Big Data Research Practices and Needs at North Carolina State University: An Ithaka S+R Local Report

September 2021

North Carolina State University Libraries

Susan Ivey
Karen Ciccone
John Vickery

Contents

Background

In the Fall of 2020, the North Carolina State University Libraries conducted a study on the research practices of faculty and research staff in a variety of fields who utilize data science or big data methodologies in their research. The purpose of the study was to better understand researcher processes in working with big data in order to inform the development of resources and services at NC State University to support this work. This project was part of a larger study organized by Ithaka S+R, a not-for-profit research and consulting service.[1] Twenty other universities conducted parallel studies about big data research at their institutions during the same time period.[2]

This report provides contextual information about the landscape of big data research at NC State and describes the methodology and the findings of the study. It focuses on needs and challenges uncovered during the interviews, as well as predictions about evolving trends in the field, as described by the interviewees.

Study Methodology

The research team, which consisted of three librarians, conducted one-on-one semi-structured interviews between September 24, 2020 and December 2, 2020. Each interview was approximately sixty minutes in length. Due to the COVID-19 restrictions, all interviews were conducted remotely via the video teleconferencing application Zoom.[3] Digital audio recordings were produced with the consent of the participants and transcripts were de-identified. A detailed description of the study methodology is provided in Appendix A.

We asked campus administrators that are stakeholders in the area of “big data” to suggest potential participants who create and/or utilize big data in their research. Participants were selected purposefully to capture the breadth of big data and data science research at NC State. The research team invited twenty researchers; a total of ten researchers agreed to participate. Interviewees came from a variety of disciplines including marketing, materials science, digital humanities, plant sciences, bioinformatics, biological and agricultural engineering, and geospatial analytics. Two participants were Assistant Professors, three participants were Associate Professors, four participants were Professors, and one participant was a Postdoctoral Research Scholar.

---

[1] https://sr.ithaka.org/about/

[2] Atlanta University Center Consortium, Boston University, Carnegie Mellon University, Case Western Reserve University, Georgia State University, New York University, North Carolina A&T State University, Northeastern University, Pennsylvania State University-Main Campus, Temple University, Texas A&M University-College Station, University of California-Berkeley, University of California-San Diego, University of Colorado Boulder, University of Illinois at Urbana-Champaign, University of Massachusetts Amherst, University of Oklahoma, University of Rochester, University of Virginia, University of Wisconsin-Madison

[3] https://zoom.us/

Participants were asked about the following topics: their current research project(s); how they work with data; research communication; training and support needs; and perceived trends in their field. A copy of the interview questionnaire is provided in Appendix B.

The interview included a question about how researchers disseminate their findings and stay abreast of developments in their field. Because the responses to this question were not specific to big data research practices, we omitted them from the body of this report but have included these findings in Appendix C.

Limitations

This study took place during the fall of 2020, occurring simultaneously with the COVID-19 pandemic, which caused great professional and personal stress and shifts in priorities. Out of the twenty invited participants, ten agreed to participate, while eight invitees did not respond and two declined. It is possible that the participation rate was affected by the additional stresses and time pressures of teaching and doing research during a completely remote fall semester.

Big Data Research

Participating researchers use a variety of different approaches and encounter a corresponding variety of different challenges related to their specific data needs and methodologies. These differences highlighted the fact that “big data” means different things in different disciplines. It is unclear whether “big data” is a truly meaningful category for combining these various types of data-intensive research.

“Big data” research is often defined as research conducted with data that is high in the “three Vs” of volume, variety, and velocity. For the purposes of this study, we considered research to be “big data research” if it is high in at least two of these three Vs. Depending on the nature of their work, researchers grapple with different challenges related to these aspects of big data. The quotes below reflect some of the ways our interviewees work with and think about “big data” in their respective areas of inquiry.

> Marketing: *“Yes, so big data does play a big role for me, right? We are often looking on the order of millions, if not, in some cases, billions of tweets or social media utterances. … [E]ven a small amount of images can suddenly start to scale up quite a ways in terms of the amount of information encoded in the amount of data that's there. Yes, we're using a lot of large scale data. In fact, in the next project, we're moving on to video analysis.”*
>
> Materials Science: *“So there's … the four V's of big data—veracity, velocity, volume, and variety. And what we deal in is variety of data. So we have a whole bunch of different types of data for different types of materials. And the question is how to integrate all of it.”*
>
> Digital Humanities: *“Not everyone in [humanities fields] buys this anyway. There are a lot of suspicions that data modelling and analysis tips over into a kind of empiricist approach*

to literature and history that is simplistic and does not at all engage the kind of complications of contexts, or even cannot accommodate the perspectives of various critical theories that we use to talk about our stuff. So, that's a challenge."

Plant Sciences: *"[W]e're going to be dealing with lots of image data and also lots of other data that are going to be describing things like field characteristics, weather, some characteristics of the orders that were placed, how things were sorted, what type of value or pricing was associated with the orders that were filled, etc. And that in combination with the actual data from the [crop] is going to provide a rich amount of information for us to then try to do our relationships. … [T]here are going to be situations where we really are going to have no choice but to use more data-driven models because some of the mechanisms associated with the processes are not known."*

Bioinformatics: *"[C]onifers have very large genomes, anywhere from five to eight or ten times the size of a human genome. So there's gigabytes of sequence for a single individual. And … we're interested in analyzing hundreds to thousands of trees. ... [S]o it becomes a very large problem and a computationally very complex problem. … [I]t was a huge computational challenge to assemble the first human genome sequence in 2001. And that's a trivial problem compared to the problem that we have of working with a thousand individuals, each of which has a genome eight times larger than the human genome."*

Epidemiology: *"For us the definition [of big data] will be around integration of different sources of data in the same database and using the different sources to make some sort of decision at the end. ... Every week there is feed coming to the farm, there's a lot of measurements happening for the farms to achieve their performance. So, there's a large amount of data they collect. People are not connecting the dots yet."*

Geospatial Analytics: *"[O]ne of the interesting things is that depending on who you're talking to, big data means very, very different things. If you're talking to a remote sensing community, big data has terabytes and petabytes of data. If you're talking to people that are dealing with outbreak data, big data is maybe tens of thousands of records. So, it's not as cut and dry as I would have originally thought when dealing with these cross-disciplinary kinds of things in terms of what people think big data is."*

The researchers we spoke with use a wide variety of different methods in working with data, including text and data mining, mathematical and statistical modeling, and machine learning and computer vision approaches. A common need across disciplines is to integrate various data streams and data sources for analysis and to formulate inferences and models for forecasting and decision making. Several of our interviewees are building predictive models to better understand such disparate phenomena as phenotypic selection, microbial communities, pest and pathogen spread, and disease transmission. Not surprisingly, researchers using similar approaches face similar challenges in working with data.

Working with Big Data

While working with big data has unique challenges, many of the challenges mentioned by our interviewees are not specific to big data, but are encountered by anyone doing data-intensive research. Nonetheless, working with a high volume or variety of data can increase the difficulties associated with acquiring, sharing, organizing, and managing data. For example, it can be difficult to share large amounts of data with collaborators, and researchers still resort to transporting terabytes of data on hard drives in some situations. Common themes in the interviews were challenges related to accessing needed data, data quality and cleaning, integrating disparate data sources, organizing and managing large amounts of data from various sources, and data security and privacy.

Data access

Interviewees frequently mentioned challenges associated with accessing data owned by companies or commercial providers of data. Even when data is commercially available, there are sometimes limitations on data access imposed by commercial providers, whose primary customers are often not academic researchers. Researchers working with data produced by companies may also face the challenge of convincing CEOs to entrust them with their data.

