ITHAKA: Supporting Big Data Research
Data and Analysis from UW-Madison Researchers

Cameron Cook
Tom Durkin
Jennifer Patiño
With contributions by Tobin Magle
UW-Madison General Library System
2021

Introduction

This study seeks to better understand the practices and needs of University of Wisconsin-Madison researchers working with big data and data science methods with an aim of helping to improve big data research support services locally. It was conducted in partnership between the University of Wisconsin-Madison's General Library System and the Division of Information Technology's (DoIT) Research Cyberinfrastructure initiative as part of Ithaka S&R's

Supporting Big Data Research project. This multi-institutional project was designed and led by Ithaka S&R and culminated in a local report from each participating institution as well as a meta-analysis of the interviews and findings published by Ithaka S&R.

Methods

Researchers working with big data or data science methods were identified from 10 unique departments and across all domain areas, though the Natural Sciences, Applied & Formal Sciences, and Social Sciences represent the majority of the interviews. One-hour semi-structured interviews were conducted with an interview guide provided by Ithaka S&R. Transcripts were auto-generated from the interview recordings, with manual cleaning for correctness, completeness, and typos conducted by the research team. Themes were identified from a subset of transcripts and appropriate codes developed. The transcripts were then coded and analyzed using NVivo.

Accessing Research Data

Of thirteen researchers interviewed at UW-Madison, the majority worked with a mix of both generated and secondary datasets, two exclusively generated their own data, and one primarily worked with secondary datasets. Researchers interviewed generated data in the following way:

- Experiments, observations, imaging (4)
- Computer simulations (1)
- Digitization (1)
- Surveys (3)
- Scraping open databases (1)
- Directing experiments with collaborators in other departments (1)
- Partnering with external collaborators (2)

Sources for secondary datasets used by researchers included:

- Government data (7)
- Publicly available papers, databases, or websites (5)
- Data from external collaborators (1)
- Subscription databases (1)
- Private data from third party (1)

Of the two researchers who exclusively generate their own data, one researcher described lack of access to curated secondary data sets as a significant barrier.

> “Essentially, the only data [I have] access to is data that I generate a model. And that's an obstacle to some of these algorithmic developments that I talked about earlier. Right,

> if I could somehow collect all the data that I know all of my colleagues have and use it for algorithm development, neural network training things would be a lot faster."

When researchers require the use of external datasets, acquisition can include additional complications and responsibilities for the researcher. One researcher mentioned the extra administrative labor that going back and forth around legal documents took to be able to get access to a third party's data. "It was really effective to have a couple of people who I was able to compensate for the hours spent poring over administrative details, and the people who spent a lot of time doing the programming and initial processing on some of the environmental data." Another researcher spoke about the challenge of getting corporations or institutions who see their data as proprietary to agree to share it.

> "It took years of saying pretty please, etc. and we're currently trying to get more data from them and they are willing to do that, but we're working with the University lawyers in order to get to an agreement that they would find acceptable, right, sort of protecting their intellectual property, and that the University would find acceptable from a research perspective. So I think navigating that trade off has been difficult to say the least."

A different researcher described how compared to publicly funded research data, industry data in their field is inaccessible. "That's also an arena where the pharmaceutical industry probably has a 100 or a 1000 times more data that would be really great for us and is certainly not publicly funded and not in those repositories so it's just inaccessible for us."

Even when researchers are able to acquire data, they can encounter other barriers to access such as having to merge data from a number of sources, unintuitive data interfaces that make cleaning data challenging, restrictions on scraping, and the impermanence of some of their data sources. One researcher using public databases mentions their completeness as a challenge along with errors that make merging between datasets difficult. Another researcher describes the extra work required when using a complicated data portal: "[The data owners] have a very complicated and un-intuitive data interface website that requires a lot more sort of clean up of data, reshaping of data in spreadsheets to make it usable." Another researcher who has access to subscription databases is finding that restrictions on scraping are a barrier. "I'm trying to move away from [databases] because they don't allow scraping and downloading things manually from there one-by-one is really time consuming, so I'm trying to rely on these public databases. I have done some scraping as well from the web but that can be pretty messy." Lastly, a researcher spoke about the challenge of impermanence when it came to accessing data in their area of study. "You kind of grab the data when you can, because in an authoritarian country, you never know how long it's gonna be there."

Open Data: Dissemination, Sharing, Incentives, and Rewards

Researchers at UW-Madison working on big data projects are very interested in and concerned about issues related to the dissemination and sharing of data, coding, and outputs such as

papers and conference presentations. Researcher interests and concerns go beyond technical issues, and focus on expressions related to risks, frustrations, and rewards connected to their web of professional relationships. In our coding for “Dissemination and Sharing,” we used ten base codes. The most frequently used code was “External Collaborators.” Other frequently used codes included “Data Sharing,” “Incentives and Disincentives,” “Internal Collaboration,” and “Publications.” Academia is an intensely social profession. The strongest pattern to emerge from analyzing the co-occurrence of themes related to the dissemination and sharing of big data research is that researchers overwhelmingly see their research not as a primarily mechanical, technical, or technological activity, but as a particularly social activity. Their big-picture goals are strongly associated with their social relationships to colleagues and students.

