Georgia State University
ScholarWorks @ Georgia State University

University Library Faculty Publications | Georgia State University Library

10-1-2021

Supporting "Big Data" Research at Georgia State University (GSU)

Kelsey Jordan

Bryan Sinclair
Georgia State University, bsinclair@gsu.edu

Mandy J. Swygart-Hobaugh M.L.S., Ph.D.
Georgia State University, aswygarthobaugh@gsu.edu

Jeremy Walker
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/univ_lib_facpub

Part of the Data Science Commons, and the Library and Information Science Commons

Recommended Citation

Jordan, K., Sinclair, B., Swygart-Hobaugh, M., & Walker, J. (2021). Supporting "Big Data" Research at Georgia State University (GSU). Local report from "Supporting Big Data Research" ITHAKA S+R project.

This Report is brought to you for free and open access by the Georgia State University Library at ScholarWorks @ Georgia State University. It has been accepted for inclusion in University Library Faculty Publications by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

Supporting “Big Data” Research at Georgia State University (GSU)

Local Report of GSU-specific insights gleaned from participation in a multi-institutional research study coordinated by the Ithaka S+R research and consulting organization

Conducted and reported by the following GSU University Library Faculty members:
Kelsey Jordan, Bryan Sinclair, Mandy Swygart-Hobaugh, Jeremy Walker

Contents

Study Objectives 3

Study Methodology in Brief 3

Key Insights re: "Big Data" Research at GSU 4

Defining "Big Data" Research at GSU 4

Infrastructure – Needs & Challenges 6

Data Access – Needs & Challenges 8

Data/Code Sharing – Needs & Challenges 10

Learning & Professional Networking – Needs & Challenges 12

Paths toward Improved Support of GSU "Big Data" Researchers 15

Appendices 16

Appendix A: Institutions in Ithaka S+R "Supporting Big Data Research" Study 16

Appendix B: Study Methodology in Detail 16

Appendix C: Final Codebook 21

Appendix D: Hierarchical Chart Visualization of Coding 22

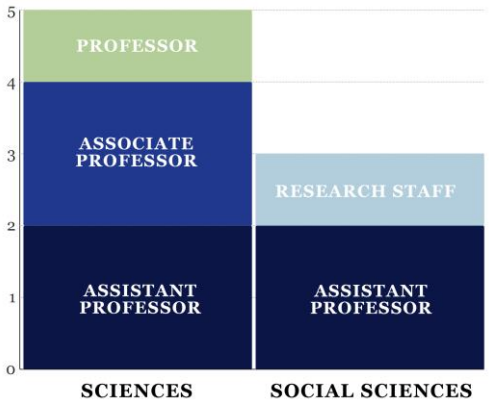
Study Objectives

From Summer 2020 to Summer 2021, a team of Georgia State University (GSU) University Library faculty[1] took part in a multi-institutional research study[2] coordinated by the Ithaka S+R[3] research and consulting organization to examine the research support needs of faculty doing "big data" research.[4] This report offers the following insights from participation in the study:

- identifies the key research support needs and associated challenges faced by GSU faculty who engage in "big data" research
- offers possible paths toward improved support of GSU researchers in this area that capitalize on the Library's strengths and have feasible return on investment

Study Methodology in Brief

The team conducted semi-structured interviews with eight GSU researchers representing a diverse cross-section of academic fields: Accountancy, Chemistry, Communication, Computer Science (2), Physics & Astronomy, Policy Studies, and one cross-disciplinary researcher with joint appointments in Gerontology, Psychology, and Neuroscience.[5]



[1] GSU University Library faculty research team: Kelsey Jordan, Data Services Librarian [formerly with GSU]; Bryan Sinclair, Associate Dean of Public Services (bsinclair@gsu.edu); Mandy Swygart-Hobaugh, Research Data Services Team Leader (aswygarthobaugh@gsu.edu); Jeremy Walker, Data Services Librarian (jwalker184@gsu.edu).

[2] See Appendix A: Institutions in Ithaka S+R "Supporting Big Data Research" Study for other participating institutions.

[3] Ithaka S+R (sr.ithaka.org), a not-for-profit research and consulting organization that helps the academic, cultural, and publishing communities, coordinated this parallel effort and provided guidance on research methodology and data analysis.

[4] Per the scope of this study as delineated by Ithaka S+R, "big data" methodologies refer to research conducted with data that is high in volume, velocity, and variety – that is, with large, diverse, semi/unstructured datasets that typically require high performance computing infrastructure for data processing and advanced computational methods for data analysis. These methods are used by scholars across the STEM, social sciences, and digital humanities, and may include machine learning/artificial intelligence, data/text mining, modeling, analytics, informatics, and other data science techniques, and the data types may include but are not limited to physical specimens, text, or images. **NOTE:** Throughout this report, the GSU study team denotes "big data" with quotations marks to reflect our critique of this term as representing a cohesive, demarcated, universally-defined and agreed-upon research area.

[5] See Appendix B: Study Methodology in Detail for detailed discussion of participant recruitment, data collection, and data analysis processes.

Key Insights re: "Big Data" Research at GSU

Defining "Big Data" Research at GSU

Throughout each of the interviews, a variety of notable themes emerged with respect to the definition and scope of the term "big data" as it applies to academic research.