> *“A lot of the material I work on really ought to be in the public domain. However, it has been digitized and monetized by commercial providers ... that then sell it to the library at exorbitant prices and prevent libraries from knowing what other libraries have bought it for and just do things in a way that's unsatisfying. I sympathize with the notion that this is an enclosure of the cultural commons, and that these things ought to be more accessible. … The problem ends up being that you have all sorts of materials scattered over different kinds of providers that have varying access, depending on what institution you're at, and all these things can make it difficult to get at things and also to collaborate with other people.”*

> *“The challenge with [using a commercial provider] is that we're now bound by whatever the limitations of that commercial provider are, right? So, we run into issues like, we can't collect more than 10,000 tweets from a single day, which most times isn’t an issue. But like, if it's election night ... For the predominant amount of data collection we do right now, we still have to supplement on a regular basis with our own tools and our own collections, just to kind of make sure we're getting a full picture of what's going on … [Q]uite honestly, a lot of those tools are built for firms that are interested in social media monitoring and not really research ... As tools they haven’t been really built for us.”*

> *“The challenge is to convince the CEO of the company that we're gonna protect the data and not gonna sell their data. That's the most challenging one, trust. … [In another country it] is much more government driven, here in the U.S. it is much more private sector driven. So, we need to go to every company to convince every company about it. It's more difficult.”*

Researchers mentioned other challenges related to data access. Sometimes the data is difficult to find or simply isn't available, or systems for finding the data are inadequate. It can be challenging accessing data from an agency that does not have sufficient guidelines for data archiving in place, making data discovery and reuse more difficult. The necessity of knowing how to code in order to collect or download the data can also be a problem.

> *"Especially with images, it's hard to sort of find the things that you might be trying to look for."*

> *"There are certain types of data that aren't required to be reported up through different data archiving systems, like with different repositories. For most environmental data, the systems are pretty good. But for example, I've been increasingly doing work in shellfish waters. So this is an unusual case where they're collecting environmental data, but the data are not under the jurisdiction of the EPA. It's under the FDA, and the FDA, because they don't do as much environmental monitoring, they don't really seem to have as much of a robust program for data archiving. It's more on a state level. So the states are doing an excellent job, but they're not cohesive in the way that they report their data up. So sometimes that kind of thing can be a problem."*

> *"Access can be an issue. Also, depending on the types of data we're working with, you often have to write quite a bit of code just to access the data. So that's pretty particularly true for big data. And for students who are just starting out, that turns out to be a pretty big hurdle. By the end of their programs, usually they're really adept, and they can download data pretty quickly. But it is kind of daunting at the start, when you tell a student, alright, you're going to have to download all these satellite images. So just to be able to access the data, you're first going to learn how to program in, like Google Earth engine or something."*

Data cleaning and quality

While not specific to big data, tasks like data cleaning, and problems related to data quality, were mentioned by several of the researchers we interviewed. Unsurprisingly, lack of standards, including formatting and metadata description, were mentioned as primary drivers exacerbating these problems. One researcher said that because there is not a standard approach for running experiments, they have to first spend time assessing the quality of the data. As a way to solve this problem, they are trying to create "an expected template of what we need." Another talked about how, even when using the same technique, different instruments and facilities will produce different formats.

> *"[O]ne of the biggest … tasks and efforts that's involved with big data science has nothing to do with actually storing the data or analyzing the data. It's the process in between where you clean the data. I feel like that's 90% of what actually goes on."*

> *"Poor databases from the companies. … [For] one company we find 18 different entries for a specific farm. So then, of course, we write some calls to recover that, but that's the*

most common thing. And the way they maintain their data, they contain the data in different tables, different places, wide format, long formats, you name it. It is all over the place. So, it takes some time to figure out what's going on and they don't know sometimes what's going on. So, that's the biggest challenge. Understanding who is the person who actually knows the system."

"In terms of formatting standards, if the data are not collected by us, researchers use their own way. Most of the time, it's an Excel spreadsheet with 10,000 tabs. So, most of the time is [spent] trying to understand because there's really not a good standard experiment. If not run properly with the idea of a modeling approach, that may not have good quality data. So, what we have to always do is first quality assess the data to remove the biological replicates and the technical replicates, etc. The standards are all over the place so now we're trying to provide more of an expected template of what we need."

"Even with the same technique, even the electron microscopy image you collect here, you may get in a different type of format, than that electron microscopy image you get at Oak Ridge National Lab. And the same with X-ray diffraction instruments. We had three X-ray diffractometers in our facility, and if you put the same sample in each of them, the data would come out in slightly different formats."

"Newspapers are really messy. [W]ith OCR, errors galore, and other kinds of limited metadata description of their contents. … [T]he data is also historically biased. And this is something that my field and many, many others are trying to reckon with, in terms of the institutionalization of certain exclusive and specifically white perspectives, based on what gets digitized and what's saleable. Clearly digital resources are wrapped up in a whole big issue of ... representation to the way access to certain resources tends to amplify the stories from those resources at the cost of consulting other things. So, that's an issue."

Data integration

The focus of work of several of the researchers we interviewed, across various disciplines, is developing pipelines and methods for integrating and analyzing disparate data sources in order to gain new insights and answers to questions. Integrating different types of data, including continuous output from various types of instruments, as well as data collected by different organizations and different researchers, is extremely challenging.

"[O]ne of the biggest difficulties ... is obviously the notion of data integration. I mean, you have different data sets, different experiments, in some cases different conditions, and being able to merge those types of data. It's just extremely challenging. With that particular project, they've identified outside vendors to be able to store not only the data but metadata. But one of the other challenges is that some of these datasets are also going to be housed and located on different clusters. … and then there's the centralized

repository for the project itself. So, you have all of these different potential locations and keeping track of who put what where is going to be challenging."

"[W]e have a whole bunch of different types of data for different types of materials. And the question is how to integrate all of it. There are very few data analysis approaches for materials structure determination, for example, that can pull all of these different types of data together in one single solution. For example, you could take a particle that you synthesized in your chemistry lab, you can put it in the microscope, and you get a microscope image of the atomic structure. And then you take it over to an X-ray instrument, and you get a scattering pattern, which tells you about atomic structure. How do you put these two pieces of data together? To tell you a single answer as to what the material is? There are very few tools available to do that right now."

"[Our project] tries to answer a methodological problem in data-driven analysis of historical newspapers, which is that they are effectively siloed because of their source history, collected by organizations associated with certain national initiatives, national libraries, this and that, all of which tend to use different kinds of schema and metadata standards. So, there's problems of institutional and technical interoperability that [we] want to experiment with about how to resolve it."

Data storage, organization, and management

A majority of researchers we spoke with mentioned using Google Drive[4] and GitHub for storing and sharing data and code, and they find that those tools generally work really well. However, storing, organizing, working with, and archiving large amounts of data, and different kinds of data, are major challenges that were mentioned by a majority of the researchers we interviewed.

"[O]ne problem that I feel like is solvable, but we haven't solved it yet, in my own group, is just having a standard way of collecting all that data and putting it in a place where we can all access it and kind of keeping good archival tabs on all of it over time … We just dump it in a folder and we leave it there, right? Now, in some cases, where we know people [would like to] have wide access to that data afterwards, we've actually created a GitHub repository for it and put it up there and LinkedIn, and the paper and all that. And so those datasets are much more easily indexable and findable. But it would be great if there was kind of a standard practice for just organizing these kinds of datasets."

"But one of the key challenges, obviously, is being able to not only acquire all of that data but obviously be able to put it in a place where we can readily access it. So, dealing with or formulating databases that would allow us to do that. This isn't something that we had done previously because, you know, we could kind of get by with just kind of having the information stored with the microscopy images, that actually was kind of hard. We would

[4] NC State provides users access to Google products through its purchased Google Workspace for Education edition.

have graduate students walking around with terabyte drives, trying to move that information from one place to the other, but we managed. With the data from [new project] that's not going to be possible."