Data Openness

As data intensive research grows, so does data sharing as a practice. When asked about what incentives exist at the institution or in the field to share data, many of the researchers noted that there are cultural expectations within many disciplines and local departments, but at the institutional level, there exists a lack of explicit incentives and rewards for these activities.

> “In the institution I feel like, at least in my department, it's viewed favorably to share code and share data. It's part of the culture. I wrote up my research statement for my tenure application recently and had a paragraph about open science and how we've shared code and data and that being part of what has defined my research. So, I think that, that, that's valued at least, at the departmental level, I don't really know about at the university level. I would be very curious what happens when my tenure application goes to the [redacted] divisional committee and if they care about these things at all,”

> “[The open access project] has gotten me a tremendous amount of goodwill. When I get invited to go give talks, that's what they always want to hear about and talk with me about. It has been really good for my own career as well as something good more broadly for other folks. I don't know if that's how it plays out for everyone certainly. But, I feel like, even though it is not incentivized in the same way, there has been a lot of encouragement along the way, and it's been great for my own kind of professional network building.”

While there may be fewer explicit incentives, one researcher did note that these cultural expectations and professional benefits can have a positive indirect impact on the recognition and reward one receives on campus.

> “I think to the extent that the field gives incentive and that, therefore you get prestige and recognition and that the campus recognizes when you are recognized in your field, That's incentive. So, for example, I do a certain amount of data and code sharing and believe in that and promote it. And that is a fairly buzzword compliant thing in my field

and gets me some recognition and that recognition, I think helped- helps me get promotion and raises on campus."

External and Internal Collaboration:

When researchers discussed collaboration, they were more likely to focus on collaborators outside their research unit or outside the university. Research domains are greatly specialized and this tendency may reflect the reality that researchers share greater affinity with a small and highly dispersed global community. We separated researchers into social science and science fields, but researchers from both broad research areas were equally likely to discuss both external and internal collaborations. Researchers recognize that there is much to gain from collaboration, but also significant risks and frustrations. Several instructive quotes are included below:

> "...<workers at a certain government agency> have a lot more awareness of what's going on. So I find out about a lot of stuff that's going on in industry through my collaborations with people [there]. And being part of the <project> that I mentioned before, I learn a lot about what's going on in industry too, because industry is going to participate as well. I learn about it through meetings, reading, and relationships with vendors…"

> "I encourage <students> to try to think about the competencies you need to learn to be able to really lead this project and not be dependent on some super software developer who's just going to show up at your door and be ready to code all your all dreams into reality because that just generally doesn't happen to the chagrin of some students. Yeah, there have been some memorable encounters, especially with undergrads who said 'I have this idea. I just need the person to code the app for me.' Yeah, I want to ask them, 'have you seen The Social Network? Do you remember the Winklevoss twins? That is you right now.'"

> "COVID has definitely made it harder to figure out who the person is that I should be interacting with and responding to. From a computational standpoint, COVID has been frustrating. Since people are working remotely from home, they're using remote, desktop applications. We've had to do quite a bit with IT and with <supercomputing cluster> [to get this to work]. When you're working from home, you don't have staff meetings where people are all in the same room, and can all get on the same page. Zoom or Webex meetings just are not the same."

Data and Code Sharing - Incentives and Disincentives:

Academia is an intensely competitive environment and a great deal is at stake for researchers. Many researchers work within a context where data sharing is expected, and can be very useful. Researchers understand that sharing is increasingly required by funding agencies as well as individual journals, and those requirements are a powerful incentive to share. Sharing

can be frustrating, but can also lead to further collaboration and creativity. Deciding to not share can hurt a researcher's reputation.

> "The incentives from the [federal] agencies are that the agencies require you to distribute your data and distribute your code, as do the journals now. The incentives are there. There are people I know who are very good in my discipline, who do not distribute their data and code. They are not appreciated by colleagues in the discipline as a consequence."

> "Some journals require you to upload the code on their website. These are the top 5 journals [in the field] as well. So it's a requirement as a publication process. And if you don't agree with it upon submission you have to file an exception. That exception will only be granted if there's a reason for doing so. For posting code, I've never had an exception granted. For posting data, we've received exceptions before because the data might have been proprietary."

> "One thing I am concerned about is sharing data that allows other people to work on projects that I've got on the back burner that I want to work on. There is really a sense of ownership over data because I have done the work in collecting or building the data set, and thinking about the limitations and what it offers. I think you can overshare. There is a real risk that you've done all the work and then somebody else comes in and publishes an article using your data, and you don't really get credit for it besides the citation."

One researcher imagines a publishing model that would allow for big data releases and more access.

> "I think when we write papers, it would be nice if it's part of the paper, every single paper, or at least most of the papers that came out if there were a big data release right? That accompany that paper. Or you could even think of maybe, you know, the big data release itself could sort of justify a journal publication. You can kind of flip it around and say, you know, ...this is an opportunity by releasing, you know, the last 5 years of data. You can have a paper, a journal publication that kind of explains how to use the data."

