 24 August 2021).

Portability and Accountability Act (HIPAA) data in addition to non-sensitive data, balancing data in two different environments with different security regulations and requirements.

The tools that researchers use are split between proprietary, often-expensive, software that is licensed at the research group or campus level and free open-source packages. Many mentioned Jupyter Notebooks, both as an entry point and as a teaching tool, as well as GitHub as ways to manage their code. Others mentioned multiple steps across different platforms and machines: their needs for cleaning, data analysis, and discipline-specific analysis required the ability to shift between SAS, Stata, and R, or perhaps Python, NVivo, and Excel. One researcher commented that the programming languages needed to conduct their research could run the gamut: "It's everything, starting from C and C++ all the way up to Python. Whatever comes in handy, right?" Common across all of our discussions was the role of iteration: our interviewees explicitly mentioned the need to continually refine code and to be flexible in their workflows.

Figure 1. Researchers use multiple resources in a single project.

We saw echoes of this flexibility—"whatever comes in handy"—in researchers' scholarly communication practices. While most pursue traditional academic journals and professional society conferences as venues for sharing their research, many actively share in non-academic venues such as Twitter and *The Conversation*. Some of this reflects an active investment in communicating science to the general public, while others noted that it was becoming a norm in their discipline. However, researchers who are doing more applied work expressed some frustration with the expectations of the traditional "publish or perish" model, since their stakeholders are much more interested in project reports than in peer-reviewed articles, and yet the expectations of tenure review committees remain traditionally focused. Many expressed a wish for incentives within their department or at the university level that would tangibly recognize the value of their work to communicate science to the general public, carry out applied research and translational entrepreneurship, foster interdisciplinary research, and contribute to open-source coding and open-access publishing.

Collecting Data

Researchers who generate original data do so in various ways, including crowdsourcing, sequencing data, and gathering social media data. Their most common challenges revolve around survey and sampling methodologies and elements of fundamental research design, such as ensuring their instruments are designed well, rather than any shortcomings of campus infrastructure or support.

Researchers who use datasets generated by others (secondary data) rely on a variety of venues and methods to locate and acquire their data, including combining datasets from multiple single cases in order to make a much larger, comprehensive dataset; accessing datasets that are behind a paywall; and using curated datasets that are publicly available. Their key challenges often have to do with how relevant the secondary datasets are and how much work is needed to prepare them for re-use. They must decide: Do they really want to use this data? Is there something better out there? Is it well described? Does the dataset actually contain what it says it contains? One interviewee commented that even if a data dictionary is included, it isn't always helpful, especially if a piece of the dictionary is missing and they need to go back to the source. This slows down their work and can sometimes lead to a project being dropped.

This is a piece of the cultural challenge facing researchers today: the path to data reuse is extremely challenging. There are few opportunities for education about the FAIR principles (to make data Findable, Accessible, Interoperable, and Reusable) or the concepts that underlie them, and there are also few incentives to make data align with those principles. For example, metadata is a critical component of the FAIR principles, but it is at such a nascent stage in many fields that there is little guidance or incentive to adopt a new and time-consuming practice that may or may not pan out. Adopting better approaches to metadata would require a huge cultural shift that would change how our scholars are educated from the bottom up. And the state of their data is at times out of their control; for example, not all agencies who provide access to data need to report the same things, or over time they are no longer beholden to previous reporting measures. The data degrades over time. Sometimes, the data simply does not exist in large enough quantities for scholars to ask their questions, especially if they are reusing specific sub-segments of existing data.

Those who rely on data from state agencies, national agencies, and business and industry note particular barriers to working with third-party data, including guidelines about where data can be stored, variation in data fields over time, and understanding the nuances of a Data Use Agreement (DUA), as well as how long it can take to establish a DUA (often it is a years-long process). HIPAA data can be particularly challenging due to the sensitive nature of this data. As one interviewee remarked:

> *To my knowledge there are no healthcare organizations that have a process in place where an outside person can come in and say, "I'm interested in this question; can we solve it?" It's a very individual person, relationship-building process. It's very painful and very slow and very inefficient, but it is the way that it is with HIPAA.*

Researchers also commented on the assistance they receive in finding or accessing datasets. Many researchers receive little or no help—perhaps they have an undergraduate or graduate student, a

collaborator, or a consultant, but little else. One researcher was a 'gatekeeper' to data that is under layers of regulations: students and others who were not the PIs on the project did not have access to the data, so they as 'gatekeeper' were responsible for querying the data on their behalf, thus taking time away from their research. And even if a student is available, it is sometimes for a limited term:

> *This semester I actually have an undergraduate research assistant helping me … that has been somewhat, a little bit helpful.... Just took a little bit of burden off, but other than that it's mostly been myself doing it.*

Some have slightly more help, which consists of collaborators across disciplines and labs:

> *I have a collaborator in Biochemistry and Molecular Biology and we're both funded on my grant. And I have a graduate student and his lab and I have a graduate student in my lab.*

Others actively involve graduate students in the data-collection process, tasking them with finding relevant data for the study they are working on. This can provide the students with hands-on experience writing querying scripts, assessing data quality, and revising the scripts to refine the collected data. Data collection may be even viewed as a PhD-level training exercise.

None of our participants explicitly mentioned the Libraries as providing assistance with data collection. While this may be because we had a relatively small pool of interview participants, this is also not totally unexpected. Those working with big data have long had to fend for themselves, and they may not consider the library—or campus resources more broadly—as a logical place for help. At the same time, our Libraries face staffing limitations that would make expanding our support in this arena particularly challenging.

CHALLENGES IN COLLECTING DATA

Not all researchers need help finding or accessing data; many have the necessary skills to access the data themselves by writing scripts to query large databases. Having control over the data-collection process is an important role for some big data researchers: one described themself as a "bridge" between the data provider and the research question they are trying to answer:

> *That's often my role to be the bridge between the folks that provide … a computer science technical expertise. I've played the role of trying to figure out what data can answer the question that I think is interesting.*

This being said, many researchers highlighted concerns and challenges around data generation, collection, and use. Broadly, their concerns can be distilled into 6 areas: **lack of training**, **quality and quantity of data**, **infrastructure and services**, **regulations and administrative burden**, **cost**, and **data security**.

Lack of Training

Researchers across the sciences noted that they need access to a specialist, even if the data already exists. This can be difficult when there are not enough specialists to meet demand, or when relationships dissipate after a few months' time. Relatedly, several researchers cited a lack of coding skills (their own, as well as among their colleagues and students) as a major barrier to accessing and querying large datasets. One researcher regularly collaborates with a bioinformatician to write codes that can retrieve the data, while another uses Python scripts that someone else wrote for them. Others are self-taught, and often turn to Twitter or Google to get pointers. These individuals often expressed a desire to learn more and develop their coding skills, but time and energy are barriers that prevent them from investing in this area.