"[We are not managing data] in the most effective and efficient way. A lot of it is … one-off solutions, and it's enough to be able to get us through the project. But in terms of the long term, even as far as integrating across projects, that's not simple. It's not simple."

"Organizational strategies, how we how we store the data, how we gain access to the subsets of it that we want at any particular time, and how we process it from DNA sequences into tables of values that can be calculated on or manipulated using statistical or mathematical approaches are all aspects of what I think of as data science. … Storage and organization probably are the key issues."

"I was not good about [organizing data for collaboration] when I started … [S]tarting earlier in 2020, this basically became kind of like a pandemic pet project. … Basically, what I found is that I really want to implement these kinds of practices. I think it's much easier to do if you start from the beginning. So one student I have who's pretty new still, she picked it all up immediately and ran with it. I think it's a combination of personality, but also that she never did it another way. But I do have, for example, two PhD students who have been here for a little while longer. I think I'm gonna have to force it on them more ... [I]t's still kind of hard to get buy in. You know, I've dealt with a lot of lab groups, in my time here, and other institutions. And so it's hard, it's hard to kind of change workflows in the middle when you're used to something. It's hard for every human to do that."

Data security and privacy

Security and privacy are major concerns for some researchers who work with data that contains names, locations, or other sensitive information. Researchers who work with corporate data must ensure companies that their data will be kept secure. While for most of the researchers we interviewed, password protected Google Drive folders suffice, some mentioned that they will be starting new projects involving more sensitive data, and that they will need to figure out new ways to securely work with their data.

"Now, the issue is that as we start to move more and more into [sensitive data] projects, ... that is going to start to draw upon stakeholder data. Then all of a sudden now there is going to be this inherent need for privacy and security, because we want to be able to provide some level of assurance or at least be able to articulate to our stakeholders that, you know, we have their information, and it's not going to be accessed accidentally or distributed accidentally, etc. So, it's still not necessarily a quote unquote requirement, but it is something that we want to be able to convey to our stakeholders. So, as such, we are going to need to start to think about at least if we're going to be creating these types

of environments for storing these data, how do we do this in a secure way such that the stakeholder security and privacy is adhered to?"

"So, we have the agreement between NCSU, the company, and if the project is with the state officials, we have a three-way agreement. And in the agreement, we say how the data is going to be stored and where, and [how] we're gonna have access to it. It is not encrypted, we do create a list of one to X of the number for the farms, so we can hide some of the information. But, since we work with the location of the farm it is a challenge to blind the location of the farm because I need to plot the farm and my ten people need to draw the polygons of the barns and everything. So, that bit to me, I cannot figure a way to protect that completely. So there is some risk.

"[T]here are five separate security walls that we have to go through, including a double authentication. So we're actually looking at whether it's necessary for us to use the five-layer security or whether we want to drop down to maybe just a double security level. But for all of our data within the project itself, you will be required to ask permission to access that data. … [T]hey can see the data but they can't access the data without permission."

"[W]hen we're forecasting our pest and pathogens spread, and we're doing it at fairly small spatial resolutions … we always have to be sure that we're scaling it up to a large enough spatial resolution, where you can't identify a single grower out of it. Because you don't want to end up having pushback from the grower and that industry. And this comes from the past with the National Agricultural Statistics data, if they have only one grower in a county, they just won't report that county's production at all. So, we tried to do similar things to that, to protect the privacy of individual growers when we go to publish. Now we'll share it with our APHIS counterparts, because they collected all the data and they know where it's at anyway, right? But just for public consumption, we have to scale it up."

Infrastructure

The researchers we interviewed use several types of research computing and data infrastructure solutions. They mentioned the following:

- High Performance Computing (HPC) cluster ("Henry2") provided by NC State's central Office of Information Technology (OIT)
- HPC systems located, owned, and maintained by other U.S. and international universities
- Researchers' own local HPC clusters and servers
- Fee-based, free, and/or community shared cloud computing resources, e.g., Google Earth Engine, Amazon Web Services (AWS), RStudio Cloud, XSEDE cyberinfrastructure resources

Which research computing and data infrastructure solutions researchers integrate into their workflows depends on a lot of factors. One interviewee highlighted that simply knowing what is

available to them, both locally and externally, and keeping abreast of new advances that should be integrated into current workflows, are driving factors in what they choose to use. Another interviewee indicated that their solutions had evolved as their research became collaborative beyond NC State, stating “[T]hey now put in my budget cloud space, HPC, SQL database because now you start from your lab, and then the labs around NC State, and then you go across multiple institutions and then you go across the pond and it is more international, so it's expanding fast.” Other factors that contributed to selecting computing and data infrastructure solutions include lack of experience or expertise, ease of control over software installation and system preferences, and goals of the project. Various challenges related to infrastructure, including paying for fee-based resources, locating solutions that scale appropriately to meet the required computing and storage needs, and having an appropriate level of staffing and support, also influenced decisions about which solutions researchers utilize (discussed in more detail in the “Challenges with Infrastructure” section below). All participants reported that they utilize multiple storage and computing infrastructure solutions.

> *“I would say for maybe 75 to 80% of our work, we're managing in our local machines … [W]e haven't used HPC yet, primarily just because I haven't had a student who is really adept at that. I do have one student now who's considering using it for some of the image-based analysis that we're doing. But actually, a lot of times, what we've done more commonly is actually purchased cloud services. … I would say that's probably the other 20 to 25% of the time—we're using some form of cloud services to do our analysis.”*

> *“The reason that I [maintain my own] rather than using [NC State’s] HPC is that I can install software, I can modify the computing environment, I can set things up the way that I want it and I have something like 50 terabytes of disk space that's immediately accessible to those compute nodes that I can use for storing the datasets. And so, the combination of large amounts of storage space and the flexibility to control the computing environment is worth what it costs in terms of the time and effort to maintain those systems.*”

> *“[We use a] full combination of computing suites. So, we have some fairly high performance desktop machines that are custom built for our lab that we use for kind of like, fairly large development stuff, and then we'll push it to the HPC. And we also do quite a bit on the cloud front in terms of just quick deployment stuff. So, the interactive interfaces are all deployed to the cloud, and are all able to be run from anywhere, anytime we want to access them. So, it runs the gamut. And it really depends on what the goal of the particular project is. And then I guess, Google Earth Engine is also a cloud-based platform. So, we do a lot of processing on Google Earth Engine ... because they have a lot of the compute available for researchers. And it's already connected to a lot of the data.”*

Some interviewees not currently utilizing HPC and cloud resources expressed interest in doing so, with one expressing regret for not beginning to integrate cloud resources sooner:

“[A]s of right now, we have not fully integrated cloud computing into the stuff that we've done. And in truth, if there are some things that I would have changed, especially over the last couple of years, that's one of the things that I would have changed. I was even talking to one of my former students. She just got a position at [university]. And she was asking me about what should you think about in terms of a lab? And I was like, look, you really should invest in some aspect of high performance computing and really start to understand and investigate how you can integrate cloud computing into what it is that you're doing. And the reason is because it really impedes our ability to scale the algorithms accordingly.”

Challenges with infrastructure

A major challenge facing many researchers is transferring data between or off of multiple machines, storage systems, and compute environments. Interviewees cited the lack of automation and networking issues, which cause time lags between data collection and accessibility for use. Several interviewees emphasized that they utilize hard drives to address this challenge, while one interviewee suggested that a specific tool (Oxford MinION) might help address this issue in their field. Multiple interviewees noted that they transfer data generated at core facilities using hard drives, because it is currently the quickest and easiest way to get access to the data. Interviewees expressed a mixture of frustration and disbelief about this ongoing challenge.