Another researcher also noted that tensions exist between protecting intellectual property and broader cultural expectations of making things open source.

> "This university is especially aggressive at wanting you to [patent and trademark] the intellectual property, and I'm not against it. It's an interesting road to travel, but I'm very much a believer in open source. I just think that's the way that science moves forward."

Privacy and Ethics

Researchers working with big data at the University of Wisconsin-Madison are navigating a number of issues around privacy and ethics including ensuring reproducibility, the legal aspects of privacy and intellectual property, consent of participants, and disparities.

Ensuring Reproducibility

Four researchers raised issues related to ethics and reproducibility. One researcher emphasized the need for tracking workflows as an ethical consideration when distributing data. “I think we have the same ethical considerations that everybody else does; that it's very easy to be [ethical] if you're disciplined about keeping track of what you've done to the data.” They went on to describe being sensitive to the impact that carelessness in workflow documentation would have on other researchers and how in their own practice they would confirm again and again that collaborators had the correct data set.

A researcher talked about their decision to use publicly available data to ensure that others would be able to replicate their findings.

> “There are a lot of potential data sources, but all of them are imperfect, and some of them are private, and the ones that are quality are private, and I'd rather use an available data source, so that people could replicate my findings and use the data for their own purposes. So, right now I'm merging a number of data sources, which are publicly available.”

Another researcher highlighted how their field is still evolving on issues of ethics and reproducibility.

> “They don't want to publish the structure and then have somebody else write the paper about what the protein does and the field finally came to the conclusion that we're really sorry you have to deposit the structure when you submit the paper that describes it when it's published, because it turned out that the data were not so reliable, that without checks and without the community being able to look at the structure immediately it just made the scientific process not work.”

One researcher spoke about ethical issues around assessing claims being made to ensure they actually reflect the performance of models.

> “It’s hard to do that kind of assessment. We’re learning how to do it. We need to learn as a community how to do it better. It begins to drift into the ethical because you begin to start claiming things about your models in order to get your paper into a high profile journal in order to make it look valuable that just really aren't true. That your model cannot give the kind of performance you are sort of claiming it's giving. So I think there's

> that issue, but I think it's mostly that we just don't know how to make those claims well, not that people are being intentionally unethical."

Legal aspects of privacy and intellectual property

Researchers across different fields are navigating legal issues around privacy and intellectual property as they enter into agreements to access data, make decisions about whether to use copyrighted material in datasets, or come up against restrictions on data scraping. One researcher who works with proprietary data through a data use agreement spoke about the steps they take to meet compliance around security and privacy concerns. They spoke about needing to be physically onsite with the provider and needing to work with protocols to ensure the data they used didn't violate any privacy concerns. "The provider has a place where we're essentially processing the data on location and we have somebody we're working with at the provider that [processes] the data in a secure way so that it can be used by us."

A researcher described the steps they take when working with copyrighted material. "When we're working with copyrighted material I want to be careful about how it's shared. I mean, if it's still protected by copyright then it shouldn't be paid or shouldn't be shared publicly unless we're doing it in a way that's fair use. We're only doing a sample of it, or we've somehow gotten the licensing rights."

A different researcher said they are unable to scrape the subscription databases that would provide quality data for their research because the terms of those databases' policies and contracts don't take into account research needs.

> "I think that's really problematic and it's too bad because researchers like me would never be scraping for commercial purposes, and I think that the policies and contracts of those websites are meant to stop people from reusing and selling the data, which is reasonable, but there are many researchers like me who are just doing this for academic purposes, not for commercial purposes. And yet the policies and contracts prohibit us from bulk downloads or scraping."

Consent of participants

Some researchers are finding themselves having to make new considerations about issues of consent around sharing data and its potential impact. One researcher is taking into account ethical issues of labor and securing permission before sharing.

> "I mean, I think with the [other project], [we] certainly find moments where it's legally permissible to share old magazines because they're public domain, but I try to think about the ethics of who put it in the time to actually scan this stuff. And so I try to ask permission, and if people don't give permission, then just not use it just because they want to respect the labor, especially if it's something they had just like recently digitized."

A researcher faced challenges with differences in understanding of consent in another country, "The whole concept of privacy is something that is very fluid in a place like [country]. You know,

we think that if you're in a case, you know, then you pretty much consented that what's going on in that case is public information, but that's a debatable question in [country], so we had that issue."

A different researcher is facing ethical issues with wanting to share past surveys.

> "The survey data contains theoretically, personally identifiable information and even though we're talking about things that are 30 or 40 years old, the respondents may still be alive, and the original researchers promised confidentiality but in a kind of sloppy or confusing way. Because, you know, at the time the surveys were implemented, IRB sorts of rules were not nearly as, I think, as well developed and robust."