First, the concept of "big data" varied notably in both definition and importance between different researchers. GSU1, GSU4, GSU6, and GSU7 explicitly noted that a common characteristic of their research is the large sample sizes used in their research. GSU4 noted that in a less-productive academic year, they produced 500 terabytes of data. GSU6 described a project in which they had so many samples that they could not decompress the dataset due to the sheer volume of data samples. Conversely, GSU2 and GSU8 described projects in which the number of samples available for study was relatively small, but where each sample represented terabytes and sometimes petabytes of data. GSU2 provided the following discussion of this point:

> So, when we talk about big data and when we read it in popular magazines, popular press and how it started, it's sort of Twitter, right? It's a lot a lot of small cheap samples and it's like 140 characters or whatever, but there are millions of people who create them for free. And even with images there are a lot of free sample, free and cheap samples. In brain imaging, the big data is kind of flipped on its head because our data is expensive and each sample is like very expensive to collect. And it's very heavy, it's not like small. It's 60 to 100,000 numbers per human brain for per time point. So it's big data even there to start when you have even a few subjects that went through the research program you're going to have a lot of data, although you don't have enough numbers for statistics to kind of say, oh, now I can tell everything about everyone on the planet. Although with like millions and billions of tweets, you may draw those inferences. So it's kind of a paradoxical situation. [GSU2]

According to interviewees, another significant component of "big data" research revolves around how data is analyzed. Specifically, GSU1, GSU4, GSU8 explicitly noted that in order to conduct their analyses, they needed to use or develop analytic code that was suited to parallel processing. GSU8 specifically noted that this posed additional challenges when it came time to translate analytic code into a structure that could work well with university high performance computing systems:

> I've run a number of analyses on all the various high-performance clusters. The challenges are things like learning to use the high-performance cluster, because whenever you're starting to work with the analyses that you can do on 25 or 30 datasets, right. Neuroimaging datasets are complicated and you need powerful computers. You can't do them on a laptop. Right. You need a Linux system. Fine. We've all learned to work in Linux systems and to run the analyses we need to run on 20 to 30 subjects. You then want to scale it up to 800 to 1,000 subjects. You're going to need a high-performance cluster. But the approach that you were using, the commands that you were using, the interface that you were using is no longer what you can use when you're starting to work in a parallel computing environment, things like that. There's a whole new interface that comes with that. And many of our pipelines are not built to work in those environments. They were built to work on 50 people, you know, and you start parallelizing it doesn't...so there's been a lot of work. [GSU8]

Lastly, interviewees provided discrepant responses and feedback regarding their perceptions of "big data" research and what it means for them in their field. GSU6 and GSU7, both social science researchers, described advances in "big data" research tools and methods as being a major boon for their field. Conversely, GSU5, a computer science researcher, expressed a disregard for "big data" as a "buzzword":

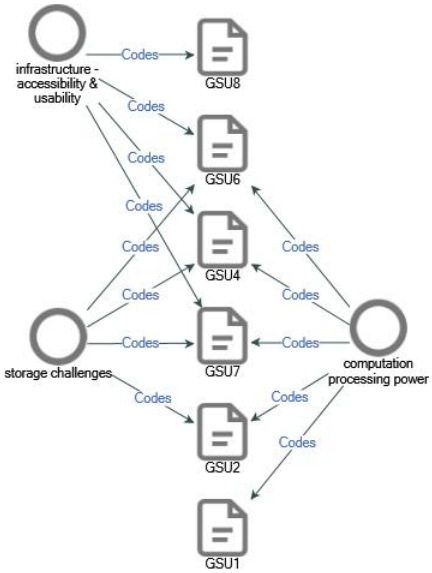
> So that is a buzzword created as many buzzwords have been created in the past. And I think in the future that's going to die as many others have died. But the data is still there. Big or small, and we will still have the same challenges that we have over the decades, they will still be with us and we have to still solve these problems. And that's why I'm just saying it's a, it's a buzzword, but it doesn't mean anything. [GSU5]

GSU2 described initiatives in which the goal was to develop pipelines, which would enable researchers to conduct analyses on datasets without needing to worry about the nuanced and niche technical elements of working with extremely large data samples. Finally, GSU1 noted that while there are constantly new developments with respect to the technologies and methods for analyzing data in bigger and more complex ways, they expressed hesitance to align their research to these advances:

> So this is maybe one of the points that I've been thinking about is how to be able to handle this big amount of data and extract the meaningful information in a *smart* way, but *automated* way so that we're not limited to how much we can actually process ourselves. Because right now we can only do as many systems and molecules and we can only analyze as much data as we're able to physically do it ourselves. And it's not that automated yet. And we're hoping to automate it more and more. And I think this is probably the direction where a lot of people are moving in our field is automating this process as much as possible to be able to compare with larger datasets and do more experiments and calculations at the same time. [GSU1]

Ultimately, the ways in which interviewees described and understood "big data" defied easy and straightforward definitions. Interviewees' understanding of "big data" research was largely contingent on the peculiarities of their specific research areas.

Infrastructure – Needs & Challenges



Many of the interviewed researchers cited a variety of strengths and challenges with respect to the computing infrastructure available on campus and what was needed for research and teaching.