“Well, I don’t know if things have changed or not, but we were doing the same thing in [project name], mailing around hard drives, given bandwidth issues and the complications of permissions and not wanting to expose anything over the internet. So yeah, I think, ironically, data comes in the mail.”

“[A]s of now, [transferring data] may just take a day or two because we go around with our portable discs so that we know that we have that right away. Data that are in the core facility should be put somewhere so that we can access them remotely or not. So the problem is to access data that are in servers that are not available remotely. It drives me crazy when my postdoc says, ‘Well, I don't have the data because it's still on the computer of the confocal and I need to go there and I need to download it.’ And I'm like, I can't believe this is still happening these days but…”

“[T]here's no automation in the field, there is no really high-speed internet that we can quickly upload data on the cloud, [and] when you use hyperspectral images or other phenotypic processes, then the data are difficult to quickly move.”

“[N]eeding to ... basically get some hard drives and walk around literally, you know, with the images on their person became a challenge. And again, what we're finding is that particularly with [project name], our perspective is, well, let's see if we can identify other solutions to be able to get that done. And I know that there exist some solutions now. [Faculty member] was telling me about this software called Globus that the university

has access to where you can transfer large amounts of data. So, it's things like that are things that I honestly was not aware of."

"Data storage and processing speed typically ends up being a challenge. I've had two students now who have had pretty major challenges just associated with moving their data from their computers to HPC."

"Oxford MinION is a device that you plug into your computer. You can sequence any sample on the fly, and you can already export that to the cloud directly. I think that is something, if that becomes the state of the art of doing diagnostics, across the species, the ones that I work with, at least, then we may be in a good position on trying to eradicate or trying to control better diseases, because right now, by the time you get the diagnostic it has been a month, and it's too late. And I know there are some NSF funds going into that right now."

Challenges associated with infrastructure solutions for basic data storage were also discussed. Most interviewees indicated that they use Google Drive due to its collaborative nature and easy-to-use interface, highlighting that it makes sharing with other NC State researchers and external collaborators relatively easy. However, there were several challenges associated with Google Drive in terms of big data, including file, transfer, and download limits. One interviewee described working with microscopy images in Google Drive as "a beast", noting that transferring such large files was difficult because "you can only transfer up to a certain amount or move up to a certain amount of data and then also, even at that point, it starts to throttle you before it shuts you down." Another researcher highlighted the potential negative impact on securing grant funding of not having available and affordable solutions for working with large amounts of data.

"I think more and more a challenge to researchers in big data is just having access to data storage space that's affordable and having infrastructure on campus for large projects. For [research project], we had to look for data storage outside of the university because there just wasn't anything available within the university that could support our needs. And I think increasingly [that is] going to be an issue [and] I think that there's more awareness that it's an issue and there's more awareness that there are more and more researchers that are becoming involved in very large data generation projects, so it's something that does need to be addressed or else it's going to negatively impact our ability to do the work we want to do and win the awards we want to win in order to do the work we want to do."

Another challenge identified by interviewees was the lack of a powerful enough compute environment or a compute environment that satisfies their disciplinary needs at NC State. One of those interviewees highlighted the lack of required domain-specific software on NC State OIT's HPC (Henry2) and shortage of available support staff. Many interviewees' have solved this problem by building up their lab's computing infrastructure, as well as by utilizing supercomputers external to NC State, though one interviewee noted that external resources are not without their own challenges. Another interviewee suggested that utilizing or partnering with cloud-based platform providers, specifically citing Google Earth Engine, might help with scale.

“I had only three CPUs before, it was taking 30 days to finish them all. I don't have 30 days to run, so that's why I have now increased. I came from [university] and a supercomputer they had, I think they had about almost 1000 CPU. So, it was that easy to run big jobs, to just use 200 CPUs was not difficult over the weekends. Here, I know that there is an effort to increase Henry2 to make it bigger. But still, it's just too fragmented ... So, that's the biggest challenge, I would say. I have other sources. I have access to a supercomputer in [country] as well. But the lines are just, I need to plan like a month before I send a job and I … cannot plan jobs a month ahead, you know, I need to run something this weekend, I need to run a call, that's it, there’s just no way.”

“One thing that has been a factor in my decisions, for example, about maintaining my own computer hardware as opposed to using community resources has been that the centralized computing facilities at my institution are developed by engineers for engineers and bioinformatics and life sciences computing is an afterthought ... So one of the advantages of my collaboration with a colleague at [university] is that I have access to the computing cluster there. And that institution has two computer clusters, one that's built by engineers for engineers, and the other one is built by bioinformaticians strictly for the use of life scientists who were analyzing biological data. ... So if I have a problem that's too big for me to solve on my own hardware, then the first place that I go to to try and solve the problem is to [university] because I happen to have access to that cluster.”

“[D]ata storage and processing speed typically ends up being a challenge. ... [I]n one case, I have a student who's running a pretty simple model, but it's for all of the coastal contiguous United States. So it just takes a really long time. So, you know, it's kind of funny, because the model is not very complicated. It's pretty simple. But once you start applying it at those scales, it takes her, I think, several days to run her models. And I have another student who's working over a small spatial scale, just [town], but she's working with really high resolution data. And so because of that, her model also takes several hours to run. … It just adds up.”

A few interviewees mentioned financial challenges associated with data storage and computational infrastructure. A couple of those mentioned that they have written data storage and computational infrastructure costs into some of their grant budgets. However, another interviewee explained that their understanding was that computing costs were not allowed to be charged to grants. This could be because of a difference in funder policies or it could be indicative of the complex and sometimes confusing budget requirements imposed by NC State, which is an issue that the authors understand the university is currently working to address.

“The other aspect of cloud computing is it’s very powerful, but it requires monthly payments. And the problem is that I don't have a budget that allows me to make monthly payments for computing costs. Computing costs are supposed to be covered by the university, and so they're not something that can typically be charged to research grants. And the way the university distributes returned overhead, it comes, you know, somewhere typically in October or November, and it all has to be spent by April. And so I have no money for six months of the year. And then the rest of the year, six months or

so, I have some unpredictable amount of money which may or may not be enough to cover the costs of whatever I need to cover with it."

Challenges with infrastructure support

Several interviewees highlighted challenges associated with finding local support at the university for setting up new individual solutions or utilizing existing infrastructure solutions. One interviewee described the difficult process of working with one of the NC State laboratories to mount data and run on their computing cluster. They described it as *"*a really tremendous pain and caused us inordinate delays with the complexity of the back end, and these sort of server-side management systems, I mean... sysadmin level complexity." Another interviewee noted that they have had issues with response times for installing software packages on central OIT's HPC, noting that they knew they were busy but it was disappointing that there was a lack of staff to address issues more quickly. It is challenging to not have local support for planning and executing multiple projects with different collaborators internal and external to NC State, specifically related to determining the best storage and computing infrastructure on a project-by-project basis. One interviewee expressed that this is very common across research groups and research programs, resulting in "these little piecemeal type of solutions, you know, things that are kind of weird things that kind of work and things that kind of don't." This interviewee explained that their external partners, including private industry partners, have proven useful for educating them about how to streamline solutions that scale:

> *"[T]he unfortunate truth is, I know most of the things that we did and that we do, we pretty much do it internal as the lab. And the only help that we actually are receiving now are actually through our external partners. So, for example, with [project], [private industry] has been extremely helpful in terms of helping to convey a more structured solution to some of the things that we're doing, particularly for that project. And obviously the same comes across in terms of [private industry] and the overall environment that they're trying to put up. But now, unfortunately, we have not effectively utilized some of the services that I now know exist where honestly, I really did not know that they existed before.*

However, other interviewees mentioned that the assistance that they have received from both NC State's central OIT staff and NC State's Bioinformatics Consulting and Service Core staff has been extremely useful. One interviewee noted that the assistance received from HPC staff was instrumental in not only solving HPC issues, but also in connecting the interviewee and their students to other university support on campus. Another interviewee explained that they consult often with individuals in the biotechnology program for assistance filling their research teams' "knowledge gaps."