One researcher allowed participants to make their own decisions around privacy in scans of their homes. "We didn't want to be the ones to decide what's private and what's not. So we had the people in the homes label what they wanted to have basically blocked, so things like photographs, phone numbers, because a lot of these scans are good enough you can read text on paper. So, what do we want to have removed from these scans? We never disclose where any of our scans are exactly from to this point, shipwrecks, homes, any of these kinds of things."

Disparities and Ethics

Three researchers highlighted issues around disparities and ethics. One researcher mentioned diversity as a recurring issue, "There's things like, I mean, we have diversity, you know, releasing diversity surveys in the collaboration that comes up every now and then and it's always kind of a hot button topic." Another researcher raised the issue of disparity of access to data and tools, noting that there was a "marked inequality for access to different kinds of data and the tools and techniques to use it." They also mentioned needing a legal team and other resources to secure data entrusted to them by a third party. "There are all sorts of ways in which researchers at some institutions have a lot better access to the 'data revolution' than other researchers and institutions. I'm curious about ways that we could increase accessibility of both the types of resources you need to secure data, and also to use and analyze it."

A researcher described considerations they make about the impact of their approach to data collection in the context of global disparities.

> "With this [other project] [the grant] is partly about enhancing the database, but it's also about globalizing the collections. And I've been trying to think about the ethics there too. In other words we're trying to diversify the collection and add more non-English language material, but we also don't want it to turn into a new form of empire building where we're just like, swooping in and taking people's stuff and then putting it on this site that's hosted here in the US. So, yeah, just trying to have some perspective and kind of approach those things... in a sensitive manner for all those reasons."

Institutional Incentives & Challenges

Challenges in Research Infrastructure & Resources

Throughout the interviews, researchers raised the challenges they face in having the infrastructure and resources to power their research in the ways they desire. Multiple researchers pointed out the high expense, both in terms of cost and resources, to manage computational infrastructure. Multiple researchers also expressed interest in more centralized support.

> “Efficiency of having a centralized resource is so much higher than doing it separately. .... I don't know what the multiple is, but I suspect you could do a little bit of calculation to figure out. You're saving, like, a factor of 2 in money, kind of scale. And efficiency of using your resources. The problem is, if I do it in my group and I waste half of a graduate student's career managing it. The campus doesn't see it. I mean, our research all pays a price, our prestige pays a price, but nobody can see it. Whereas, if you have to pay a staff at [computing service] to manage computers, everybody sees it because it's a cost in front of you. So, I think there are a lot of hidden costs of a lot of people building their own cooling rooms, spending all that money, and maintaining their own computers, and all learning all their own stuff, and hiring their own little system administrators, that the campus doesn't - it feels it indirectly, but they don't see it directly. And then when you centralize it, you see it directly and it looks expensive but if we could quantify that and really make rational decisions, I think we end up supporting a lot more centralized efforts”

Researchers face other resource challenges as well. One researcher described how storage and computational infrastructure solutions that work for small projects often aren't extensible as projects grow, “I guess one thing that these projects have in common is they usually begin small and then they get larger and sometimes you find that the strategies and techniques that worked when they were still pretty small, just like, they don't scale at all.”

Others shared resources issues with staff turnover impacting the ability to maintain software contribution and difficulty with cost constraints and justifications when accessing statistical consulting services.

> “[...] in kind of the maintenance of the analysis and the framework software. One of the challenges we have is just, uh, you know, we get all this data, all this code that gets contributed by transients grad students and postdocs. They hang around for a couple of years, and they take off and get gigs elsewhere and then the 3 or 4 software engineers that are kind of, on the core team are gonna have to maintain all the code and make sure, you know, all the orphan code gets it's love and care.”

> “And then [previous institution] also had a statistics consulting service, which I relied on a lot. And I wish we had that here. It was free for faculty members and students there.

> Here, I think that the price of the statistical consulting is unreasonably high. There, they had it staffed by Ph.D. students in statistics. So maybe a model like that would work to help keep costs down because I know here, the staff members are not students, but I think that the rate is like, 100 or more dollars an hour, and sometimes it's just time consuming. So costs add up very quickly. And it's hard to justify them."

Two researchers raised explicit opportunities for helping researchers scale their work, both centering on services having more defined, set options or concrete recommendations for their storage and computing infrastructure.

> "So I think having skilled IT personnel, right, that can sort of like advise on so say you're starting a new project and it's computationally intensive and you don't [have] the kind of experience in that field right and so should you do that in your local machine or should you push it in the cloud to tap some of the computing resources inside the university, what are those computing resources. Right, so the people who are knowledgeable in that area, I think that's important. [...] so like a researcher needs to pivot a little bit and use the data techniques and so far has been running things on their local machine and so sort of some training wheels on getting, say an AWS account set up or something that seems valuable."

> "I'm not interested at all in figuring out how to store my data. I would like to have some sort of set options and sort of known costs for the data that need to have fast access versus the stuff that's in cold storage and so forth."

Challenges in Public Communication

The challenge of communicating research to the public arose throughout the interviews. One researcher mentioned that public communication and translational work with the public lacks direct and explicit incentives at the institutional level and that this causes tension between their ethical responsibilities as researchers and the common reward and promotion systems in academia. Communication with both peers and the public via social media was also brought up multiple times as an area with which researchers face resource constraints and seek support.