Multiple researchers reflected positively on their experiences using high performance computing (HPC) resources and engaging with staff who administer and support HPC infrastructure. Specifically, they noted that HPC support staff have been exceptionally responsive to researchers' needs and have been instrumental in advancing research projects:

> So first let me say that [GSU staff member in HPC unit] is amazing. His help has been invaluable, seriously… And they [students] had to call on [GSU staff member in HPC unit] every time, like "Can you fix that for me? Can you fix that for me?" And just to say, he usually would have solved it in like an hour or two, like immediately. They write in Slack, he asks what exactly is the problem, then he solves it. [GSU7]

> I work with [GSU staff member in HPC unit] and [GSU staff member in HPC unit] and the folks over at HPC and I've got a grant with them that helped fund buying a whole bunch of new clusters and things like that. [GSU8]

> We work with [GSU staff member in HPC unit] and [GSU staff member in HPC unit]…they're constantly on call. Well, Slack window with one of them, is always open…and yeah, we cannot work without them. [GSU2]

The primary challenges that researchers cited centered on limitations related to the costs of storing and moving large datasets, the relative lack of necessary computational power to conduct advanced analyses, and the relative rigidity of the campus' advanced computing ecosystems.

Multiple researchers mentioned infrastructural challenges associated with storing and transferring research data. GSU2 noted that a single sample of data in their research was approximately one petabyte in size. GSU4 noted that their research will generate between 500 terabytes and multiple petabytes of data each year. Consequently,

researchers also noted that there are notable financial challenges associated with buying or licensing data storage infrastructure. Although cloud computing services like Amazon Web Services (AWS) are convenient and scalable, the cost and impermanence of these services requires researchers to focus on developing local infrastructure:

> We can collect even more data of higher, finer resolution. And we're getting to petabytes per brain and petabytes on AWS Cloud will cost twenty thousand a month to store. [GSU2]

Beyond simply storing data, researchers noted that being able to access and transfer data in a timely manner also represented an infrastructural bottleneck with respect to their research:

> ...we have about 100 researchers of different levels, and about 50 of them are daily accessing this data and are writing their models. So it's back and forth on the same network, on the same cluster. And it's, it's like terabyte traffic per second. And that creates a lot of issues for the IT actually. And for us too because we clog the system. [GSU2]

> When I was setting up the computer that I mentioned, I had to buy a 10 gigabyte cable, so otherwise the computer system would have been fairly useless. I couldn't move data anyplace else. So it would be great if the university were going to upgrade the speed of Internet connections everywhere, that would be a help for everybody. [GSU4]

With respect to computational power, researchers described a variety of ways in which their research required access to abnormally powerful computational resources. GSU1 noted that some computational research could be facilitated using extant high performance computing resources on campus. However, some researchers noted that they access additional computational power not always available on campus. GSU1, GSU4, and GSU8 all explicitly mentioned using research funds to use XSEDE supercomputers. Researchers also provided examples of needing to use grant and research funds to purchase additional powerful computing resources:

> So in the United States, I run on XSEDE machines and have also been purchasing and administering a small supercomputer on GSU's campus. Think that this is a small supercomputer, as supercomputers go, but as I understand it's the largest computer system at GSU right now. So this computer system we're investing in... Currently, it's actually, I think, installed under [building with centralized data systems]. The data center is 30 nodes that are Intel Skylake Silver nodes. [GSU4]

> And luckily with the money from the grant, I was able to buy stronger machines, which are really satisfactory. I mean, 500 gigabytes of RAM and 48 cores. I'm well set for my analysis. [GSU7]

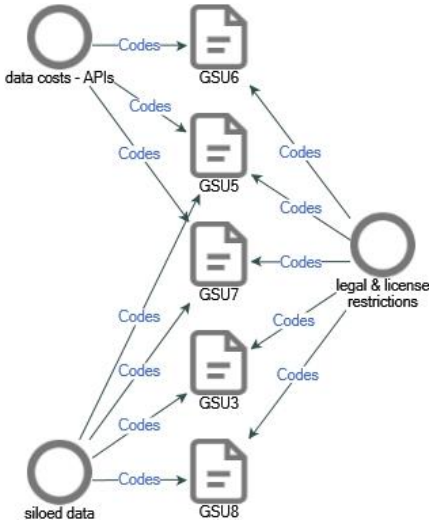
In multiple interviews, researchers noted that existing campus computational infrastructure presented a few peculiar challenges and barriers to research and learning. GSU6 and GSU7, mentioned that improvements to the accessibility, usability, and malleability of existing high performance computing resources would greatly aid both their individual research and in their work with students in the field:

> I would certainly benefit a lot from maybe a more open ecosystem of things. Like, so, like more interactive computing resources where I could do stuff on GSU servers where I can do it in real time instead batch computing. [GSU6]

> I think the main problem, and I don't know how to solve it because I've never tried to manage something on that size and scale, but that's the malleability of the system. I find that students usually need to tinker packages. They need to install packages. They need to install a wider range of packages that some of them require tinkering. And this tinkering is something they're unable to do within the DICE system [local high performance computing system]. And I think that's why a lot of my students started working with that kind of faltered off it. [GSU7]

Data Access – Needs & Challenges

Of the interviewed researchers, their data use was essentially as follows: (1) the science researchers were working with biomedical and scientific data, (2) the computer scientists were developing computational processes to run against test datasets, and (3) the social science researchers were analyzing social media or traditional news media. The majority relied on secondary datasets, with a minority collecting original data. Several researchers noted the unique challenges they face as "big data" researchers in terms of data access, citing recurring woes related to infrastructure, data licensing, and costs.