> *But for the coding part, I need of course a lot of help. And I'm not alone. I think that most biological labs in UMass now experience that significant deficiency in [the] support and training of students, [to] be able to analyze biological data using software.*

The key for these individuals is to establish a "bread and butter tool setup" that they can reuse or tweak as necessary, without needing to put in the significant investment that would be required to develop their own coding skills. Furthermore, if individuals in a field like Biology are experiencing difficulty in setting up reusable workflows, those in other fields that lack access to outside funding may be feeling these challenges even more acutely, as they may lack access to any funding or support streams to help remedy this challenge.

Even when researchers try to expand their skill set, they face barriers. Some scholars tried to take a graduate-level course in coding, only to find that enrollment in the course was limited to those within the department (i.e. they lacked the appropriate department affiliation). Others explored computing and programming workshops and peer discussion/collaboration groups in their own or related departments. A consistent issue was the gap between basic skills workshops and discipline-specific peer groups; it was unclear where a researcher could go to develop intermediate skills and expand their foundational knowledge before diving into highly-specific discussions:

> *I tried to join the R studies group in the medical biology program, which was organized by graduate students. But I realized that my level is not on par with them. Because they already discuss some—I don't know—technical difficulties, while I did not even know how to start.*

Quality and Quantity of Data

Several interviewees commented on data not being a perfect fit for their work. Perhaps the data is not in the correct format, so it needs to be modified with a script before in order to be interoperable with other datasets, or it is difficult to parse out the noise in a dataset, or the data contains mistakes. It could be that there are a series of unknowns they are facing. For example, private companies like Twitter or Facebook can be a bit of a black box — they can remove APIs that researchers relied upon previously, since private companies are under no obligations to provide continuous access to data. Researchers seem to be at the whim of industry: they noted that while their hands may be tied at a later point in

time, for now, the access they have is 'good enough.' One researcher explicitly mentioned that data dictionaries are not even a cure-all. Data dictionaries do not always contain the relevant information necessary to use a particular dataset; variables or other pieces of information are sometimes missing, which requires tracking down the original data generator.

Having to clean and prepare the data before analysis is a major challenge that researchers face, whether they are generating their own data or reusing secondary datasets. When working with people in the real world, the inevitable variables that come from human-generated processes create imperfect data. These range from incentives to completing a task to individuals missing appointments to take a survey. Researchers work to mitigate these issues by properly designing tasks for crowdsourced workers, or by treating research participants as partners in the research process and working to give back to them. When working with secondary datasets, variability in data formats is a common phenomenon; when a multitude of pieces and types of equipment are generating data in different ways, the variation needs to be accounted for in order to make one synergistic dataset.

Organizing data is a perpetual challenge because researchers often lack a systemic way to label and organize their data. Some mentioned that their data is so complex that no structured database would make sense, or that some available solutions—like a MySQL database—are prohibitively time-intensive to implement before even beginning to work with the data. Some mentioned a lack of guidance about organizing data as a pain point: they have searched for guidance, but none seems to exist. One researcher commented:

> *I would reiterate that the biggest challenge is organizing the data that I collect and that I acquire through public resources for my research. I don't know how to do it and I don't know anybody that has looked. I've searched all the sources I could find and I cannot find an answer to that problem and I feel like that would be a huge amplifier for my research because it would enable me to develop new methods for these types of datasets and test them and make new biological discoveries.*

Additional challenges related to quality and quantity of data include the time component—it can take a great deal of time to crawl very large datasets for answers or to perfect an incredibly refined query in order to extract a high-quality dataset—and the fact that sometimes there simply is not enough data available yet, especially when trying to dig into subsets of populations or certain parameters for thousands of chemicals that are not consistently listed.

Infrastructure and Services

Many researchers were able to describe computing and analytics services for hire on campus, such as the Center for Data Science, the Statistical Consulting & Collaboration Services, and even the Center for Research on Families. Most could recall support that used to be offered to those using MGHPCC's services. But the majority were vague on details, and several admitted that they wouldn't know quite where to start when deciding which service could meet their needs or how to get started using one, and they expressed a wish for a centralized service point:

> *It could be useful to have the kind of centralized place at UMass that would help people connect to the resources that do exist. So, for instance, I could imagine if I had more computationally intensive work in the future, I know they have the [Massachusetts Green High Performance Computing Center], but I wouldn't necessarily know where to begin using it. And if there was some kind of central place that had some consultants that know about computer science and know about statistics that could help with that, that could help me understand things like how do I interact—I've found that always a little bit confusing…. I think that would help a lot for the day-to-day stuff.*

One researcher noted the tension between maintaining hardware and funders not supporting hardware maintenance or upgrades, even if this hardware is important to the project. This is compounded when the group lacks dedicated IT support. This issue can tie into too much data—when scholars need server space that is secure but also has the computational power they need. Other researchers comment that finding the data and hosting the data are two separate problems: they can find a dataset easily but need to pay for the hosting and computing resources, such as with Amazon Web Services (AWS), so it becomes a variable cost they need to factor into their research program. And while the MGHPCC is available, it is not considered a catch-all—while they can offer computing and storage resources, high-performance computing centers do not have a universal configuration, so there are times when outside computing centers need to be consulted.

There is also a cultural piece to the issue of infrastructure. Another researcher commented on doing work with colleagues in parallel: they may store documentation in a secure storage solution, like Box, but adopting other technology is slow or even impossible—while the researcher finds Jira Confluence and Jupyter Notebooks to be beneficial, they are not being adopted by other collaborators and so they end up working separately.

One researcher specifically pointed out that the infrastructure piece is a huge challenge that needs to be undertaken by the university. In other words, it is a university-wide problem that, taken on by individual colleges, leads to the duplication of effort and even the duplication of costly audits. One researcher observed: "It's not silos exactly, it's like colleges can't figure out how to work together."

<u>Regulations and Administrative Burden</u>

While researchers note that they are important, security regulations are a barrier to research. This ties in with the infrastructure issue, in some ways: if a university does not have the appropriate infrastructure, data cannot be hosted at the university. Data must either be worked on through a mediated source or through mediated means, meaning scholars may not have full access to a dataset or they may be a sole gatekeeper (i.e. they must run queries on the full dataset on behalf of others). There is also a great deal of power with those that hold the HIPAA data—the organization can say no to a researcher if they wish. One researcher commented that "HIPAA has cast this incredibly long shadow over an organization who has the sort of potential to violate HIPAA or to be penalized under HIPAA protection." Another researcher commented on their inflexibility, which can limit what questions they ask:

> *So on our data, people that are charged with protecting UMass's liability in getting HIPAA data there all these … mental barriers, but also these—it's just almost lack of willingness to go and follow the letter of the law…. If there's any possibility that this could be interpreted in a different way, they're not willing to sign off and facilitate this research project.… There are data costs to standardize datasets that are often made barriers.... The biggest barriers are just the inefficient layers that are required to get—in my view—interesting data…. It goes beyond "let's look at some canned data and do this again."*

Other researchers—in particular, those that were not working with highly sensitive data—felt confident they were following appropriate protocols.