> "I think the biggest hurdle unambiguously is that right now doing a lot of public communication that knowledge translation work is not rewarded in the academy either in hiring or promotion. For example, I have a whole group of colleagues who have basically set their careers aside for the last 7 months to do really important public science communication work around the pandemic. They spend 60 hours a week during this, and none of those efforts are relevant for getting jobs, keeping jobs or being promoted within their jobs."

> "One of my goals is to do better with frequent brief communications, like blogging, social media. I think I'm going to hire a PA, and just make that the PA's top job. That is something that is hard to find the time in the day to do. Frequent, brief, social media communications have been a lost opportunity."

This theme arose throughout the interviews and even raised suggestions for further support for public communications. Multiple researchers across varied disciplines mentioned challenges or interest for further support in this area.

> “And I think most of us, What is missing, is about really to communicate the result with the general public. [...] because every day every researcher or every department has some different research outcome, What kind of, you know, research we should really connect, right, to the research office and to further communication.“

> “You know, it's tricky. We do have this really fabulous person in university communications, [Redacted] [They’re] incredible but that's one person and there are how many thousands of researchers on campus. I think it would have to be something around having a group of people who are well trained to support someone when the Internet turns in a weird way, and some training and support for faculty about if you're going to be engaged in the public communication of research, here are some lessons learned.”

Institutional Opportunities

Researchers suggested a number of interesting ideas and potential opportunities for the institution to support big data research specifically, but that could also be applicable to research support broadly. Specific ideas included more support for building industry relationships, modular IRB approval, and more connection between campus research support offices for data sharing.

> “Maybe another thing just again, these are kind of like vague ideas I'm just throwing out there, not very specific, but stronger connections to industry around these things. Like, I have almost none, but I feel like every company in Wisconsin and the world is sitting around trying to figure out what it needs to do about data science and machine learning and artificial intelligence. Of almost every size. And if academics could be connecting to that, maybe there'd be an opportunity to forge much stronger relationships. I'm thinking, you know, internships for students, funded PhDs, funded research projects, increased relevance of academic research, just stronger synergy between the academic and industrial activities in the state and beyond.”

> “Um, so, I think, yeah, for any research, for instance, the 1st step is we need to go through the IRB for every single project. Um, so we, we understand that, but Uh, from speaking, I think there might be, or there should be some, More convenient way. For example, Because most of my projects would involve the same logic, right? ...Know that okay, we'll have this module and this module will serve as a base layer that actually will support some downstream tasks or some other, or multiple projects. And then actually can save A lot of researchers' energy to focus on the upper level project design.”

> “Well, in some ways, I think I already said that that if there someplace on campus where you know, when you get an NSF and it's a survey, there would be someone at Research

and Sponsored Programs that would say, you know, here are the, the people on campus that can help you post this data set. “

Training & Skills

Helping researchers develop computational skill sets is critical to empowering current and future researchers in their work with big data. Of the interviewed researchers, many had none-to-little formal training. 12 of the 13 interviewees recounted that they had not received formal training to work in data science or working with big data. Most researchers said they were self-taught learning from:

- Projects or hands-on learning,
- Professional and governmental organizations,
- On-demand resources such as books, blogs, YouTube,
- Seminars or local workshops such as the R workshops from the Libraries, the local Carpentries community, and medical imaging workshops
- Human resources such as the Center for High Throughput Computing (CHTC) facilitators, campus colleagues, and the communities that evolve around data science tools

The interviewee who had formal training stated that training came from their graduate program and former institution. Multiple researchers said that they felt that younger scholars and students now have more formal training opportunities available to them through their academic programs, campus resources, and both local and national workshops.

To keep abreast of technological developments that may benefit their work, interviewees mentioned that they learn about the developments through:

- Collaborations with agencies and industry; vendor relationships
- Peers and collaborators
- Conferences, with the following specific topical conferences mentioned: instrumentation, software, digital humanities, digital libraries
- Onlines resources such as pre-print servers, publisher news feeds, disciplinary literature, Google Scholar alerts, Twitter, and hardware and software releases

Computing support

A number of researchers interviewed relied on campus computing resources largely via the Center for High Throughput Computing (CHTC) though some still maintain their own compute clusters at the local level. The breakdown of computing resource mentions is as follows: 5 of the interviewees use or have used the help of the CHTC, while 3 mentioned using their own local (departmental or lab) computing clusters, 1 mentioned using national XSEDE computing support, and 4 mentioned using cloud services of either their own or through other campus service providers.

