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

Jennifer Huck
University of Virginia Library, jah2ax@virginia.edu

Jacalyn Huband
University of Virginia Research Computing, jmh5ad@virginia.edu

Introduction

This report is an investigation of research practices of Big Data and Data Science Researchers at the University of Virginia (UVA). The study was conducted by a team of staff from the UVA Academic Library and UVA Research Computing (RC). Ithaka S+R, a not-for-profit research organization, oversaw this study, along with comparable studies at 20 other academic institutions. The goal of Ithaka S+R was to understand the resources and services that researchers who use Big Data and Data Science practices need to be successful in their work. Ithaka S+R set the overall research agenda, created the semi-structured interview guide, and provided training for local researchers.

The 20 other local institutions includes 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, North Carolina State University at Raleigh, 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, and University of Wisconsin-Madison (Ithaka 2020). Each of these institutions also conducted local studies and compiled their own results and recommendations. Ithaka S+R will combine this report with the other local reports to generate a more comprehensive overview of the services and resources needed by the researchers who handle Big Data or use Data Science practices.

The University of Virginia Library is home to the Research Data Services team. This team provides a variety of services related to research data, including data discovery, statistical consulting, data management planning, and research software support. This team works closely with the Scholarly Communication team which manages the library's repository services, including our data repository, LibraData. Research Computing at the University of Virginia is a separate team outside of the library that administers advanced research computing platforms, such as the high-performance computing system. Research Data Services at UVA Library and Research Computing jointly provide workshop training. The School of Data Science is a new school at UVA, formally established in 2019. They provide data science training to their students, but do not have the same service orientation that UVA Library and Research Computing each have.

For the UVA study, the research team interviewed 11 participants at different levels of research and from a variety of schools. The participants included five full professors, three associate professors, one assistant professor, and two research scientists. The schools that were represented were Engineering (3 individuals), Arts & Sciences (2 individuals), Data Science (2 individuals), Medicine (2 individuals), Commerce (1 individual) and Architecture (1 individual). Our attempt was to interview a broad mix of researchers in different careers and disciplines.

The participants were asked questions about a variety of topics, including i) how they acquired their data; ii) how they stored and analyzed the data; iii) how they shared data and research outputs; and iv) what kind of ethical and data sensitivity concerns they faced. The full interview guide can be found in Appendix I.

The remainder of the report summarizes the following themes and then presents recommendations and conclusions:

- Data Acquisition and Preprocessing;
- Ethics of Data Handling;
- Data Analysis and Programming Languages;
- Compute Infrastructure;
- Data/Code/Research Sharing;
- Teaching and Learning; and
- Advancing the Discipline.

Data Acquisition

The data acquisition process varies according to the discipline and the type of data being collected. The most common method used by the participants for obtaining data was through companies or government organizations. Collecting data from sensors – ranging from soil moisture sensors to hospital bedside monitors – was tied with receiving data from colleagues or alumni. One participant described actively collecting data by coordinating a team of interviewers to record sessions with participants.

In general, the participants did not have problems finding sources of data. The challenges began with the quality of the datasets and how to pre-process them.

Quality, Verification, and Pre-Processing

Some respondents, especially the ones more clearly trained in data science methods, mentioned issues of quality, and the time it takes to verify and clean data. **A considerable amount of time is spent verifying the data, attempting to discover errors, cleaning up non-numeric data, and understanding the dataset as best as they can.** They spend time looking at distributions, looking for outliers, determining a particular variable meets expectations, and understanding limitations that may come from dirty or missing data.

In addition to data quality, data completeness also came up as a concern. For many machine learning techniques, the data must have labels or classifications so that it can be used to train the model. This activity can be time-consuming and costly. One respondent was specifically researching the differences between human assessment versus outputs from AI coders, and more importantly, what is the threshold of error they can accept.

Finally, because many of our respondents use medical data, they must spend time de-identifying the data. In addition to removing names and social security numbers, they may be required to modify event times, so that patients cannot be identified by the dates and time spent in the hospital.