Related to regulations is administrative burden, including taking the necessary precautions to receive data, such as data applications that are reviewed by several committees. It can take up to a year or longer to complete this process. This is also true in collaborations, with researchers citing drawn-out processes in order to align workflows between different universities. Several mentioned how arduous the process can be to acquire their data in the first place.

Cost

Another theme throughout the researchers' narratives was cost. Cost comes into play in a variety of ways: as time, hardware upkeep, or even just the up-front expense of the datasets. One researcher commented that the most current data is expensive to obtain, and that sometimes the frequency of the data output does not match what is purchased (e.g. data comes out daily but it is purchased on a weekly basis, so the most current data is not achievable).

Data Security

Sometimes researchers don't know where to put their very large datasets—they need secure space as well as speed. Data security focuses on how scholars keep their data managed in such a way that unauthorized access or use is prevented. Many researchers commented that this is a topic they take into consideration, either because they have to (e.g. HIPAA requirements, data use agreements) or because it is a good practice (e.g. contacting another researcher and commenting that the data will only be used for research, following IRB and other data security best practice protocols). Some scholars mention they do not have personal data to secure but highlight that they are still concerned about data breaches, especially as they are conducting research.

However, others mention it is a gray area or admit to not fully understanding data security. For example, one scholar mentioned they don't understand exactly how the mechanisms for sharing data works, just that their data use agreement states that their servers should not be connected to the internet. Researchers in this position admit they must lean on the technical and IT experts in their area.

Analyzing Data

Scholars rely on a variety of methods for data analysis, including traditional regression analysis, gene sequencing, income distribution, demographic analysis, social network analysis, simulations, and text analysis. One researcher explicitly mentioned the importance of visualizing their data ahead of analysis to help them determine if there are any outliers or unexpected variations in their data. Most commented on the importance of coding scripts to perform their analyses.

While we do not go into detail about their particular processes, throughout our interviews, researchers highlighted their concerns and challenges. One of the major challenges they face has to do with preparing data for analysis. As with data collection, imperfect data can make data analysis challenging. Some interviewees cited that there is a great deal of work that goes into preparing a dataset for analysis. Several mentioned the bottleneck that is created by limited staffing for data cleaning and analysis or limited availability of consultants, which in turn can limit their ability to ask questions. One researcher took this a step further, commenting that even if they could find trained individuals, because they are working with high security data, there is a layer of vetting involved—even honest mistakes can end a career.

Another major challenge researchers face in analyzing their data revolves around the variety of methods that are available and a lack of cross-training. In particular, researchers pointed to the lack of a strong statistical background as a challenge—either their own as it relates to the work they are now doing, or their current students as it relates to the current and future work. Many mentioned the variety of methods that are available and the need to pair the right methods with the right data:

> *So there are probably 20 different ways to run a particular statistical regression and 20 ways that might actually be perfectly valid and useful for answering different questions, but the process of adjudicating which of those is going to be best for your given situation is, I think, something that I've always found to be challenging…. Pairing the right method to go with the data that you have available and trying to communicate it to an audience that might not be very familiar with those methods is, I think, kind of the generic challenge that … a lot of us face.*

This ambiguity could also be due in part to some uncertainty about big data as an evolving field:

> *I would say we are still trying to figure out what big data research can do, can contribute to the field.*

Researchers also commented on the inefficiency of learning a new skill without dedicated leave such as a sabbatical—they would rather hire or bring in an expert to tackle a particular new method. This being said, they encourage their students to take classes that will provide them with a variety of skillsets, such as a Health Policy student taking a Bioinformatics course. There is a recognition both that there is a gap in the necessary skills, and that they can prepare their students to enter the field with a robust methodological background.

Using High-Performance Computing Resources

The two biggest sources of support for computing resources are UMass' Research Computing—which recently launched the new Unity cluster—and the Massachusetts Green High Performance Computing Center (MGHPCC). Both provide access to high-performance computing resources that allow researchers to store massive datasets and process them much more rapidly than would be possible on local machines.

Most big data researchers on campus use one or both of these high-performance computing resources to host and analyze their data, with the exception of the Computer Science department (which has its own GPU cluster) and the School of Public Health (which has its own HIPAA-compliant cluster housed at the MGHPCC). Some researchers maintain their own servers, either because the MGHPCC was not yet available or because it does not allow for the specific configuration they need, or they use Amazon Web Services to configure virtual machines in the cloud. Lab-specific servers are often maintained at the MGHPCC facility, but are occasionally housed on-campus.

Table 1. Articulated Pros and Cons of Computing Cluster Options for UMass Amherst Researchers

	Pros	Cons
PI-purchased servers	• control over hardware and software configuration	• high startup capital • ongoing funding for maintenance personnel
Departmental clusters	• department pays certification (e.g. HIPAA) and maintenance • free to departmental researchers	• department controls available installed packages • departmental support personnel may be unversed in needed areas or not funded by the department • not available all departments
Research Computing Unity Cluster	• dedicated, skilled staff for software and server maintenance • free to UMass researchers	• Research Computing controls available installed packages • training in use is lacking • computing capabilities limited
MGHPCC	• dedicated, skilled staff for software and server maintenance	• MGHPCC controls available installed packages • training in use is lacking • paid service
Collaborators' infrastructure	• skilled researcher overseeing process (collaborator) • saves UMass researcher time and effort	• UMass researcher has little to no control over process • transfering completed data analysis between institutions can be tricky

CHALLENGES IN USING HIGH PERFORMANCE COMPUTING RESOURCES

While our interviewees expressed a wide range of positive attitudes towards the Unity cluster and the MGHPCC, they also acknowledged several limitations. High-performance computing resources require a knowledge of coding languages, and the documentation on how to use them can be overwhelming to researchers who have never used cloud computing resources before. There are limitations as to how the virtual machines and clusters can be customized and configured. Some pointed out that they occasionally run out of local resources, such as storage space or sequencing processing in the genomics core, or that they have a bottleneck, such as the need to do computationally intensive work before they can move on to the next task in their analysis workflow. Others cite that early adopters of the Unity cluster at UMass encountered challenges and setbacks, but that it has ultimately been beneficial to the group. While the currently available resources are great in many ways, they are not a one-size-fits-all solution:

> *I understand that our Research Computing folks are always striving to make more resources readily available to the campus community, and I appreciate their efforts and it's something that we're all pulling for. It's just a very expensive proposition to have clusters that are spun up that can do something that can serve everybody.*