> *"[W]ithout [HPC staff member], my students wouldn't have been able to figure out HPC issues. And [HPC staff member] has helped us with so many things, like connecting us with other resources on campus. So there are definitely on-campus resources also that we've plugged into for support. It wasn't as much like formal training. It was more like*

how we don't know how to do this thing. And that's happened many times where I've gotten support from campus offices."

"So, typically we'll meet in our project groups, [and] we'll identify exactly what it is that we need to do and where the knowledge gaps are. And we will talk. So, we interact quite a bit with the folks at the biotechnology program that are responsible for putting together packages of software for bioinformatics pipelines. So we consult with those folks quite a bit."

Lastly, several interviewees described needs and the challenges associated with software application development, statistical analysis, and computer programming. Researchers often need help designing and creating the tools they need to do their research, but they have a difficult time finding collaborators with the needed expertise, or finding affordable assistance.

"[W]e've been working on a web application that requires way more computer programming types of experience than what my group had. We have quite a bit of experience with data analysis, but less on the decision support tooling side of things, like how do you actually produce decision support tools that are driven by the models that we produce. And it's an important step, because it's not like in computer science. They're doing computer science research, they're not creating computer science applications. So one thing that I think is often lost is that if I want to create a decision support tool that's Web based, that's interactive, that provides a lot of important functions and features, I can't go and collaborate with someone in computer science for that, because that's really not what they do. There's not a research question embedded there. And there are probably ways of embedding research questions. But I have sometimes felt like I either have to have like a massive project with many collaborators, like from computer science, that what we're doing is really interdisciplinary in terms of the research questions we're asking, in order to create those kinds of tools that help us basically put big data analysis to use. But there really is no support. ... That's basically, as far as I can tell, nonexistent. And this isn't unique to NC State. This is everywhere."

"[W]e have the capability to really understand the biology and to really understand the mathematics, [but] in order to really design the model on the data for some specific aspect like specific algorithm development, [that] is not our bread and butter. ... So, we rely, for example, on classes over the fall semester or spring semester, say from bioinformatics that have graduate students focus on a semester-long project. So, just recently, we just collaborated with two graduate students from a bioinformatics course to develop a graphical user interface in R, which we were well versed in MATLAB but not in R, so they help us with that. … [I]f a postdoc is not well versed in the statistical approach and analysis, then we often use the Statistical Consulting Service."

"We don't have a C+ person, but I'm looking at hiring a computer person to actually transform the functions on our end to C, because for the [research project] we have about, in one state 300,000 farms to model, about 2.2 billion movements a year. So, running that in R takes quite a lot. So, I was talking to the main campus to see if I could

have hours in somebody to actually help with transforming the functions in R into C. And apparently that is an option there. And I'm not ready for that, fund-wise, I don't have the funding to do that, quite expensive … I don't even know if I'm going to be able to do it, but it is in the proposal. So, I guess I need to figure out a way to do it."

Interviewees that struggled with finding this kind of support explained that hiring external firms came with expenses that were often prohibitive. One interviewee noted that while the high cost for hiring external firms is understandable, they can neither afford those prices nor do they need the amount of support those firms' fees include, such as a project manager and a graphic manager. Some solutions that were mentioned included learning to do the work themselves or hiring students to help. However, one interviewee cautioned that hiring students is not always a sustainable or effective solution:

"We also hired a computer science student. And that worked out well. But I know that it's only working out well 'cause the student is really awesome. And I think it would have been easy to hire someone that maybe just didn't get into the project and wouldn't have really helped us the way the student has. So I wish that there were more resources like that, where it's not just someone who could advise us on who to hire, but like someone who we could actually work with, who can help to do some of these things, almost like [staff member], but for like web applications or something."

Another interviewee discussed a similar challenge, but in terms of finding people that "are well versed in IT infrastructure". They emphasized that the ability and expertise necessary to identify and implement appropriate infrastructure solutions is very different from the skills and knowledge necessary for data analysis, noting that there is a perceived misconception that researchers are capable of doing both:

"So, one thing that ... I face these days, and you may know, is just these misconceptions that people that are well versed in data analysis should be also well versed in IT infrastructure. But to me, they come with two completely different expertise and knowledge. ... [T]he other aspect that we probably alluded [to] at the beginning is the infrastructure for all of these data analytics, and we touch[ed] on HPC, and the cloud, etc. But all of these come with their own understanding that I myself don't much understand and I don't understand the intensive work that has to be put in place to allow all of these data analysis etc. So, sometimes it comes like, we're all one thing but I always see this as separate. So, I was wondering, is there anybody that is approaching these as one same problem or two different problems? ... Are there people or are there capabilities around that bridge between IT infrastructure and data analytics here?"

Skills and Training

Interviewees named a wide variety of skills needed by researchers in order to do big data research in their fields:

- General data science skills and "computational thinking"

- Coding (e.g., R, Python, JavaScript)
- Machine learning
- Predictive modeling
- Statistics
- Data cleaning and processing
- Unix command line
- GitHub
- Database systems (e.g., MySQL)
- Research computing (e.g., HPC, cloud computing, Apache Spark)
- Data visualization
- Data management

Some interviewees emphasized the need for researchers to have a broad understanding of data science concepts, including everything from data cleaning and processing to statistics and computational methods. Knowing how to work with data and how to break problems down into steps that can be solved with computational tools are critical skills.

> *"I think there's some pre-concepts that need to be taught there about just what is a computer model versus a mathematical or econometric model … Just understanding what computer modeling is about and how to use it is probably the aspect I think would be most important."*

> *"I feel like [data cleaning is] 90% of what actually goes on. ... And it's an aspect that there is, to a large extent, zero training on in the modern world. ... we either tell them to go out and collect data and look at it or we tell them to analyze a perfectly clean set of data. And there's very little where we say go out, collect the data, clean it up and then analyze it from start to finish. And I think that as both students and research scientists, that's something we're not trained as to how to do and what the best practices are."*

> *"Well, one of the things that I firmly believe is that familiarity with the Unix command line and with the basic principles of what's called computational thinking, how to approach large problems and break them down into individual steps that can be solved with individual software tools is a skill that all scientists should have, whether they're life scientists or any other kind of scientists. And so, that's one of the things that I focus on in the data analysis class that I teach in the biotechnology program is providing students with a foundation in those sort of basic skills of working with large datasets. Using tools that are more capable of handling gigabyte-sized files than the usual office productivity suite packages will tolerate."*

> *"Every researcher should be given a four-year sabbatical, so they can go back and study details of statistics, data science, big data, mathematics, and computer science and all the associated disciplines. … I think what is needed to get up to speed, all of the disciplinary experts, right, on the broad picture of big data, is to have high-level introductions. Like the 10,000-foot view … It's very important to find these sort of big picture introductions into the right terminology and the right categorization of how the*

field is organized, so you know how to frame questions. Otherwise, you're just paralysed by all of these different resources and not knowing exactly what question to ask or where to go. So, the best conversations I've had are with people that think the broadest about the field of data science. They can really say, okay, because you asked that question, you know, think about drilling down to this keyword here and searching just for that. It's difficult to see that from the outside."

The importance of coding skills was mentioned by a majority of interviewees. Due to the critical need for graduate students to develop these skills immediately in order to be able to do research, some researchers have begun teaching their students how to code.