Fifty-five percent of the participants are working with sensitive or protected data. They must follow IRB guidelines or guidelines defined by organizations, such as HIPAA, for storage and handling of the data. Of the participants working with sensitive/protected data, 67% mentioned that they are responsible for de-identifying or anonymizing the data.

Challenges of Data Acquisition

Many challenges were discussed relative to data acquisition; however, there was not a common thread among the participants. Some of the challenges that were mentioned are highlighted below.

Funding proprietary data purchases

Researchers in less-resourced disciplines (e.g., urban planning) have a harder time acquiring data because they lack the necessary funds, or at least have to work harder to get those funds. The urban planning respondent expressed a desire to access proprietary data more easily. It can be expensive to license proprietary data, plus there can sometimes be time spent on communicating with vendors and negotiating terms. There is room for improvement when it comes to working with proprietary datasets from for-profit industries.

Receiving data in inconsistent formats

Respondents who use multiple data sources may find that the formats of the data are not consistent or even different names can be used to define the same characteristic. The researchers must spend their time getting the data into one common form: "all the data looks slightly different and just getting it all into a common set up takes work." This additional amount of work takes time away from the analysis of the data and can cause errors due to misinterpretation of the formats.

Downloading large datasets in a reasonable amount of time

One medical health researcher mentioned the size of data as being a problem. The respondent mentioned that it took one of their postdocs a month to download a set of Alzheimer's data. Transfer rates are dependent on the underlying network. Although the university has invested in faster data transfer networks, the speed on the network at the source can be the limiting factor in transfer speeds.

Acquiring datasets from the medical system

One respondent had some very specific challenges with the way datasets are distributed in the medical system. According to the respondent, it used to be that you could go directly to departments so that they would pull, for example, Public Health Data or Anesthesia data directly from the Central Data Repository. It used to take a couple of weeks to receive data; it was free; and it was pulled by staff who were intimately familiar with the data because they were the ones maintaining the databases. Now, it takes months, requires payment, and is pulled by centralized staff who are not as familiar with the data. This is an extra challenge to healthcare researchers at UVA. The pandemic likely exacerbated these problems since some staff on the medical side of UVA were furloughed for some months in 2020.

Receiving data as periodic dumps, rather than an up-to-date stream

One respondent noted that they get access to a third party's data, and that the third party is happy to share, but the sharing mechanism does not work as smoothly as they would like: "It's a data dump every once in a while. No real-time access to the information for us to be able to pull it on demand."

Ethics of Data Handling

When asked if there were ethical concerns when handling the data, 45% of the participants did have serious concerns, 27% of the participants had mostly no concerns because of the protocols that they

were following and 27% had no concerns at all because their data did not involve humans or were based on historical artifacts. Of the participants who had concerns, **the biggest concern was that the advancement of machine learning algorithms could reveal more about individuals than was intended**. In the cases where the researchers raised this concern, they indicated that it was the responsibility of the researchers to act ethically.

Other concerns that were mentioned included:

Are results sufficient for making life and death decisions?
The respondent who worked with flooding models commented: "How should you report [your result]? What if you're not very confident in it? What if you know a street is going to flood but you're not very confident? Should you still report it? Those types of issues will start coming [up]."

Can findings that are intended to help society be used for bad, instead of good?
The respondent working on policy-oriented topics, such as immigration and refugee camps, pointed out that any research published on those topics "can be used in a way that would be counter to what I think would be normatively good policy."

Can machine learning algorithms have biases that promote gender or race inequality?
If the data used for training the models have embedded biases, then the models will learn the biases for predictive outcomes. In general, there are concerns that machine learning algorithms are not tested under a wide range of inputs to determine if gender or race biases exist. One of the respondents talked about getting experts to help with studying characteristic or feature variables to "make sure that there is no inherent bias" in their models.