> *And I do remember [the IT support person] sort of throwing me into the deep end of MGHPCC. Like, 'here are the keys!' Like, oh, yeah, we can do this and do that. The capacity was mind-boggling at the time, but … I do think that it would be invaluable to have this resource, or even a really simple program … to bridge that gap, especially with faculty.*

Staffing is a challenge in its own right. Many researchers mentioned that even if computing clusters and infrastructure are available, if the appropriate network administrators or staff are lacking, it will fail. Good infrastructure managers, well-versed and vetted consultants, and adequate staffing levels are critical to the success of our researchers. One researcher highlights this:

> *When I arrived at UMass [there was a] complete lack of big data research and infrastructure on campus. And so it has changed a little bit, but it hasn't changed enough that there is adequate support on campus for all of the parts of research. And so some of these things that I gloss over are actually huge barriers to doing research. Like getting a data use agreement signed. I had an update to a DUA.... It contains two sentences, and that took six weeks and four reminders, including escalating through the levels of administration. It doesn't really make me want to deal with this. And so this kind of administrative burden can be really really tough.*

Our interviewees acknowledged the role that undergraduate and graduate students play but commented that they are educating their students. The need for on-demand advanced assistance is inadequately met by the current staffing infrastructure.

In terms of getting help using these resources, setting up infrastructure for individual research projects, or having access to technical expertise to maintain project clusters and servers, our interviewees had vastly different experiences depending upon which department they were based in. One referred to this

landscape of technical help as “fragmented across different departments.” For example, a researcher in Environmental Conservation recalled that the department used to have dedicated technical staff or researchers with specialized technical skill sets who could provide individualized support to their colleagues. The College of Engineering is currently supported by the Engineering Computer Services, but the researcher we interviewed who uses this service believed that a key individual was retiring or that the unit would be absorbed into campus IT. In short, there has been a pattern of department-based IT support being retired and/or absorbed into the centralized Research Services. This is viewed as a reduction in services, since the move to central IT means the expert will have more demands on their time:

> *And so we installed [a] $100,000 cluster just for that project.... The key to that and the resource that most people don't have is we have a very good network administrator who manages our computing infrastructure.… And so other departments do not have this person, and they struggle mightily with computing and providing computing resources for their faculty. But he's like a 10X improvement on my research, just having him.*

And yet others mentioned the challenge of hiring qualified personnel with soft money or pursuing interdepartmental collaboration with colleagues who may not be able to commit to ongoing maintenance.

Other departments also provide discipline-specific support for data collection and analysis, but even the subject-specific high-performance resources fail to meet the needs of all researchers. For example, the work one researcher had done at the IALS Genomics Resource Laboratory is run on their Illumina sequencing system, which feeds to a BaseSpace cloud analysis architecture. From there, the researcher moves their data to free tools for more in-depth and customized analysis, but the free tools are not the best option for the work:

> *Ingenuity Pathway Analysis, IPA, it's one of the tools which every big university has access to. We don't. And I think that if we will, we will immediately have—I don't know—a hundred users. Because it's one of the best tools for bioinformatic analysis of gene expression data.... But we don't have access to it.*

Several researchers mentioned that the high performance computing needs of their project were being met by the infrastructure that is available at collaborators’ institutions. A variety of factors seem to influence this choice, including the extra effort involved in moving the data, whether the collaborator has the necessary coding skills, and a lack of adequate infrastructure at UMass.

Staying Up-To-Date in the Field

The general strategy that big data researchers use to stay up-to-date on developments in the field is to maximize their efficiency: they use different tactics to learn about new and relevant studies while minimizing the amount of time needed to do so. Researchers have limited time to spend on learning, and the “explosion of information” or “information overload” is a major obstacle to staying current:

> *Honestly at this point it's like a firehose of information at all times. And so it's borderline impossible to keep up with the field, broadly defined.*

The strategies that our interviewees cited to maximize their time efficiency include enlisting graduate students to carry out literature reviews, either independently or as assignments in seminars; forming lab or project reading groups to read relevant articles; subscribing to arXiv feeds, email digests, and Google alerts tied to specific keywords to filter content to only the most relevant studies; and using Twitter to learn about the most recent advancements in the field.

Although a few of our interviewees do not use social media at all, a majority reported using Twitter regularly. The benefit of having access to curated, field-specific content is generally seen as outweighing potential negative aspects, such as algorithmic bias toward more influential accounts with high follower counts. It also provides a quick and easy way for researchers to learn about developments in fields that are on the "fringe" or tangential to their core interests:

> *Some people actually monitor the arXiv feed every day and look at the hundreds of papers that have been put out, and that's not really feasible for most people. So you need at least someone to filter and find interesting papers. And, yeah, right now Twitter is probably the place to go for that.*

Only one of our interviewees mentioned carrying out a literature review of specific journals as a part of the information-gathering stage for a project. And while many discussed conferences as a key venue for sharing about their research findings (see "Sharing Project Findings"), only two highlighted the value of conferences as a place to learn about developments and network with other researchers in their fields. For them, it was the one-on-one discussions with colleagues that provided the best context for learning.

Learning about developments in industry was less of a priority for most of the researchers we interviewed; some have RSS feeds to follow mainstream news or learn about developments outside academia tangentially, such as through the patent application process. One expressed an ambivalence to learning about non-academic pursuits because they felt the tools and methods they used were essentially the same, but that the questions their projects were seeking to answer were too different to be comparable. However, researchers who have direct links to industry—such as through colleagues who formerly worked for tech companies—expressed excitement and interest in these connections.

> *Some of the equipment … is actually industry-donated, so they can use it for some of their R&D also. And so that's a nice two-way street where they say, "Hey, we have this new technology, do you want to try it out?" We just had a talk about some Intel FPGA boards that we might want to put in a testbed. And I don't know if we would have learned about this that early on if we hadn't had that person-to-person relationship with Intel.*

Sharing Project Findings

The ways in which researchers share the findings from their projects vary widely and are largely dependent on whether the research is more academic or more applied in nature and what field the researchers belong to. The majority of our interviewees are engaged in academic projects and share their project findings in traditional academic venues: as working papers shared internally or at brown bags, preprints, conferences, journal articles, and (less commonly) books. Social media, university press offices, and traditional media are all ways in which they communicate their findings to the broader public. Our interviewees who are engaging in applied projects typically share their findings as reports to their clients or as white papers, and (less commonly, or at least aspirationally) in journal articles. Press releases are also important ways of communicating more widely about their outputs (Figure 2).