GSU3, GSU6, and GSU7 noted challenges with securing and adhering to data licensing agreements from multiple and varied stakeholders, noting how these agreements often limit the volume of data accessible altogether or in each access iteration. Relatedly GSU5, GSU6, and GSU7 noted the high cost and often siloed licensing agreements

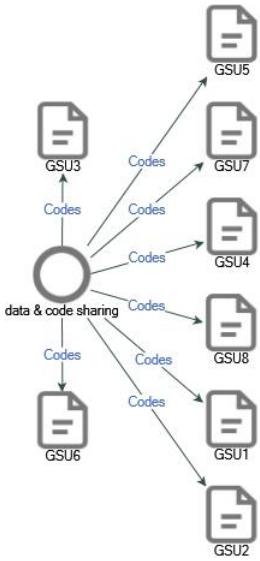
associated with secondary dataset access from social media and other data outlets. To overcome these challenges, GSU7 suggested that perhaps the library or other campus entities could support institution-wide data licensing and access agreements:

> What I see in many other universities, there are enclaves and departments where professors in that department may have a firehose access to Twitter, but they don't or can't share it with the rest of the community in the university. And this is something which I know is the Hail Mary because it's costly [securing a campus-wide "firehose" access to Twitter]. I'm well aware of the cost of something like that, but I think something like that would push the work of a lot of potential social scientists in a very strong way… Oh, and again, I will say again, better infrastructure for data acquisition in GSU library. If this is something you can solve, that will put us ahead of even bigger, better institutions that don't have anything like that. And this is something which, if we can get done, it's basically, it's fertilizing papers. Like basically the second that you have this Twitter data, you're going to find [researchers] who are going to take this Twitter data and going to churn out papers from it. And I think that's, in the end, the goal of every department, because every department is actually being evaluated by their research output. And I think that's one of the best ways to get research output. Because data creates output. [GSU7]

GSU4 and GSU7 noted that collaborating with researchers at other institutions with better financial and infrastructural resources was common practice to facilitate data access. GSU7 explicitly noted that, while collaboration is in part driven by necessity in order for them to access the requisite data for their research, the added benefit of these collaborations include an increased breadth of theoretical and methodological expertise:

> For myself, when I work with data, I usually have to work with coauthors who have access to the data. So I find myself working with people in where I got my PhD… [with] other institutions, other departments, and even other countries. We're working with whoever has the data and the theories and the idea that we care about or whoever has the bit of method that we don't have and we want to borrow. [GSU7]

Data/Code Sharing – Needs & Challenges



Many of the researchers expressed an interest in increasing the amount of data and code they shared but noted that there were often significant hurdles to doing so, such as legal boundaries, high cost and physical challenges of transferring and storing data, and the preparation and maintenance needed to support useful open-access code. Some researchers felt that there were strong ethical and personal motivations to share, which others expressed a lack of incentives and concern over being properly credited. Platforms such as Git, GitHub, and DataLab were cited as common sharing tools.

The researchers noted that sharing data and sharing code are different processes, each with their own challenges. GSU7 viewed sharing data and code as a multifaceted positive because they were interested in promoting open scholarship and having their work cited by more people. On the other hand, GSU6 expressed concerns about "scooping", "maintaining one's...advantage over competitors in a space", and not being credited when sharing publicly:

> Yeah, I mean, I do think right now there's kind of the pressure now coming from the journals, from open access, and maybe that's more of a stick than a carrot. And so it's an incentive, but it's not a positive one. The other pieces, of course, citation activity, right? If you provide a dataset lots of people use. That said, I do think that sometimes, like if you provide, if you provide like a clean set of data and people use it, chances are they're going to cite you. I do think that, like, if someone looks at your code, there's a, people don't like to be perceived as having done work that they can't do themselves. And so, I do think that there is a strong chance you put your code out there, that no one will credit you for it. They'll look at it, they'll reproduce it, and then they'll move on. And so and to some extent, that could be true, I think with data, too, if it's publicly available, it was "Oh, I scraped Seeking Alpha [crowd-sourced content service for financial markets] myself," and maybe did some of that. Right. But so I don't know that there are actually that many positive incentives because you have to be worried that people simply won't bother to credit you. [GSU6]

GSU1 said that the granting agencies they worked with seemed more interested in sharing code than sharing data, because a large volume of the data generated is "junk" that wouldn't be useful to other researchers, versus the smaller amount that is culled for meaningful analysis. GSU5 and GSU7 both do research that involves a significant amount of original coding, and they noted an increasing emphasis in their field on sharing and publishing code when possible, with GSU7 noting both extrinsic and intrinsic motivations for sharing code:

> There is more external now, you see more and more external calls for that. But that's something I was taught to do. And I really think research is often useless without that. Like, if I develop, most of my papers develop a new method, not only a new kind of investigation. I have a new way of measuring internationalization. If I don't share my script with everyone, then my paper is kind of useless. You can read it and learn from it, but you can't really apply anything from it to your own research. And you get more citations if you put your methods out there. [GSU7]

While citing the positives of sharing code, researchers also noted that sharing code often requires a large amount of preparation and documentation before it is shareable, and subsequently requires management and upkeep to remain useful in an open-source environment – i.e., the work does not end once researchers upload the code.