Internal & External Resources Used

Throughout the interviews, researchers mentioned a number of local campus resources that they rely on for various training, consultations, infrastructure, or other support:

- Local campus computing resources and staff
    - Center for High Throughput Computing (CHTC),
    - Campus Computing Initiative (CCI),
    - Cloud services,
    - Social Sciences Computing Cooperative
- IT professionals whether from DoIT, departmental or project-specific,
- Local data science support via the Data Science Institute & the Data Science Hub,
- Librarians through the General Library System, departmental libraries such as the Law Library, and The Wisconsin Historical Society
- Campus cores, facilities, and services such as
    - The Survey Center;
    - The Small Molecule Screening Facility;
    - Virtual environment 'CAVE'
- Departmentally hosted resources (computing, storage, staff)
- Campus colleagues (specifically, Medical Imaging was mentioned multiple times)
- Students who assist across many tasks and during all phases of the research lifecycle
- Institutional storage services such as Google Drive, Box, off-site storage, etc.

External support and resources that researchers mentioned included:

- Training from professional organizations, government organizations, and conferences
- Online resources such as YouTube videos, blogs, Coursera courses
- Peers, collaborators, colleagues (such as applied mathematicians or statisticians and those working with similar methods or data)
- Pre-print servers for publications and public data repositories for data sharing; and specific repositories such as The Materials Data Facility
- Former institutions' digital scholarship resources
- Public feedback on shared data

Opportunities for Training and Skills Development

An important note is that while many of the interviewed researchers named resources from all across campus, three researchers from a less traditionally data-heavy discipline stated that they did not know who to go to for big data or computational support on campus, which suggests that there may be opportunity for outreach and support in such areas. One of these researchers also stated that they had used digital scholarship services for training and support at their previous institution which may be indicative of what types of services or disciplinary-focused support researchers from non-STEM disciplines may be looking for.

An extremely fruitful portion of the interview came when the researchers were asked to imagine what training would be useful considering the evolving trends in their fields. Machine learning was mentioned multiple times by researchers as an area for further training, with concerns raised over both its own and other computational tools potential to be used as a ‘black box’.

> “as the software becomes more powerful and easier to use there is a greater and greater temptation to use it as a black box. And also to mix metaphors horribly: to find a piece of software that makes for a nice hammer and then hit all your nails just use whatever you've got your hands on to solve all your problems. And what worries me is especially on the machine learning side all of these approaches have some assumptions built in. Either assumptions about the mathematical structure of the data or implicit assumptions in the form of the training datasets and without equipping students with enough basic mathematics and an intellectual framework to understand those assumptions and limitations, even if they don't understand really in detail how every last little bit of the algorithm works.”

Data visualization was raised twice, noting the importance of being able to communicate research effectively.

> “I think a training [...] in data visualization. So I think how to extract useful and relatively basic information from large data sets would be really useful because our research has to be communicable to people that don't have a lot of methodological training oftentimes. The standard way of presenting, sort of data results and regression analysis, and that kind of thing, correlations or coefficients is not not always the most, I think, effective way of communicating sort of what we found the relationships that we've uncovered.”

Another researcher raised that data management and reproducibility were important to have for researchers, especially those working with bigger teams, bigger data, and interdisciplinarily.

> “...and some of it is honestly the adoption of stuff for reproducibility and visualization and so, I'd give all of us training and how to start using things like Git and R Markdown. And then you'd want to have a bunch of things related to best practices with respect to organization, because I think everyone has a way of managing data that they've used in the past with smaller teams, but now, as you start tackling bigger problems with these really interesting forms of data, there's often a lot of people involved, and the way to harness this data in a really interesting way often involves interdisciplinary collaboration and so that means working across computing systems across programming platforms and so to me, the way to do this work well, is to keep it organized and reproducible, and I think some training. But also, norms, adoption, these practices would be really, really important.”

Two researchers mentioned that communities for peers to share and learn from each other would be useful. Another mentioned that campus-wide communications and sharing could be helpful. In the same vein, one researcher mentioned that cross-domain and interdisciplinary opportunities would be beneficial.

> “And so, like, sort of lunchtime talks or that kind of thing, you know, and this helps us solve that chicken and egg problem I was talking about, right? Until you know what the methods are capable of and what their weaknesses are, it's hard to know how they'd be used. And so I think that kind of basic introductions to some of these developments and how to begin thinking about research projects in light of those developments could be, I think, really quite useful.”

> “I guess I do think there is an opportunity for sort of cross-domain things that could be very fruitful on a campus like ours in particular. [...] So, creating that education across those disciplines. Same thing for computer science. It's hard for me to get access maybe to people in computer science, doing innovative things and statistics doing innovative, like cutting edge research, and trying to bring that into what I'm doing. There's always a bit of a divide there and helping to educate those groups about each other's problems and opportunities is, I think, useful and good for the campus, because we get kind of the synergistic plus of doing that, otherwise they go do it with other people. I find other machine learning experts somewhere else are into, the machine learning experts find other applied domain-specific people to work with at other campuses or something.”

In line with developing communities, another researcher stated that access to human resources for assistance was useful, a theme that arose throughout the interviews. They also felt that developing literacy in data science was important, which echoes the earlier calls in this section for helping researchers understand the methods, limitations, and algorithms that are being used in data science tools, “...so for graduate students, and for faculty, I think, access to people and literacy is key. If you can get pointed in the right direction, It's actually pretty fast as I said, to learn this stuff. But the space can seem daunting.”