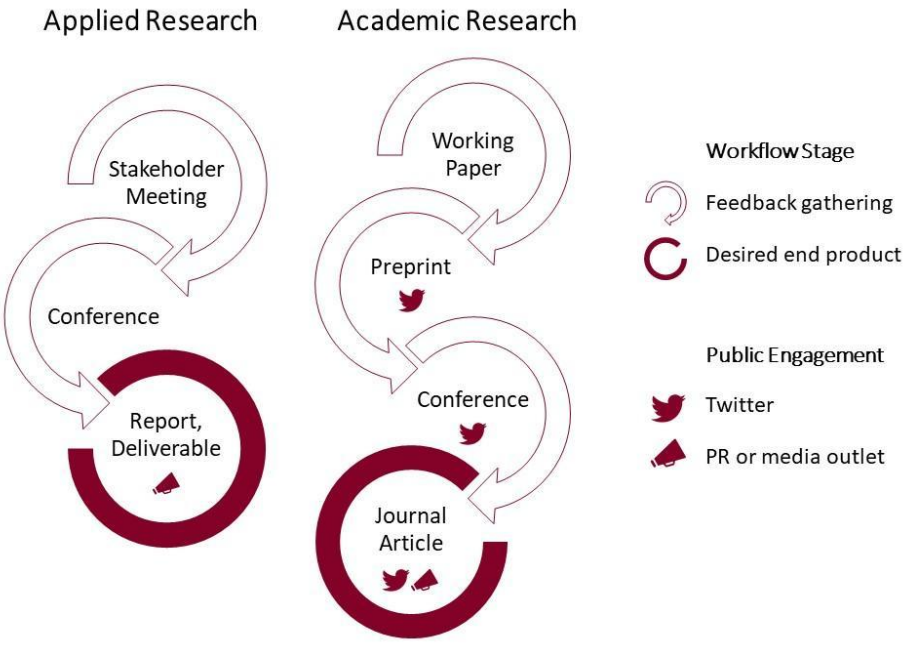

Figure 2. Typical workflows for sharing project findings: applied vs. academic research.

For those in the academic track, the researchers' field plays a key role in determining which of the primary venues is prioritized. For fast-paced fields, speed is of utmost importance—these researchers tend to share their research on preprint services like arXiv which are not peer-reviewed, allowing their work to start being cited immediately even if it takes several years for the final publication to appear. Conferences also provide a relatively quick turnaround time for sharing research findings, but often have the added advantage of providing peer-review—depending on the conference, researchers may be required to submit an extended abstract or even a full paper to be reviewed before acceptance. Our interviewees reported attending between 2 and 4 conferences in a typical year. Slower-paced fields continue to prioritize publications in peer-reviewed journal articles, which usually subjects the study to more rigorous peer-review. But regardless of the venue that is considered the "top priority," there was general consensus that the workflow in Figure 2 represents the typical trajectory for an academic

researcher: starting with low-stakes venues like working papers and conferences allows the researcher to gather feedback and refine the paper into a publishable format.

There was general acknowledgment that research shared in traditional formats (like journal articles) is rarely accessible to people outside academia, and therefore is unlikely to enter public discourse. In order to achieve broader public attention for a particular project, researchers need to engage directly with the public by writing pieces on public platforms like Medium.com, *The Conversation*, or *Boston Review*. Some maintain personal or project websites where they can share media blasts or videos. Or they might enlist the help of the university's press office or the marketing and outreach teams for their specific school or college, if one exists. Sometimes a story advertised by the press office will be picked up by media outlets; this is how one of our interviewees was contacted by a local newspaper and asked to give an interview on a topic related to their research. But by far the most common public engagement tool that our researchers are using is Twitter.

A minority of our interviewees talked about public engagement as something they wanted to do or felt an ethical obligation to do (cf. attitudes toward code and data sharing, which a majority shared). One shared that public engagement was exciting for them and something they felt compelled to do: "if the public does not have faith in science, it's our failure." Another—who is involved in a more applied project—cited public engagement as a strategic means of promoting their work and getting ahead of the competition. But the majority expressed ambivalence; for example, "it's not something that I actively try to do."

A few expressed a desire to engage with the public, but believed that they did not have sufficient time, incentive, or skill to do so:

> *Since I'm pretty junior, I've been more focused on just publishing than actually translating results to practice. But that's something that I think is a future goal.*
>
> *I would say I'm learning to try to go beyond the journals and academic meetings, but it's not what we're trained to do. I mean, as you know, we have this very narrow skill set.*

CHALLENGES IN SHARING RESEARCH FINDINGS

There was no clear singular challenge that all researchers face in sharing their research findings. Some cited the bias towards prestigious universities or programs as a limiting factor, whereby well-known individuals are more likely to have large followings on Twitter and have their tweets promoted widely, or to encounter a smoother peer-review process for submissions to conferences and journals. Researchers on the fringe of traditional academic research—whether because their projects are more applied in nature, or because their work is interdisciplinary—have their own unique challenges that make it hard to be assessed according to a traditional academic rubric of success. Applied researchers—particularly those in non-tenure track positions—must prioritize writing for non-academic audiences (reports, white papers) which are not valued as highly as peer-reviewed papers and make it hard for those researchers to pursue tenure track careers in the future. Interdisciplinary researchers likewise find it difficult to identify appropriate journals for their research or to find adequate peer reviewers.

Sharing Code and Data

The practice of sharing code and data varies widely across campus and is largely determined by the funding researchers receive for the project, the platform they use to publish their research findings, the field in which they are situated, and/or their own attitudes and ethical beliefs. One of the major motivators is external pressure: some federal grant agencies (like the National Science Foundation or the National Institutes of Health) now require data to be made available to the public, and increasingly journals and even conferences are requiring code and data to be shared as part of the peer-review process. The goal of these institutionalized incentives is to ensure reproducibility, accessibility, and transparency. A few of our interviewees pointed out that adhering to these requirements meant an increased likelihood of receiving funding in the future or a research paper being treated benevolently by peer reviewers.

At the same time, many of our interviewees are motivated by intangible cultural attitudes about the value of open-access research. These attitudes appear to vary widely according to field; in certain fields (such as Computer Science, Bioinformatics) researchers are encouraged or even expected to share their research products, while other fields (such as Economics) are just now starting to move in that direction. One researcher succinctly stated: “It’s the right thing to do.” Others cited “community norms,” “prosocial behavior,” the “moral imperative,” contributing to the “community of movement,” being a part of “the growing trend of open scholarship,” creating an “open culture,” and being a “good citizen” as motivations for sharing their code and data. Others pointed to increased citation counts and enhanced impact within the field (particularly for data relevant to pressing issues like climate change) as additional reasons for sharing.