"I teach an R programming course. And now most of my students, when they start out ... that's becoming basically like part of their onboarding, because it's so critical to our lab to have those skills. And they need to get it immediately. So I teach that class in the fall, and then they can take that course and over the span of the semester, start to gain some of those foundational skills."

"[Most importantly] they need to learn coding ... Maybe they are in high school, if it could go back to high school or you know just having to learn how to write code efficiently, not just something that works. … I cannot hire anybody that doesn't know how to code in my lab. So, it's very difficult for me to hire people, because I don't have time to teach them right from scratch."

Multiple interviewees said that machine learning is an area where training and support are needed:

"I think [machine learning] is fast becoming an area of interest and application in a lot of different disciplines and percolating into the humanities. But it invites a complexity that I think is really beyond even people who are pretty handy with a computer … it's definitely where people will need expert support."

"[It would be useful to have a] TensorFlow workshop or something along those lines. So, PyTorch, basically things on how to handle things like the big computational machine learning algorithms, and what are best practices on those would probably be really helpful."

While data science and coding skills are critical for working with big data, not all researchers view it as important to learn these skills themselves. Instead, they stressed the importance of domain experts working with data science people:

"We've worked with [colleagues in the Departments of Statistics and Mathematics] for now about six years on developing new algorithms and statistical approaches … We need to have application people realizing there's an opportunity and trying to engage people that understand data science solutions, but we need the data science people

also to say, oh, here's what we do … and these are examples of how we put this into practice. So providing some ideas to the application people. I think that the driver for this needs to come from both directions."

"[I]t can be not only overwhelming, but almost in some cases apprehensive in terms of admitting, you know, the deficits associated with our overall knowledge of how best to do some of these things. And I think at times we don't understand that there are experts that spend their life working on this."

Most interviewees had little or no formal training in how to do big data research. While some mentioned courses taken in graduate school, most said they gained their skills on the job and through informal channels such as colleagues, community members, and online courses and materials. Most Interviewees mentioned workshops and online courses as means for gaining new technical skills. Others mentioned learning from seminars, communities of practice (e.g., RStudio community, Stack overflow), podcasts, articles, Software and Data Carpentry materials, and instructional videos on YouTube.

"I did not receive specific training to work on big data. However ... even if it was not a specific training, it was part of my postdoctoral training career to learn how to deal with big data."

"[M]ost of what I picked up, I picked up through doing it and through working with students. Not formally, you know, availed myself of training from the university. There have been times when I've looked up information available on the Web."

"[O]n Twitter, I follow a lot of people who work for RStudio and things like this, and they all the time post about open science. There are a few people in my world of environmental and agricultural data analysis, who post about it, but it's mostly the people who are working in R that I follow. … There were certain, basically, like computational skills, where I sought out training materials online, primarily. So, for example, I learned how to code in Unix through a Lynda course."

When asked where they refer graduate students who need to develop data-related skills, some interviewees said they turn first to "someone in my group who was an expert on the subject" or "somebody that I know, that knows how to do it." In addition, university courses, online courses (e.g., Codecademy Python, Coursera), and library workshops were mentioned. However, finding ways to help students develop needed skills was viewed as a challenge.

"I have no idea whatsoever [where to point students]. And indeed, there's not a real, say, center or environment that is around us. I would tell them to learn by experience or hands-on data and be mentored by a senior postdoc or a senior graduate student that has been trained. That would be a short-term solution. Second, to really be brought up to speed there are some courses that NC State offers, some Coursera courses, some workshops, the class there for systems and synthetic biology, but it's not really a center,

so there are not that many people that are just in the same environment where you can have strategic discussion or a common training, so I would not know where to send them. Other than send them or establish collaboration or send them to classes or something like that."

While library workshops were mentioned by several interviewees, and recommended by some, they do not often meet researchers' needs for ongoing and more advanced training. Library workshops are only offered periodically, and they are generally aimed at a beginner audience. In addition, some researchers might not be aware of the Libraries' offerings.

"[M]aybe because I work with [tool and librarian] I may be more aware of those than my colleagues and my faculty are, or my co-workers are, but I think a lot of them aren't aware that those tools exist, right?"

"I've benefited a lot from the workshops that NC State University librarians have been very generous to offer to these classes. And tried to learn alongside the students about some of these basic approaches to cleaning and visualizing relatively simple and small datasets, but still like CSV files that you can handle on your computer and run on personal software."

"I have pointed students to the Libraries when it comes to R programming and I think a course on Python. I know that I've had students who have attended the workshops through the Libraries, and they have great things to say about the workshops. I think maybe a limitation of the workshops is that they are more bite-sized and happen over your schedule, which is totally understandable. I think for the Libraries to really meet their needs, it would have to be like a lot—like nonstop."

"In the library here, I know that there are these walk-in hours for R and Python. And there are some workshops that I know the librarians do. I went to one of them to get the feeling. But, I got the feeling that people are just starting. So I cannot commit my time to these because they're just starting from scratch. I had one of my postdocs reach out to somebody. And he said he felt like I already know all that, so it's not what I need. I need more advanced than that. So, and I understand that, most of the people are in the beginning so you need to help them get started. Once you pass the beginning, that is more difficult. To find the advanced thing like this is very, very hard. I don't know if there is not an audience for that. But it's very hard to find. Like if I want to run my own MCMC sampler from scratch myself, you know I can do that, but to find help to teach someone to do that, I really cannot."

Researchers face a number of challenges in taking advantage of training opportunities, most especially limited time and competing priorities. Established researchers have different needs and training priorities than graduate students and early career researchers.

“So in theory, training is great and there were times when I definitely would like to avail myself of it. My biggest challenge is available time, really.”

“[T]he number one factor is that it's kind of interesting. [A] small factor is, do I need to understand this to do the research I want to do? And if that's the case, then yes, I will often pursue it. I think even more importantly, I'm personally someone who, before I teach something, likes to see how someone else teaches it. So if I am going to be teaching that particular thing, right, I like to see the way someone else is teaching it first. And so going to training is often motivated by the fact that I actually need to teach that subject.”

“Most of the training I've attended since I've been a faculty member, it hasn't been as focused on how to do research. It's more so on other things like grant writing and whatnot. And that's probably just a reflection of where I'm at in my career stage. So if I get tenure, then at that point I might end up pursuing more of those types of workshops and training opportunities.”

Sharing Data and Code

In addition to comments about disseminating research findings (see Appendix C), many of our interviewees spoke specifically about their experiences with and practices for sharing their data and code. In general, the researchers we spoke with were willing to share their data and code when it might be useful to others. While not many mentioned sharing and publishing data and code through repositories and databases, many did mention sharing code through GitHub.

“I mean, we'll share our data. We'll share information with anybody that wants it. Especially once it's been published, I don't think that there's any particular use in terms of holding onto it, at least from our perspective.”

“For the most part, there are other people who work in the same general area. And so it's common practice for people to deposit high-throughput DNA sequencing results in the National Center for Biotechnology Information.”

One of our interviewees mentioned that sharing what might be considered as “bad data” is becoming more common within their scientific community.

”There's a push in our scientific community to more broadly share data, even bad data. And I got a call yesterday from a collaborator at [university] who's writing a proposal to create a software platform whereby people can share all of their bad data as well, for the purposes of going through with machine learning algorithms to look at all of these different types of data.”

Related to the push to share data, some interviewees mentioned that sharing data and code is a requirement for publication by some journals. For the Interviewees who mentioned sharing their code, GitHub is the most frequently used platform.