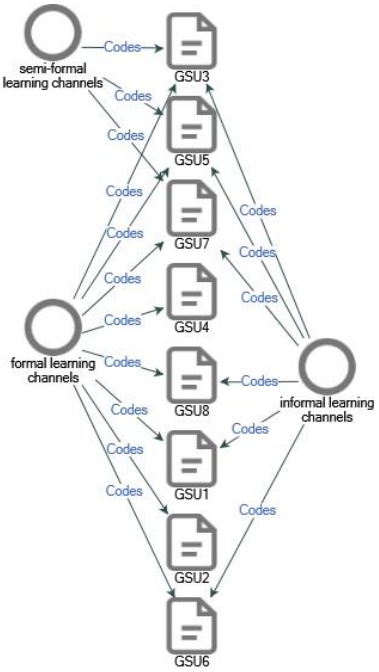
In terms of sharing data, the interviewed researchers noted that datasets can be difficult to share because of the volume of data, and that the best solution is often to have both parties apply for an allocation on the same computing system, so the data does not have to be transferred. In terms of making "big" data open to the public, as grants and journals are increasingly pushing, GSU4 noted the impracticality of those policies:

> I think sharing...this period of history we're falling in, this, into this period of time where things are ill-defined. Still ill-defined. So there are lots of grants or publications, journals that will require that you make your data publicly available. And they don't really define at this point what that means or how you can do it. Um, as I was telling you, I, in a very productive year, might produce a petabyte of data, and there's no place that will let you save a petabyte of data and make it open to the public. So usually the raw data is something that it's impossible to follow those rules about. The kind of data that I do share then in those cases is the data that goes directly into making figures or papers. And that's small amounts of data. This is an issue that might really impact on sort of big data practices just because there's no place to share big data, there's no functionality to do that... Raw data is nearly impossible to share, just because it's so large. When someone wants that, it's something that you can usually produce on request. But there's always this question of how could you transfer it to somebody? It's very much larger than you could email, it's very much larger than you could put on something like Dropbox. It might take years to upload to a box, even if the Dropbox account was large enough. And so I don't have a good solution for sharing large amounts of data. [GSU4]

In addition to infrastructural barriers to data sharing, some researchers [GSU3, GSU5, GSU6, GSU7] cited that licensing agreements tied to accessing secondary data in large volumes via APIs or wholesale data purchase often prohibit data sharing. Similarly, GSU8 works with an international research group and commented on the legal and cultural difficulties with sharing data between countries. While some countries have a

more permissive attitude toward data sharing, the European Union's General Data Protection Regulation (GDPR) is one example of a law that makes it more challenging to share data. However, GSU8 expressed an interest in making data more available despite these challenges, saying, "There's a number of motivations. One is, it's ethical to share data, right? It's unethical to hang on to your data and never share it."

Learning & Professional Networking – Needs & Challenges



When asked how they learned "big data" methods and tools, and what they recommend to others (colleagues, students) for learning/training, the researchers noted a combination of informal, semi-formal, and formal channels. Being "self-taught" prefaced much discussion of informal learning, and that an immediate need-at-hand largely drives informal learning. Thus, informal learning primarily took the form of googling, consulting trusted resources such as Stack Overflow and other online forums, using YouTube videos, etc. Similarly, the researchers mentioned online webinars and courses as a go-to resource to build foundations in tools or methods. Some researchers lauded the benefits of semi-formal networking for learning, such as peer groups, meetups, and so on. GSU7 particular advocated for the benefit of networking opportunities within and across disciplines, specifically praising an event collaboratively organized by the GSU High Performance Computing Facilitator and the University Library's Research Data Services (RDS) Team Leader and another organized and hosted by the University Library:

> We had the Scientific Computing Day [a mini conference showcasing faculty and student research in the area]. And it was great. I mean, you got to meet people in all kinds of other departments, kind of share methods with them. I found people doing in psychology, and people in the business school, doing things which are very similar and helpful to me, which I would not get to meet the other way. There was the library event where people presented their research, which was – aside from the ability to present my networks on the huge screen, which was just so pretty, like that was the best part – but also we get to see what other people are researching and you can now get to talk with people and see whether you can collaborate. So, I think jumping from these events and finding a way to have a more kind of like an established collaborative network of computational researchers in [GSU] would be amazing. Because, again, I get to meet the person from Andrew Young [School for Policy Studies], for example, I speak with them,

> and they do unsupervised machine learning analysis of protocols, of discussions in the Congress. And this is so fitting to what I do. And we don't know each other. We just met with third person who says, "Oh, I think you would love talking to each other because it seems like you're doing similar stuff." So if you can do something like that socially and that's all these events, and even if you can find a way to kind of facilitate more, especially physically a way of putting these people together, I think that, again, can benefit. [GSU7]

The researchers typically pointed to formal coursework or workshops as potential learning resources. That said, some noted that, during their own graduate studies, they had to go outside of their disciplines to find courses on the computational skills necessary for their work – and that this challenge continues for current students in their departments:

> So I'm mostly self-taught in these areas. In my PhD program, I had access to a couple of computer science courses that I found useful, and they allowed me to kind of fold it into a PhD program that would normally incorporate that material… So I would love to be able to provide like a real computer science education to every graduate student that comes through our program, even master's students… So like have kind of a structured set of courses to get them up and running in probably again in Python. But even just like basic programing, basics of working and manipulating with data, and then all the way up through like econometrics and statistics would be useful. Right now, everybody kind of gets a catch-as-catch-can. [GSU6]

> We're still at a point where we don't even require R classes yet. That's something that many students are sort of doing on their own, is taking classes in R or classes in Python. To sort of go, "You know what? These are the skills that I'm going to need moving forward and I need to do something." But we don't offer them. You know? They have to go over to [another GSU department] or something to get them. And so we haven't yet gotten to the point of saying, "OK, for those of you who are working on big data, here's really a nice structured approach to an introduction to cluster computing," and things like that. [We] just don't have it. [GSU8]