> *I'm using datasets from others, so I think others will only be willing to continue to do that if I, or if the rest of the community, contribute to it. It's the same with writing papers and reviewing papers, and so I think it's always a two-way street.*
>
> *If you produce something, if you publish it, it should be accessible for any researcher.*

It is notable that there are few, if any, institutional incentives at UMass for researchers to share their code and data. Preparing these products to be shared publicly takes a great deal of time and effort, the bulk of which usually falls to the researcher. The most commonly cited reasons for taking this work on is time savings—the people who are best capable of preparing data and code are the ones who analyzed or created it in the first place, and teaching someone else about the data or code would take longer than doing the cleaning and preparation work oneself. A few of our interviewees rely on the labor of undergraduate and graduate students who are a part of the project and are therefore familiar with the code and data. A few expressed a need for help with this part of the process, but were not aware of any existing resources.

SHARING PLATFORMS

For both code and data, there are two broad categories of sharing platforms: those used for collaboration during the course of the project, and those used to disseminate products to the public.

Code is usually shared with collaborators in private, cloud-based repositories like GitHub and Amazon Web Services. One interviewee uses Google Colab because their collaborator (who writes the code for their project) prefers it. If researchers decide to share their code publicly, GitHub is the most common platform choice, especially if that is where the code was stored during development. Some opt to share their code on their own or a colleague's website or as an appendix to a journal article. While code can be shared as simple text files, some package their code in executable formats (e.g. R packages, Docker containers, Jupyter Notebooks) to make it easier for reviewers or other researchers to implement.

Data is most often shared with collaborators via university- or department-based servers (see "High-Performance Computing Resources") or institutional cloud document storage such as Box or Google Drive. One interviewee reported using Dropbox to share data with a collaborator, but because the collaborator did not have a professional license they were unable to edit and share back the data (i.e. Dropbox provided data-sharing capabilities but not collaboration capabilities). There is much more variation in how researchers publish or share their data with the broader public. In the past, federal agencies like the National Institutes of Health (NIH) that required data-sharing provided hosting services, but this seems to have been phased out as storage requirements exceeded their capacity. Our interviewees reported sharing their data via conference tracks dedicated to data; services like Preprints, Mendeley Data, or Data In Brief; UMass-based servers that are available to researchers in specific schools (like the College of Information and Computer Sciences); and UMass Libraries' ScholarWorks.

CHALLENGES IN SHARING CODE AND DATA

One of the biggest hurdles for researchers to overcome is the time needed to clean and prepare their code and data. One researcher expressed embarrassment about the state of their code and cited this as their main reasons for not sharing it, despite a strong personal desire to do so. Another regularly shares code for their new publications, but they have been unable to find the time to clear up their code from earlier publications.

> *I look at the coding that I did from 20 years ago or 10 years ago, whenever it was, and it would require a lot of work to make it minimally acceptable to not embarrass myself.*

From an ethical standpoint, some types of data (e.g. weather measurements) can be shared without hesitation, while others (e.g. data containing Personal Health Information) cannot. But for researchers working with data in the gray zone between these two poles, the decision about whether or not to share their data can be less clear-cut, even if that data is anonymized or aggregated. For example, it is not always clear whether Institutional Review Board (IRB) approval is necessary to work with some human-generated data, such as Twitter or Facebook posts, which are technically public information but could be viewed as private in some ways:

> *I haven't heard any clear answer.… I think each discipline has its own guidelines. And sometimes you have to apply common sense.… If I'm looking at public organization accounts on Twitter, yeah, I wouldn't necessarily think about IRB because those are organizations, right? Those are their public documents. But if I'm starting assessing topics and I'm collecting the communication*

record of people talking about certain health issues or sensitive political issues, yeah, I would think about getting IRB.

More generally, there can be confusion about the researcher's obligations, and there is no external support to help them determine the best course of action. For example, one researcher cited concerns about the changing nature of federal requirements to share their data with the public:

I try to pay as little attention as possible because … it's just stressful to think about ... what one is required to do in a lot of cases, versus what one can do. You can get really caught in the middle here as a researcher.

Last but not least, one interviewee cited a lack of clear guidance about how to publish datasets as a barrier they face, specifically when working with a collaborator at a different institution. Having clearly articulated best practices about the data publication workflow—where to publish the dataset (i.e. which institutional repository, or both), how to obtain a DOI, and how to integrate that publication with other venues—would be helpful to researchers interested in sharing their data.

REASONS NOT TO SHARE

Limited time and a lack of institutional incentive are major factors for researchers who decide not to share these products. If funding agency or publisher requirements are not a concern, and if their field does not have a culture of supporting open-access scholarship, the effort needed to clean the code and datasets and prepare metadata may not be judged as worthwhile:

I'm sort of somewhat in a teaching track. I have a lecturer position... I don't have the impression that people would take a serious look at something like my GitHub profile if I put things up there. It would help if I were to leave academia, I think. It might help getting a job then.... I don't think inside academia in my field it is useful at the moment.

A common reason why researchers opt not to share their code or data is that it is not considered to be "useful" to others. Many do not share their code if it not an explicit contribution to the field: for example, code that is designed to transform their data at a specific step in the workflow (i.e. "project management" code) would not necessarily be helpful to other researchers or easily implementable in their own unique workflows. In these cases, researchers might choose to make the code available only upon request.

Similarly, researchers are unlikely to share data that is already publicly available, since other researchers can already access it directly. At most, they might include an appendix to the publication that explains any specific changes that were made to the dataset for that project.

Other reasons that researchers cited for not sharing their code or data include a fear of inadvertently exposing highly sensitive data, a fear that their research would be "scooped" by other researchers, and not being allowed by the data provider to release the data (e.g. Facebook data gathered through CrowdTangle cannot be shared).

Conclusions

There are signs at the campus level and within departments that administrators know how important supporting big data research is for the future of innovation and discovery. The partnership to create the MGHPCC, investment in the College of Information and Computer Science's Center for Data Science, and the IALS research cores were all strategic investments in supporting and advancing computing intensive research. Again and again, our interviewees described big data analysis and data science techniques such as machine learning and natural language processing as the future of their type of research. Demand for these skills, the necessary infrastructure, and support services, is only going to grow. As the 2015 President's Office Report report put it:

> *Through concerted efforts and targeted investments in talent, educational programs, and infrastructure, Massachusetts and UMass will be well equipped to emerge as an international hub of data science discoveries and innovation, home to coveted degree programs, and a primary source for the next generation of Big Data scientists. By working together we will contribute to a better world and a prosperous and forward-thinking Commonwealth.*[3]

A recurring theme in our findings is the need for more support. Whether it comes up because the researcher's department is losing its dedicated IT person and they are not sure campus IT can meet their needs, they are stymied in hiring or learning the coding skills necessary to advance their projects, they are overwhelmed by the quantity or mixed quality of information they must sift through, they are not able to dedicate time to open-science practices because it is not valued as highly as a journal manuscript is, or they simply are not getting enough training to be confident using the MGHPCC, every researcher expressed a wish for more support in some aspect or another of their work. We believe that the university will be best served to continue its commitment to the future of research, and that the next stage of investing in that future is investing in people.