> *"For most of the journals that we publish in, one of the requirements for publication is that we make our datasets publicly available. So we do that. And as much as possible, we archive our data sets in international and national databases that are accessible by anyone."*

> *"Definitely, we do our best to make the code available. ... I used to have a website that had a lot of code and data available on that website. And then we would also push it to GitHub when it was appropriate, as well."*

> *"[W]e have created two pretty great software packages for analysis of X-ray diffraction patterns. And both of those are open source and freely available online. I can't remember where the first one is located. But the second one is on GitHub."*

> *"When possible, which is largely restricted by the commercial providers with whom we have specific licenses and can't open up that research data to share … we have shared the code [and] the scripts that that we've used on GitHub, and [journal] actually requires that you submit these things for peer review and publication along with the text descriptions."*

> *"[T]ypically ... the data has to be submitted with the paper. And also, typically, the code needs to be submitted with the people or at least placed in a GitHub repository where people have access to the code."*

In other instances, interviewees said that they do not share their data or code. Reasons for not sharing data or code included lack of perceived or actual requirements, not having created any new data that is not already publicly available, and collaborators' preferences.

> *"[U]nless required by the journal, don't make our data freely and spontaneously available online. According to our data management plans with NSF, we are required to provide data when asked, so if anyone asks us for data, we'll go find it and give it to them. But we do not go through the process of putting it out there."*

> *"[S]ince we've been primarily working with public data, we haven't been really publishing any data, not dropping data in repositories or anything. But I don't always know if I think that's a good approach. Like, I do think that we might, that we probably should start dropping data into probably GitHub repositories, 'cause I don't think it would make sense to drop data into a formal data repository, like USDA Ag Data Commons or something, since we're not the owners of the data."*

“Sometimes we cannot [share] the data, that's the problem. We have confidentiality agreements. I cannot put the farm's data on it.”

“[O]ftentimes [we can’t share] the data, because of the practical side of our collaboration [is] that they prefer for those things not to be shared, unless they're kind of scaled up to a certain level.”

“Twitter has a restriction that basically says you can't share with anyone who's not in the lab that collected the data, and now the solution around that is you are allowed to share what they call dehydrated tweets, which are essentially a list of the object numbers for the tweets that you analyzed. And so we will publish that list of numbers. And we will provide code so people could use that list of numbers to go and retrieve the tweets themselves.”

“And, there's a concern on making [code] available. Like, if you make a package or something, because users are going to start sending you an email asking for help with bugs or whatever. ... And, you know, we don't have time to reply to all the emails. So, we hesitate sometimes to [share], because you need to reply to emails of the issues of the user if you're making software.”

Trends and Future Needs

In addition to providing insights into current needs and challenges, several interviewees discussed evolving trends that would be useful for NC State to consider in planning for future needs around data science and big data research.

Two interviewees talked about the important role that research utilizing big data and data science methods plays in convergence research. They perceive a growing need to continue to focus on interdisciplinary connections and workforce development for working and communicating across the disciplines:

“[D]ata science is potentially a really good convergence research project or approach. ... Problems that require the integration of multiple disciplines, diverse data, big data, it's a great tool to do that. That can be driven from the application perspective, like from my perspective, or that could be driven from the data science perspective. And I think we need to do both.”

“I think that more convergence research and workforce development will be the way to go. I think that these data, even if you are in a specific discipline and you have a core knowledge of some engineering aspect or plant aspect or etc., I think that knowing the importance of using engineering mathematical approaches to really investigate the data comes with the knowledge that you need to communicate with the experts in the plant business. At the same time, it is really trying to understand what are the economical and societal impacts, and the big vision, of why you will want to learn how to really

investigate these data to have what kind of impact. So, workforce development, more interdisciplinary rather than disciplinary, with the clear core knowledge, but other aspects that are augmented, and supplementing the knowledge of students so that they can really seamlessly communicate across disciplines."

One interviewee noted that they see data science extension as a "glaring gap" where land-grant universities, like NC State, should step in and lead. Imagining dedicated roles specialized in data science extension, the interviewee highlighted the potentially "huge" impact that these roles could have in working with end-users that depend on extension, specifically citing agricultural producers and government offices. They further suggest that these data science "extension associates" should be proactive in reaching out to their end-users, rather than "putting the onus on them to ask for [help]," and that they expect stakeholders to want greater data science extension support soon.

"I think it would be really fantastic if land-grant universities could take a lead in having more data science, educational materials, but specifically geared for some of these end-users that do depend on extension. So agricultural producers and government offices. And not putting the onus on them to ask for it, but then being there to say, hey, we have these ways of doing things. And I think it would work for people who are dedicated in extension to doing that. I don't think it could be like just one small sliver of someone's appointment. So even if it's maybe not faculty, even like Extension Associates who specialize in data science or something, I think that the benefit could be huge. [I]t feels like it's been an underexplored space, but I could be wrong about that, I don't know the full landscape. But just since I arrived here, I thought I'd mention that. I would love to see data science extension as being an initiative that NC State takes the lead on."

Lastly, one interviewee explained that there is a big opportunity to consider how to curate the massive amounts of data that are being generated at shared core facilities, so that these data can be reused by others. Currently, this data is generated for a user and then leaves the facility, meaning future users will not know that it was generated or potentially that it exists:

"At the level of a shared facility, like the materials characterization facilities that you typically see on research active campuses, the users record a lot of data, they record a lot of electron microscopy images, they record a lot of chemical spectrum, they record a lot of X-ray diffraction patterns. And there's always a question as to whether or not the facility should support the curation of that data in some way so we can derive value from it, or if it's simply the user coming in getting the data and going away, but there's a huge opportunity if somehow we can take advantage of the fact that all of these people are collecting this large amounts of data. You know, what can we learn from materials data? That's the important question. From the facility perspective, whether you're talking about a synchrotron facility where 5000 experimenters come every year, or whether you're talking about a university-level materials characterization facility where maybe 400 users come every year, you know, there's a lot of data being generated. And there's a big

opportunity to take advantage of that data. … [I]t's very rare to find where someone is pulling all of this data together from a bunch of different research groups. So that's big data lost, maybe that is what it should be called."

Conclusion

Researchers in every discipline are applying a growing number of computationally and data-intensive approaches in their research workflows. These new approaches mean that researchers require advanced support that often spans across campus units. Universities with high research activity, such as NC State, must continuously reassess needs and challenges to keep up with the growing pace and demands.

The following are key takeaways from this study. We include suggestions for potential ways to address certain challenges that were uncovered.

- Researchers face challenges related to accessing needed data, data quality and cleaning, integrating disparate data sources, data security and privacy, and organizing and managing large amounts of data from various sources. They are using a lot of "one-off" solutions for data storage and management and are feeling anxiety about how they will handle this increasingly challenging issue.
- Researchers are developing pipelines and methods for integrating and analyzing disparate data sources in order to gain new insights and answers to questions. Integrating different types of data, including continuous output from various types of instruments, as well as data collected by different organizations and different researchers, is extremely challenging.
- Researchers need help finding collaborators with different areas of expertise (e.g., statistical analysis, algorithm development, software application development, coding). They also requested additional staff support and domain-specific support for NC State OIT's HPC (Henry2), as well as assistance identifying and implementing additional cyberinfrastructure solutions. It would be beneficial to assess areas where additional technical resources are necessary.
- There is a need for advanced coding and data science training for graduate students. It could be beneficial for the Libraries, in collaboration with other university units that support data science, to coordinate efforts to determine how to best fulfill this need.
- Researchers lack awareness of currently available computational and data support, tools, and infrastructure. Additional staff resources to help increase outreach efforts, boost coordination among research computing and data service providers, and help facilitate researchers in leveraging computational and data resources to accelerate their research efforts, would be beneficial.
- Clearer guidance about how to use grant funds to pay for research computing and data infrastructure, and if grant funds can be used for these costs, would be beneficial to researchers. Providing language that could be incorporated into proposal budgets could be a relatively easy way to help researchers navigate this challenge.