> I think in terms of training, we need more computational courses. I don't think we have enough, and I don't think there is enough variability in the amount of them. Especially in terms of social sciences, because I think taking a course in computer science is fine and nice, but not of your first computational course. You can't take NLP [natural language processing] in computer science courses. You're going to be lost, you're going to drown, and you're going to basically quit after two weeks. You need to have some good basis of coding in your own discipline before [you] verge out into this one and learn from them, because they're going to be much deeper into the math and kind of general gist of the method that you don't really need to know or even able to learn at that stage. [GSU7]

GSU1 similarly noted a paucity of relevant training available, but they couched it as resulting from their methodology being new/niche within their field, in tandem with abstracted training often not adequately preparing researchers for a specific tasks or applications:

> So, yeah, that's the other thing that I struggle with is training new students. There's not enough resources to do that, so. There are a lot of opportunities for learning how to program in general, so if you want to learn how to do Python or C++, etc., there are a lot of opportunities for *that*. But of course, then learning how to take that and combine it with what we do, that's a lot more specialized. And I think, yeah, we at the moment, it's not a big enough field that, you know, there are training events in this particular, well, they exist, but they're rare. So it's still not as organized. That's again, that's changing. As the field grows, there are more concerted efforts to make training videos and online courses for learning how to run calculations. But everyone has their own code. Everyone has their own methods. They train people for those particular codes and methods. But not… So, we develop our own code. For example, we have to train our students. I've been working on throughout iCollege [GSU's course management software], just making my own makeshift tutorials, sort of just putting together different things and putting videos online so that people can use them to learn those calculations. This is the first year I've tried to do that, it's gone… I mean, it's a lot of stumbling blocks, but I'm going to try it again next year and see what happens. [GSU1]

Similarly, while GSU2 felt that computational skills training was strong within their discipline/field, they lamented the lack of contextualized statistical and methodological training necessary to ground researchers' work in theory-driven aims:

> That that would be helpful if I think to get some exposure either I want to say formal statistical training, but that's incorrect because it will be dry and remote from actual application. But and yet when people are in actual applications, it's usually it ends up as a tutorial of how to use the tools, that's not what I mean either. It's like, I wish I, in my grad school or earlier, got more formal understanding of the models and the questions of the statistics. What, what is this? Another thing that's never taught and probably never gonna be taught, because everyone is talking about it, is the scientific method. The scientific method is never taught anywhere formally… Everyone is so proud that we as scientists achieve that. And that's created the sort of field or created scientific revolution and everything. And it's never taught – reasoning, rationality, experiments, planning, maybe taught in some kind of an academic dry manner. But the core of it is not, like, I never saw anywhere, at least that kind of exposition. Maybe I wasn't looking at and peering long enough into the curriculums. But I don't see that. That would help. [GSU2]

Paths toward Improved Support of GSU "Big Data" Researchers

To conclude this report, we offer some possible paths toward improving support of GSU "big data" researchers, grounding our suggestions in this study's findings and the team's broader knowledge of the strengths and limitations of the University Library and Georgia State University (GSU) as a whole.

We have learned that the scope, scale, and diversity of what constitutes "big data" research is extremely varied and specialized across fields. As such, it is not realistic to expect the University Library's Research Data Services Team to provide substantive tool and/or method support across all permutations of what constitutes this research area on our campus. Similarly, the size of data and computational methods used by these researchers is beyond the skills of most librarians, the skills typically taught in most library schools, and the data management/curation solutions that librarians often provide as part of data services. Moreover, computing and data storage infrastructure falls under the purview of campus IT support, and thus improvements to these are beyond the scope of standard library services. Given these realities, we offer the following as possible paths the University Library can explore for improving research support in this area at GSU:

- **Data Access:** Explore options to fund and manage campus-wide data access and licensing (e.g., APIs, proprietary datasets) to alleviate cost burden and promote cross-departmental access to data that is relevant to a critical mass of campus researchers.

- **Building Research Community:** Explore possibilities of creating cross-disciplinary networking opportunities for GSU campus researchers – e.g., social events, symposiums/mini conferences, meetup groups – to foster potential research collaborations, skills-building opportunities, etc.

- **Learning/Training Support:** Explore possibilities of the Research Data Services Team to collaborate with GSU "big data" researchers to offer discipline- or method-specific workshops to contextualize and de-abstract the content and create more task- or problem-driven sessions for their specific research project goals – keeping in mind current limitations related to skills capacity, sustainability, and scalability.

- **Advocacy:** Share our report with campus IT staff who oversee computing and data storage infrastructure to advocate for improvements to system accessibility and usability.

Appendices

Appendix A: Institutions in Ithaka S+R "Supporting Big Data Research" Study

The following higher-education institutions also participated in this Ithaka S+R research study.

Atlanta University Center
Boston University
Carnegie Mellon University
Case Western Reserve University
New York University
North Carolina A&T State University
North Carolina State University
Northeastern University
Pennsylvania State University
Temple University
Texas A&M University
UC Berkeley
UC San Diego
University of Colorado Boulder
University of Illinois Urbana-Champaign
University of Massachusetts, Amherst
University of Oklahoma
University of Rochester
University of Virginia
University of Wisconsin-Madison

Appendix B: Study Methodology in Detail

Participant Recruitment

Upon receiving study approval from the GSU University Research Services & Administration (URSA) Institutional Review Board,[6] the team commenced to recruit participants. Members of the team, associated with the University Library's Research Data Services Team and library administration, already had some familiarity with key researchers on campus and the types of research funding that the University had been generating over the last several years. The initial challenge was to identify those researchers working with "big data" specifically, as defined in the Ithaka S+R Project Scope and Recruiting document.[7] As instructed by Ithaka S+R, our team members identified three key stakeholders on campus: a Vice President and former Dean, a research computing technology administrator, and a current Associate Dean for Research. Formal conversations were had with the first two stakeholders, with informal conversations and email exchanges occurring with the third. As a result of these conversations and familiarity with research on campus, the team identified and

[6] IRB Number H21042, Principal Investigator (PI) Amanda Swygart-Hobaugh; for IRB-related inquiries, email irb@gsu.edu
[7] See footnote 4 (p.3) for project definition of "big data" as delineated by Ithaka S+R.

contacted 22 possible participants for the study, representing a wide range of disciplines in the sciences, social sciences, and health fields. Perhaps due to COVID-19 and the fact that many of these researchers had not been on campus or in their labs for several months, recruitment proved to be somewhat challenging, resulting in eight participants.