Recommendations

Review institutional processes around data access

While our interviewees did not provide explicit suggestions about how to reduce administrative burden, it was clear that they perceive these burdens to be a hindrance to big data research. We recommend undertaking a review of our institutional processes and procedures to provide transparency and clarity so that researchers are set up for success. Rather than act as gatekeepers or bureaucrats, the roles of those who respond to requests for access, resources, and permissions should be reframed as opportunities to support researchers.

[3] UMass Data Science Faculty Working Group, "UMass Big Data: Data Science for the Commonwealth Powered by the University of Massachusetts," 2015, p. 3, https://archives.lib.state.ma.us/handle/2452/430344 (accessed 24 August 2021).

Mitigate information overload

Researchers are overloaded by the amount of information that is out there, and they are essentially left to their own devices to make sure they stay current. Librarians can help filter the "firehose of information" to a more manageable stream by offering workshops and information on the libraries' website on services and tools to curate researchers' reading lists, from Twitter hashtags and RSS feeds to expert review and recommendations sites.

Fund additional permanent positions in Research Computing

The retirement of school- and department-based technical staff in recent years has left critical gaps in the infrastructure of computing support that is available to researchers on campus. Right now, researchers outside the College of Information and Computer Sciences and the School of Public Health and Health Sciences (which both have their own computing clusters and staff) need the most help. Funding additional positionals in central campus IT would help researchers maintain their technology and data-generating workflows.

Offer workshops to help researchers develop technical skills

This study highlights the lack of training support for researchers who need to acquire new skills in order to work with big data. While a further survey would be helpful in determining the level of detail and specificity to disciplinary needs these workshops should aim for, for now, it is clear that workshops in coding (Python, R, Jupyter Notebooks) and data cleaning, analysis, management, organization, and re-use would be highly beneficial across all fields. One option would be to hire a Data Analysis librarian or an expert for the Research Computing group and run the workshops out of one of those departments. Another option would be for current librarians or Research Computing staff to approach campus community members with the necessary expertise and build collaborations to support ongoing workshops and training in addition to their current duties.

Develop introductory trainings for high-performance computing infrastructure

The two main computing clusters available at UMass Amherst—the Unity cluster and the MGHPCC—have a steep learning curve that can discourage new adopters. Both Research Computing and the MGHPCC should develop training resources that are geared toward users with no coding experience and that provide advice for finding scripts online or connecting them to campus services such as the Statistical Consultation & Collaboration Services or the Center for Data Science.

Facilitate interdepartmental collaboration

The work our big data researchers are undertaking touches on a variety of disciplines and would benefit from easier cross-campus collaboration. And yet, our interviewees gave mixed reviews of how well their departments, this campus, and their promotion and tenure boards view interdisciplinary work. We recommend the administration develop clear messaging, provide institutional collaboration tools, and promote a shift in promotion and tenure judgement on interdisciplinary work. The university can immediately improve opportunities for cross-training and bridging across fields by lifting department-only restrictions on courses or seminars.

Articulate best practices for the data publication workflow

Making the data publication process easy and quick will encourage junior faculty especially to publish their data. A LibGuide or other Libraries-based resource that is available to all researchers on campus would be an excellent resource to detail the steps in the process, including how to request a DOI, how to publish data in ScholarWorks, and how to streamline with data submitted on other platforms (e.g. for conferences).

Improve information quality

When attempting to reuse others' datasets, our researchers are often hampered by the insufficient education within their disciplines. Challenges they face include insufficient metadata, poor data quality in publicly available data sets, a lack of discoverability, and more basic data management and sharing inadequacies. While there is no easy answer for any of these challenges, we can invest in making UMass' data contributions high quality and reproducible. Scholars across campus would benefit from tools, methods, and other support to make their data FAIR (Findable, Accessible, Interoperable, and Reusable). Tools used at the big data scale could inform smaller datasets, as well, and best practices for FAIR are discipline neutral. We can also pressure our partners, both industry and academic, to adopt FAIR principles.

Prioritize integrating data literacy across the curriculum

Our interviewees highlighted several areas of critical big data skills that the graduate students in their disciplines tend to lack. While not a concern where data analytics and coding are part of the core curriculum, researchers in such areas as Public Health & Health Policy, Environmental Conservation, Linguistics, and more had concerns about their students' skills in data reuse, data management, coding, data analysis. Every department that produces researchers should incorporate training in these skills across the curriculum to ensure our graduates are able to contribute high quality, FAIR (Findable, Accessible, Interoperable, and Reusable) results to the scholarly record.

Provide incentives to reward alternatives to traditional publication efforts

Not all research lends itself to scholarly books and journal articles as an end product. Even interviewees who described how sharing code and data have become expected norms in their disciplines could cite no examples of such work receiving departmental or university-level recognition. Many researchers in more applied areas are producing white papers, reports, code, and other contributions to their field that improve the research of the community. Other areas where faculty are encouraged to contribute but not rewarded for their efforts within the traditional promotion and tenure structures include interdisciplinary work and collaboration between departments, communicating science to the general public, and work towards products of applied research translational work. Incentivizing these areas and incorporating them as part of the evaluation of promotion and tenure would demonstrate the University's support of all aspects of current research and help shape the researchers of tomorrow.

Appendices

Appendix I. Interviewee Solicitation Email

Subject. UMass Amherst's study on supporting big data research

Dear [*first name of researcher*],

The UMass Libraries are conducting a study on the practices of researchers who use big data or data science methods in order to improve support services for their work. Would you be willing to participate in a one-hour interview to share your unique experiences and perspective? We will be recording interviews in order to generate a transcript of our interview for analysis purposes. Transcripts will not have any identifying information, and audio recordings will be deleted immediately upon completion of the transcript. If you feel uncomfortable being recorded, we will not be able to conduct an interview, but we would be happy to speak with you informally after we have findings to report.

Our local UMass Amherst 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 UMass Amherst will also be included in a landmark capstone report by Ithaka S+R and will be essential for UMass to further understand how the support needs of big data/data science researchers are evolving more broadly.

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

Sincerely,
Thea Atwood, Melanie Radik, and Rebecca Seifried

Appendix II. Interviewee Consent Form

Researchers. Thea Atwood, Data Services Librarian (PI); Melanie Radik, Science and Engineering Librarian; Rebecca Seifried, Geospatial Information Librarian

Project title. Supporting Big Data Research

---

1. **What is this form?**

   This form is called a Consent Form. It will give you information about the study so you can make an informed decision about participation in this research. We encourage you to take some time to think this over and ask questions now and at any other time. If you decide to participate, you will be asked to sign this form and you will be given a copy for your records.