- Of the interviewees that stated that they share their research data and code, most only cited journal mandates as the reason they share, and very few mentioned sharing through a data or code repository. More education and training about the benefits of sharing research data, and particularly about the importance of good data management and stewardship to ensure data are more easily reused (e.g. how to follow the FAIR data principles), could have a positive impact on the data sharing habits of NC State researchers.

We should note that there are multiple current efforts on campus to provide additional resources for supporting big data and data-intensive research at NC State, including the recently launched Data Science Academy and a collaborative effort led by the Libraries, the Office of Information Technology (OIT), and the Office of Research and Innovation (ORI) that aims to address awareness of and access to available cyberinfrastructure/research computing and data (CI/RCD) services and support. This developing program aims to strategically improve and grow these services based on continuous feedback from stakeholders and in collaboration with CI/RCD service providing units. People from various groups across campus are involved in efforts to address several other challenges uncovered in our interviews, such as increased support and training for research with security and compliance requirements, reducing the difficulty of providing external collaborators with access to NC State cyberinfrastructure systems, and simplifying ways to utilize grant funding for infrastructure solutions and support.

Studies such as this one provide a glimpse into the current landscape. It is the authors' hope that the needs and challenges uncovered provide qualitative data that will help inform researcher-supporting units across campus in their efforts to continue to develop and grow services and support for the advancing needs of big data and data science approaches at NC State.

Appendix A: Detailed Study Methodology

The research team included three librarians who conducted one-on-one semi-structured interviews using questions provided by Ithaka S+R for participating institutions. The Semi-structured Interview Guide can be found in Appendix B. The interviews were approximately sixty minutes in length and were conducted remotely via the video teleconferencing software program Zoom[5], adhering to the NC State University guidance on in-person data collection at the time of the interviews and due to COVID-19 restrictions. Participants were asked about the following topics: their current research project(s); how they work with data; research communication; training and support needs; and predicted future trends in their field.

Participants were chosen from disciplines that create and/or utilize "big data." They were selected purposefully in order to capture the breadth of data science research at NC State University. Participants were chosen from a pool of names suggested by various campus stakeholders in the area of big data research, and they were approved by the NC State University Libraries' Head of Library Impact Analysis. This study only recruited tenured and tenure-track faculty, postdoctoral scholars, and staff researchers.

Participants received an email inviting them to participate in the study along with an informed consent form. The research team invited twenty researchers; a total of ten researchers agreed to participate. Interviewees came from several disciplines including marketing, materials science, digital humanities, plant sciences, bioinformatics, biological and agricultural engineering, and geospatial analytics. Two participants were Assistant Professors, three participants were Associate Professors, four participants were Professors, and one participant was a Postdoctoral Research Scholar.

Digital audio recordings were transcribed and anonymized. The anonymized transcripts were manually analyzed and coded by each team member. The team then identified themes in the data and divided the codes for further analysis. In agreement with Ithaka S+R's requirements, de-identified transcriptions were shared with Ithaka S+R staff on the project. De-identified transcription files will be preserved indefinitely by Ithaka S+R and the investigators so that they may be used in future research.

[5] https://zoom.us/

Appendix B: Semi-Structured Interview Guide

Introduction

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

- Give me a brief overview of the role that "big data" or data science methods play in your research.
- Is the use of "big data" common in your discipline?

Working with Data

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

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

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

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

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

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

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

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

Research Communication

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

- Do you keep abreast of technological developments outside academia in order to inform your research? If so, how?
- Do you communicate your research findings to audiences outside academia? If so, how?

- What challenges do you face in disseminating your research and keeping up with your field?

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

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

Training and Support

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

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

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

Wrapping Up

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

Appendix C: Research Dissemination Practices

Interviewees named several venues and strategies for disseminating their research to both an academic and public audience. In many respects, these strategies do not appear unique to big data research and seem to closely mirror the broader academic disciplines of each interviewee.

The typical venues for dissemination include the following:

- Peer-reviewed journals
- Conferences
- Preprint servers (e.g., bioRxiv)
- Social media
- University newsletters/announcements
- Websites (e.g., lab or research team public websites)

Unsurprisingly, every interviewee stated that they use published peer-reviewed articles as a primary means of disseminating their work. In addition to peer-reviewed articles, most interviewees also indicated that they use conference presentations and preprint servers (e.g., bioRxiv) to disseminate their research findings. Additional strategies employed to disseminate research include the following:

- Self-manage social media presence
- Coordinate with college and university communications
- Lab or project website

Social media

Beyond the more traditional methods of publishing, interviewees regularly mentioned the increased use of social media to disseminate their research. Twitter was mentioned as a method of having their research reach a broader audience both in the academy and general public.

> *“We tweet everything that we do. Either a blog post, interviews or talk or anything, we tweet that.”*
>
> *“Social media, yes, Twitter, mostly.”*
>
> *“And we usually also post through Twitter, when we publish it.”*
>
> *“I have a decent Twitter following myself now. So that is one platform by which I reach a couple thousand followers there. So they tend to pick up on my research when it's communicated there.”*

While useful for some interviewees, social media and Twitter were not always mentioned in a positive light. For two of our interviewees, they were considered challenging, confusing, or just not worth the effort.

“So, unfortunately, I am not the most adept to Twitter or other social media platforms. I swear the stuff is just so confusing to me. … And I know some of my colleagues there are so adept at Twitter and it freaks me out. I was at this one conference and somebody literally live tweeted something that I said. I was like, oh crap, really? And then when I heard it from another colleague of mine, I was like, oh, my gosh.”

“I don't do social media.”

College and university communications

Some interviewees coordinate with college or university public relations in order to help draw attention to and disseminate their research. While helpful, this takes time and requires initiative on the part of the researcher.

“We've also engaged with people like [staff member] ... he's a science writer basically. So he would write up a story on what has been done and then post it to different areas of places, not only the NCSU website, but other science-based websites and venues to advertise the release or the publication of the paper. Similar to what [staff member] does for the [university initiative].”

“And I work with [staff members] at [college]. When we get the grant that fulfils the mission of [college], they attempt to do a post on it. So, we work with them, provide the information they need, they might call, they might read the paper and pull the paper up. So, we do that. And my dean of research [staff member], she's very good. She asks for all the papers that are approved to go through the communication office to see if they flag that they might have a big impact. Then she wants that to be posted at the [college] web page and perhaps sometimes goes to the main campus. So, we do everything that they ask in that regard, you know, as much as you can get out there, the better. But, we don't have a communication person, like my graphic skills are something that is lacking in my lab. I had one postdoc, was very good in graphics, [but] he's leaving.”

“[W]e have [staff member], who's our science communicator at the Center, and she does a lot of the public communication with kind of less of the technical science angle and more of like, this is how you would talk about it for someone that doesn't have the technical expertise in your field.”

“I think the biggest issue, I would say that obviously, my own social media, it's not much of a challenge, but working with communications at the university and college level. I have to actively push on that. Like, I have to send an email to someone and say, hey, I would like to write a piece about this and set that up. And a lot of times it doesn't happen, just because I get overwhelmed, and I forget to do it (LAUGHS). So I think the biggest challenge is that there isn't like a pull on me. I have to push out that content, right?”

Lab or project websites

The use of a lab or project website was also mentioned as a method for disseminating research findings beyond the traditional peer reviewed journal articles and preprint servers.

> *“We also maintain a website for our project that provides news highlights about our findings.”*
>
> *“So, we have a website for the lab that I need to revamp a bit. But every paper we do a science page, which is basically a one-page summary, written to the producers, to the [industry] audience.”*
>
> *“So, we have a website and that's about it.”*