Data Collection – Semi-Structured Interviews

Data were collected via eight semi-structured interviews, conducted virtually via the WebEx platform and approximately 60 minutes in length per interview. The team used the below **interview guide** provided by Ithaka S+R, which contained prescribed interview questions to ensure continuity across the multiple institutions participating in the project. The research teams on the project were instructed to ask the top-level questions of all study participants and to use the bulleted questions as probing questions if necessary.

Supporting Big Data Research | Semi-Structured Interview Guide

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

Introduction

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

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

Working with Data

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

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

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

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

- What challenges do you face in finding data to use in your research?

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

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

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

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

Research Communication

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

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

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

- What factors influenced your decision to make/not to make your data or code available?
- Have you received help or support from anyone in preparing your data or code to be shared with others? Why or why not?
- What, if any, incentives exist in your department or 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?

For each interview, one team member conducted the interview, with other team members present to take notes, ask clarification questions, and corroborate interpretations during the analytical coding process if needed.

The recorded interviews were transcribed using the Trint automated transcription platform.[8] Team members then corrected errors resulting from automatic transcription and employed the following de-identifying practices:

- removed names and other co-occurring information spoken in the interview that could result in identifiable data
- applied the following anonymized interviewee identifiers in transcripts and associated metadata: GSU1, GSU2, GSU3, GSU4, GSU5, GSU6, GSU7, GSU8

Data Analysis – Qualitative Coding Process

The team employed a qualitative coding process, informed by the grounded theory method, to analyze the interview data.[9] The team's coding process proceeded along three stages: (1) developing a codebook, (2) analytical coding of interview transcripts, and (3) condensing coding insights.

1) Developing a codebook:

    a. Two team members (Jordan and Walker) both independently open coded two transcripts: one from the science interviews and one from the social science interviews. The initial open coding was done in Microsoft Word, highlighting relevant passages and commenting with themes/codes.

    b. Walker compared the open coding and created an Excel document that summarized and color-coded related themes/codes.

[8] Trint automated transcription service: trint.com

[9] For a high-level overview of the qualitative coding process in grounded theory method, see delvetool.com/blog/openaxialselective

c. Jordan and Walker then created a draft codebook organizing these themes/codes into a hierarchy of main codes and subcodes, including commentary explaining the reasoning behind chosen groupings.

d. Jordan and Walker met with Swygart-Hobaugh to discuss their open coding and condensing process and the resulting draft codebook. Jordan and Walker suspected that there were a few additional themes/codes not encapsulated in the first two transcripts, and recommended coding a third transcript to test the draft codebook and round out any remaining themes/codes.

e. Swygart-Hobaugh coded a third transcript using the draft codebook, and added new themes/codes to the final codebook[10] as needed.

2) Analytical coding of interview transcripts:

a. Swygart-Hobaugh imported all transcripts into an NVivo[11] project file, recreated the final codebook, then made three NVivo project file copies and distributed one copy file to each of the team coders (Jordan, Swygart-Hobaugh, Walker).

b. The three coders divided the coding responsibility, each independently coding 2-3 transcripts within their respective NVivo project file copy.[12] Each coder also documented insights gleaned from their coding in an NVivo 'memo' organized by codebook-informed themes/codes.

c. Once all coding was completed, Swygart-Hobaugh imported the copy NVivo project files into a master NVivo project file, resulting in one project file containing all interview transcripts, coding, and insight memos. Swygart-Hobaugh also conducted various NVivo queries and generated visualizations to explore the data and coding.[13]

3) Condensing coding insights:

a. Swygart-Hobaugh reviewed all the coding and insights memos, and then created a document that condensed the primary insights gained regarding the key research support needs and associated challenges faced by the interviewed GSU researchers.

b. The entire team (Jordan, Sinclair, Swygart-Hobaugh, Walker) met to discuss and corroborate the condensed coding insights and proceeded to organize and delegate writing responsibilities for report sections.

[10] See Appendix C: Final Codebook.

[11] NVivo is a qualitative analysis software tool developed by QSR International (qsrinternational.com/nvivo-qualitative-data-analysis-software/home).