The Data Science Hub’s Carpentries trainings were mentioned multiple times as beneficial. Support was also voiced for tutorials and workshops that could help make improvements in researcher workflows or walk researchers through common tools used with big data.

> “The sort of hands on best practices training in software and data offered by the carpentry workshops is essential and at this point, I just make all of my graduate students do it,”

> “I think even my group always benefits from, like, well-done tutorials or even workshop resources, and sometimes real workshops, that are bite size and introducing some technology that can improve our workflows.”

Other growth areas for training and skills development included further statistical support, understanding limitations of descriptive metadata, providing hands-on opportunities for undergraduates, and game engines and computer graphics.

> “...the biggest challenge is trying to learn and get up to speed on the appropriate statistical techniques. And then the commands that they're implemented with in the

programs. I'm not trained as a statistician and so I have to sort of self learn on my own, what the most appropriate techniques are."

"I think what would be great for the whole field is if there's some awareness of the value and kind of the limits and problems with descriptive metadata and how there's good things that can come from having controlled vocabularies and authoritative metadata, but how there's also biases that are built into that and so if people working with these collections or bigger datasets had kind of an awareness about that, then I think they could make better interpretive judgements themselves. But then also, if they're able to contribute metadata, do it in a way that [tries] not to reproduce those biases and does it in a way that is helpful to other users"

"I'll just jump into like, a very specific example, but I'm not sure this is a big data issue. For our virtual reality field we really are looking for are people that utilize game engines. ...So, we have on campus one computer graphics course…. It's a huge field and so that course is very helpful for establishing foundational practices. But it would be right if there would be trainings to help people through. Interactive ways, I guess, to deal with big data here on that, back to big data, interactive ways to visualize big data."

"... so I have about 20 or 30 students every semester where we just engage them in different research projects and that's something we're trying to build out to other schools and grow. And I think that's a key way to tackle a bunch of problems at once with undergraduates, it's to inform them about machine learning and data science. It's to get them using computers and it's to give them hands-on experience and it's to give them sophisticated open problem solving experience. And I call it research experience because that's the word for it, but I actually think it's the wrong way to think about it. Research makes it sound like it's the academic-y. But research is just a word for doing a hard open problem that you don't know the answer to."

Conclusion: Observations and Recommendations

Datasets have become one of the most important products that universities create. Data are valuable resources that are expensive to collect and complicated and expensive to curate. Given the expenses related to data, there are strong incentives to share and reuse data. Empowering researchers' ability to effectively work with big data, to produce new insights using computational and emerging methods, and to share the various outputs of that work is critical for furthering the research mission of the University. This study of University of Wisconsin-Madison researchers working with big data has identified a series of issues that face big data researchers as well as excellent opportunities to support research using big data and data science methods.

We have observed:

- UW-Madison researchers face a diverse set of challenges when trying to obtain access to research data for their projects. Many researchers obtain datasets from external sources and organizations, and this data requires the protection of external and industry partners' proprietary interests. Merging multiple available data sources is also difficult due to messy or poorly structured data.
- Researchers at UW-Madison working on big data projects are very interested in and concerned about issues related to the dissemination and sharing of data, coding, and outputs such as papers and conference presentations. Researchers overwhelmingly see their research not as a primarily mechanical, technical, or technological activity, but as a particularly social activity. Their big-picture goals are strongly associated with their social relationships to colleagues and students. When researchers discussed collaboration, they were more likely to focus on collaborators outside their research unit or outside the university. Research domains are greatly specialized and this tendency may reflect the reality that researchers share greater affinity with a small and highly dispersed global community.
- Academia is an intensely competitive environment and a great deal is at stake for researchers. Many researchers work within a context where data sharing is required by funding agencies as well as individual journals.Those requirements are a powerful incentive to share. Sharing can be frustrating, but can also lead to further collaboration and creativity. Deciding to not share can hurt a researcher's reputation.
- Researchers working with big data regularly encounter a number of issues related to privacy and ethics. Researchers emphasized the need for tracking workflows as an ethical consideration when distributing data. Researchers also expressed a concern with ensuring the reproducibility of results, by using public data or making data public. Researchers also face further challenges around security and privacy concerns. There are inherent problems working with raw data that contain personally identifiable information. Protocols to ensure the data used does not violate any privacy concerns can be complex. Some researchers are finding themselves having to navigate complex issues of consent, for example there are challenges with differences in understanding of consent in other countries. Curated data from several decades ago may have been collected during other eras when IRB protocols did not exist.
- Researchers face a variety of logistical problems related to working with big data, including obtaining infrastructure and resources to power their research, high expenses for resources to manage computational infrastructure, difficulty arranging statistical consulting, difficulty scaling projects, and staff turnover. Researchers expressed interest in more centralized support.
- Of the interviewed researchers, most had none-to-little formal training in working with big data or data science. Most researchers said they were self-taught. Helping researchers develop computational skill sets is critical to empowering current and future researchers

in their work with big data. Younger scholars and students now have more formal training opportunities available to them through their academic programs than had existed previously, but additional formal campus training and resources would be valuable.