2. **What are some of the important aspects of this research study that I should be aware of?**

   1) The fact that consent is being sought for research and that participation is voluntary;

   2) We are investigating researchers' practices working with big data and data science methodologies. We expect that your participation will be one interview lasting approximately sixty minutes. The interview will focus on how you complete your research, and will be conducted by an authorized research team member;

   3) There are minimal risks associated with this research study; however, a risk of breach of confidentiality always exists and we have taken the steps to minimize this risk as outlined in section 9 below; and

   4) Benefits of this study may include increased insight and awareness into your research practices and support needs, and will inform the development of new services to support your work.

3. **Why are we doing this research study?**

   This study seeks to examine researchers' practices in working with big data/data science methods in order to understand the resources and services that researchers at the University of Massachusetts Amherst need to be successful in their work.

   Further, the study at the University of Massachusetts is connected to a suite of parallel studies being developed locally at other higher education institutions. The anonymized, aggregate data shared with the coordinating organization, Ithaka S+R, will be used to compose a comprehensive report written and made publicly available by Ithaka S+R. Such a report will provide guidance to other higher education institutions in providing support and services for researchers using big data and data science methodologies.

4. Who can participate in this research study?

We are seeking researchers at UMass Amherst who work with big data or data science methods to participate in this study.

5. Where will this research study take place, and how many people will participate?

Interviews will take place either in person or virtually. In person interviews will take place somewhere private, like your office or a room designated for meetings. Virtual interviews will take place over Zoom.

We anticipate approximately 15 researchers to take part in this study.

6. What will I be asked to do and how much time will it take?

7. If you agree to take part in this study, you will be asked to participate in one 60-minute, audio-recorded interview about research practices.

Your participation in all or part of this study is completely voluntary. You are free to withdraw consent and discontinue participation in the interview at any time for any reason.

8. Will being in this research study help me in any way?

You may experience benefit in the form of increased insight and awareness into research practices and support needs. More broadly, your participation in this study will help develop resources and services in support of your research at the University of Massachusetts.

9. What are my risks of being in this research study?

We believe that there are minimal risks associated with this research study; however, a risk of breach of confidentiality always exists and we have taken the steps to minimize this risk as outlined in section 9 below.

10. How will my personal information be protected?

Your privacy and confidentiality is important to us. The following procedures will be used to protect the confidentiality of your study records.

Securely storing data

The researchers will keep all study records in a secure location. The study will largely consist of digital files, which will be stored in Box, UMass Amherst's secure storage solution. If you decide to sign a physical consent form, this will be stored securely in a locked file cabinet in the PI Atwood's office. Our procedures for protecting your personal information according to the file type is outlined in the table below.

File type	Secure Location	Privacy procedures
Audio recording of interview	UMass Box	Recorded in order to create a transcript of the interview for analysis. We immediately apply a pseudonym to the transcript of our interview. There is no key to link you to your pseudonym. Once the transcription is complete we delete the original audio recording.
Transcripts of interviews	UMass Box	Pseudonyms are used throughout this document. Shared only with members of the project.
Consent form	Physical: in a locked file cabinet Digital: on Box	Consent forms are stored securely for the duration of the study. Upon completion of the study consent forms are destroyed.

At the conclusion of this study, the results of the research will be publicly disseminated, such as through conference presentations, scholarly articles and as part of publicly available reports published online through ScholarWorks@UMass, the University of Massachusetts Amherst's dedicated institutional repository, and the Ithaka S+R website.

Information will be presented in summary format, demographic or contextual information will not be used in public reports of the research findings, and you will not be identified in any publications or presentations.

Protecting your individual privacy as a participant in this study

To protect your individual privacy, study procedures will be conducted in a private location, like your office, one of the researcher's offices, or a private meeting space. Only authorized research team members will meet with research participants.

Your signed consent documents will be stored securely and separately from the research data.

Protecting your individual privacy when sharing data with our consulting organization

The study at the University of Massachusetts 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, has been hired by the researchers to coordinate this parallel effort and to provide guidance on research methodology and data analysis. The research project will be implemented exclusively by the investigators listed on this form.

Anonymized, aggregated data and analysis will be shared with Ithaka S+R in order to compose a comprehensive report written and made publicly available by Ithaka S+R.

Ithaka S+R will have no access to the research subjects or their personal information. Ithaka S+R will only have access to de-identified interview transcripts and de-identified metadata about the transcripts, not the audio recordings.

11. Will I be given any money or other compensation for being in this research study?

Participants in this study will not receive payment or other compensation for being in this research study.

12. Who can I talk to if I have questions?

Take as long as you like before you make a decision. We will be happy to answer any question you have about this study. If you have further questions about this project or if you have a research-related problem, you may contact the researchers:

Thea Atwood, tpatwood@umass.edu, 413-545-2674
Melanie Radik, mradik@umass.edu, 413-545-6943
Rebecca Seifried, rseifried@umass.edu, 413-577-5317

If you have any questions concerning your rights as a research subject, you may contact the University of Massachusetts Amherst Human Research Protection Office (HRPO) at (413) 545-3428 or humansubjects@ora.umass.edu.

13. What happens if I say yes, but change my mind later?

You do not have to be in this study if you do not want to. If you agree to be in the study, but later change your mind, you may drop out at any time. There are no penalties or consequences of any kind if you decide that you do not want to participate.

14. Subject statement of voluntary consent

When signing this form I am agreeing to voluntarily enter this study. I have had a chance to read this consent form, and it was explained to me in a language which I use. I have had the opportunity to ask questions and have received satisfactory answers. I have been informed that I can withdraw at any time. A copy of this signed Informed Consent Form has been given to me.

_____	_____	_____
Participant Signature:	Print Name:	Date:

By signing below I indicate that the participant has read and, to the best of my knowledge, understands the details contained in this document and has been given a copy.

_____	_____	_____
Signature of Person Obtaining Consent:	Print Name:	Date:

Appendix III. Semi-Structured Interview Questions

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

As a reminder, this interview is being recorded so that we can create a transcript for data analysis. Any identifying information will not be included in the transcript, including your name.

Introduction

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

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

Working with Data

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

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

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

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

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

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

- What software or computing infrastructure do you use? *Examples: programming languages, high-performance computing, cloud computing*
- What challenges do you face in analyzing or modeling data?
- If you work with a research group or collaborators, how do you organize your data and/or code for collaboration?
- Do you take any security issues into consideration when deciding how to store and manage data and/or code in the course of your research?
- Does anyone other than your research group members or collaborators help you analyze, model, store, or manage data? *Examples: statistics consulting service, research computing staff*

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

Research Communication

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

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

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

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

Training and Support

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

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

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

Wrapping Up

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