[12] Coding responsibilities: Jordan – GSU4, GSU8; Swygart-Hobaugh – GSU1, GSU2, GSU6; Walker – GSU3, GSU5, GSU7

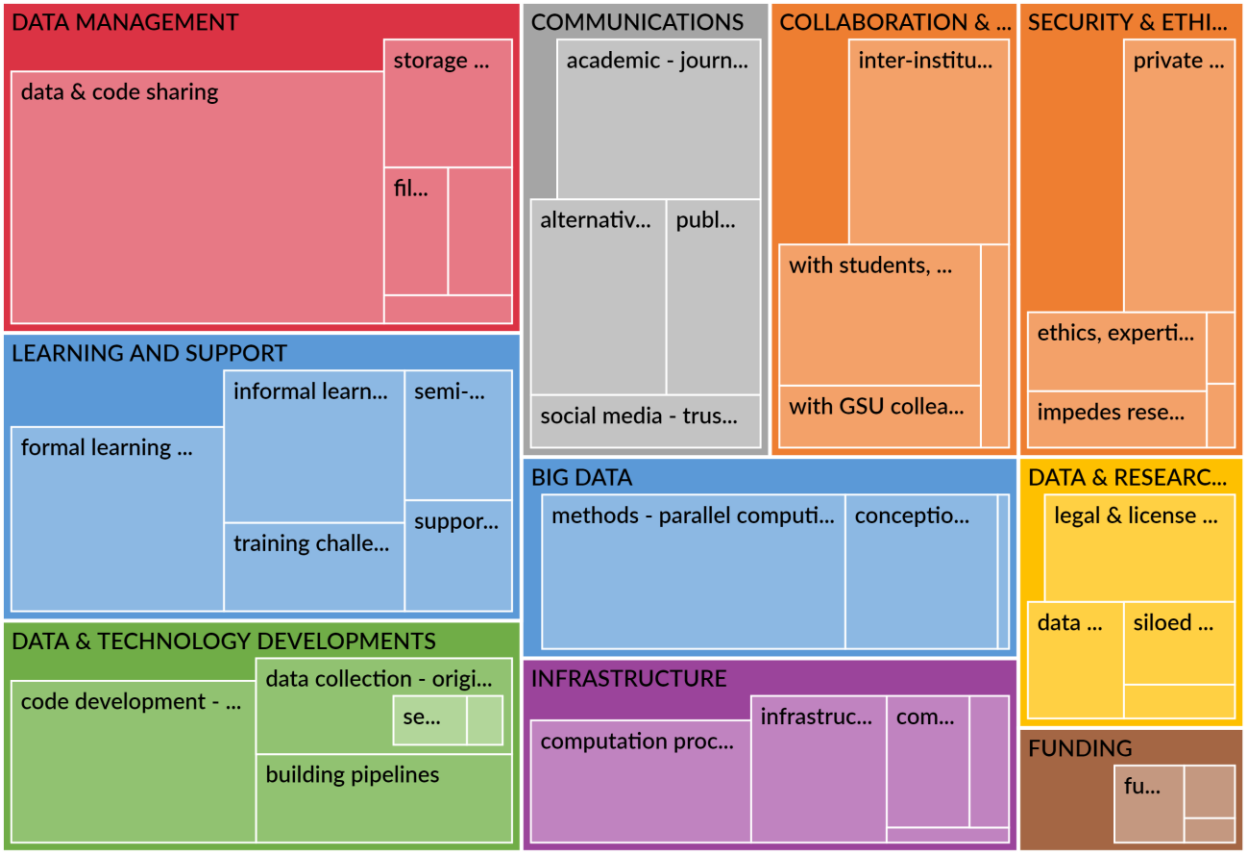
[13] See NVivo-generated charts, word clouds, and diagrams throughout the report, and Appendix D: Hierarchical Chart Visualization of Coding.

Appendix C: Final Codebook

The following are the final main and subcodes we used to analytically code the interview data using the NVivo qualitative research software tool.

- INFRASTRUCTURE
  - networking - data transfer over cable, challenges w/ network infrastructure not being able to handle (physical impediments)
  - computation power from processing view - institutional
  - computation power – students (e.g. underpowered laptops; access to institutional infrastructure)
  - specialized hardware
  - accessibility / usability - (e.g., HPC systems difficult to use; HPC responsive but lag in getting a fix)
- FUNDING
  - impediment to research
  - impediment to learning
  - Chronic - e.g. continuously having to apply more funding, reapplying for use on other computing, pay by the computational second
- SECURITY & ETHICS
  - impedes research
  - private data
  - ethics, expertise - e.g. others on my team with expertise to handle that
- DATA & TECHNOLOGY DEVELOPMENTS
  - data collection - original, secondary
  - code development - original
  - open source - utility, limitations
- BIG DATA
  - challenges - sharing, data volumes
  - methods - parallel computing, machine learning
  - conceptions - volume, variety
- COLLABORATION & TEACHING
  - involuntary - technical, logistical - GSU limitations kind of forcing them to collaborate with others, born out of necessity
  - Institutional - interinstitutional (non-GSU)
  - Colleagues - GSU colleagues
  - students, mentorship
- LEARNING AND SUPPORT
  - formal - coursework, workshops, summer camps
  - informal - webinars, online videos
  - semi-formal - networks, peer groups, clubs, meetups
  - support services - library, URSA
- DATA & RESEARCH ACCESS
  - legal & license restrictions
  - siloed data
  - data costs - APIs
  - institutional repository – lacking (or unawareness) -- in context of sharing code, or data, or pubs

- DATA MANAGEMENT
  - storage challenges
  - volumes
  - cloud / remote storage
  - data/code sharing - limitations, preferences
- COMMUNICATIONS
  - academic - journals, conferences
  - public - mainstream media, news, community
  - social media - trusted, untrusted
  - alternative - preprint servers, code repos

Appendix D: Hierarchical Chart Visualization of Coding

The following is a hierarchical tree map diagram, generated within and exported from NVivo qualitative research software tool, showing nested rectangles of code groupings (main/subcodes) sized by the number of text excerpts coded to the main and subcodes.