- The University of Wisconsin-Madison has ample ripe opportunities to support research using big data and data science methods through:
    - Better, more organized and thorough trainings for all researchers. In particular, supporting access to training and shared resources for those from traditionally non-data intensive disciplines.
    - Increased centralized access to cyberinfrastructure.
    - Increased access to dedicated human expertise in statistics, data visualization, cyberinfrastructures, and related topics.
    - Incentivizing open sharing and publishing.
    - Improved campus connections and workflows for research support.
    - New platforms and venues for discussions of methods, nascent ethical challenges, and other issues.

Recommendations:

Throughout the interviews many clear opportunities arose for improving the ways in which the UW-Madison campus supports big data and data science research, some of which were named explicitly by the researchers when answering the interview questions. This portion of the report contains a review of some of these suggested opportunities along with some that were illuminated during the analysis and discussions of the previous sections.

The authors acknowledge that multiple research support units on the UW-Madison campus have current responsibilities for portions of the recommendations listed in this section. We encourage these campus units to take a collaborative approach and framework in responding to the recommendations of this section and brainstorm areas of overlap, opportunities for collaboration, and ways to leverage one another's strengths and skills to provide inclusive and visionary support services.

- *Open Data, Dissemination, and Sharing*

  The value of openness and sharing arose as a significant theme throughout the interviews. Building rewards and incentives for sharing into formal promotion and tenure procedures could be a beneficial way to recognize work that researchers are already doing, the cultural expectations of their disciplines, and reward values and work that align with the Wisconsin Idea.

  Researchers are faced with a confusing array of institutional, granting-agency, and publication-related data-storage requirements and obligations. Researchers are also presented with an equally confusing list of storage options in a variety of locations with variable technical parameters. Much more work should be done at a national level to

thoroughly and systematically standardize and organize storage requirements, procedures, and systems across all academic institutions.

Researchers are very concerned about losing the opportunity to analyze the data that they have generated. As more institutions and organizations require datasets to be posted for reuse, there is a risk that protections for data creators will become disorganized. There should be a standardized minimum level of protection for data creators across institutions and organizations. More formal recognition and reward could also help address this.

- *Institutional Opportunities*

  Researchers face financial, staffing, and scale constraints when powering their research. Some researchers suggested there was benefit to more centralized services. One interviewee also felt that access to skilled IT personnel would be helpful for getting researchers up and running.

  Other explicit recommendations from the interviews include suggestions for more support in public communication of research findings, more support for building industry relationships, modular IRB approval systems, and more connection between campus research support offices for data sharing.

- *Privacy and Ethics*

  When asked about ethics and privacy and their data, some researchers at UW-Madison raised reproducibility as an issue in their field. Two researchers said that practices in their field were still evolving specifically regarding data sharing and assessment of models. Those supporting big data research should be mindful of conversations around best practices in the fields of the researchers they support so that they can help researchers make the best decisions they can in a shifting landscape and advocate for their needs.

  Due to privacy and intellectual property rights, some researchers are encountering the need for legal agreements, licensing rights, and subscription database terms of use in order to access data. Ensuring that our research infrastructure meets the needs of researchers working with external data providers wishing to protect their data due to proprietary or privacy concerns will continue to be important. Additionally, recommendations include increasing education around data and licensing throughout the research enterprise as well as advocacy for increased access for researchers to data for noncommercial use.

  Accessing data for research at UW-Madison is increasingly international in scope and this is an opportunity for the research enterprise to consider the access of data from institutions abroad by those in the United States alongside issues of consent.

Researchers at UW-Madison are also aware of disparities in access in terms of diversity among colleagues, and access to institutional and regional resources. The researchers we interviewed at UW-Madison are making ethical choices in a global context and we suggest that opportunities for researchers to discuss and learn from one another would be beneficial.

- *Training and Education*

  The interviews generated many ideas for increased training and educational opportunities. Researchers shared a number of specific recommendations for workshops, communities of practice, and ideas for training topics.

  Helping researchers better understand the emerging and popular data science methods and tools. Machine learning as a topic was suggested explicitly multiple times. Interviews suggested that this is important for helping researchers understand how such tools work and what their limitations are, in order to avoid using them as ‘black boxes’ and better understand their results.

  Researchers are also looking for help with communication of their research, both externally and internally. There was interest in more training on data visualization to enable them to better communicate their findings. There was also interest in developing communities for researchers to be able to share with one another especially across campus and across disciplines.

  There were multiple suggestions for other training and education topics including, improving researcher workflows, workshops for common big data tools, further statistical support, understanding limitations of descriptive metadata, providing hands-on opportunities for undergraduates, and game engines and computer graphics.

  One particular opportunity to note here is that researchers from less traditionally data-heavy or from non-STEM disciplines may benefit from more targeted outreach and support for big data and computational work. These researchers may be looking for more targeted disciplinary services or digital scholarship services, as such they may not know where to go on campus for the support they need.