The psychologist working in neuroimaging pointed out that "brain data is unique to that individual," and that even without any metadata, it is possible to identify who the brain image represents, based on the shape and spatial layout. It is very possible to link medical data back to the individual, so this research team tries to be very careful about that. At the same time, this poses a tough challenge to biomedical researchers:

> *"It makes it pretty exciting, and we appreciate that; however, we're also very sensitive about being overly protective, such that we can't do any innovative science. So, we want to make sure that we are finding a balance and striking that balance between subject protection and scientific innovation. We don't want to make it absolutely impossible for anybody to do anything. Otherwise, why are we collecting all this valuable data?"*

Data Analysis & Programming Languages

The type of analysis performed on the collected data falls into four major categories:

- Machine Learning/ Deep Learning,
- Spatial Analysis (e.g., ArcGIS),
- Natural Language Processing, and
- Image Analysis.

Some participants are using multiple types of analysis on their data. Only one participant is using traditional statistical methods that do not fall into any of the other categories.

There were three primary programming languages mentioned by the participants: Python, R, and MATLAB. Approximately 75% of the participants said that they used more than one (and in some cases all three) of these languages. One participant described using a web-based tool for doing analysis, and thus, was not able to describe the programming language being used.

For the most part, the participants apply algorithms (i.e., pre-packaged codes) to perform their analyses, as opposed to writing their own codes. **While emphasis in training is on programming languages, it is becoming more important for researchers to learn the underlying concepts of the analytic tools that they use, and to understand the limitations of the software**.

Compute Infrastructure

Data Storage

When asked about short-term storage for their data, the participants described a variety of locations, including Dropbox, Google Dive, AWS, local database servers, customized servers, and leased storage through the two centrally-operated systems (for non-sensitive and sensitive data). **The consensus was that there does not exist a reliable solution for storing and maintaining Big Data at UVA.** The challenges that the participants described included:

- It can take days, weeks, or even a month to transfer data from storage to the compute platform,
- The cost of storage (with either AWS or the university's systems) is too high or cannot be sustained for multiple years, and
- Systems provided by the university tend to go away when funding runs out or advocates for the system leave the university.

The participants had similar concerns about long-term storage of their data. However, they do believe that the university's library will be part of the solution. One participant relayed his confidence in the library in this space: "Usually when I think about long term storage, I think about libraries for that." However, from the library's point of view, the library's data repository is not set up to handle big data-datasets.

Compute Platforms

Of those interviewed, over half were using the university's high-performance computing system or secure compute system; whereas a third were using commercial systems (like Amazon Web Services or Google Cloud); and only one participant was still running on a desktop or laptop. Nearly all the participants had moved off laptops or desktops because those systems could no longer handle the size of the data, or the processing power required for the data. However, moving onto bigger computation platforms introduced new challenges.

The number one challenge when working on computational platforms was dealing with the costs. The costs associated with using cloud computing have become prohibitive for Big Data. There are also costs for using the University's resources, namely the secure infrastructure resources and the data storage

resources. Finally, there may be a licensing fee for some of the programming tools needed for data analysis.

The second most-cited challenge when working with the computational platforms was the speed needed for processing the data. Although the participants have access to high-performance computing systems with graphical processing units (GPUs), there were concerns that the algorithms still may take days to run, or the jobs would have to sit in a queue waiting for the appropriate resources. As the volume of data increases, the time and resources needed to process the data will similarly increase.
Support for Research Activities

When asked if they received support to collect, store, or analyze their data, the most common response was that they do not go outside of the research team to get help. About 25% of the participants mentioned asking for help from the Library or the School of Data Science, even fewer mentioned Research Computing.

It is unclear why the researchers do not seek more support from service organizations, such as the Library and Research Computing. Further investigation would be needed to determine if there is a lack of recognition of the support provided by these organizations or if the organizations are not providing the services needed.

Sharing

Sharing Data, Code, and More

Many respondents cannot share data for various reasons (usually because the data are sensitive or proprietary), but they often share codes.

A researcher working with medical data indicated that they were working with a School of Data Science researcher who is very committed to open science, accessibility, and reuse of data and code. Those principles were being incorporated into their research project from the start. Even though the raw Protected Health Information (PHI) data they worked with is not likely to be shared, the research team intends to share code. The greater hope is that UVA creates a "*library of data and results and algorithms for the UVA R&D community to use for discovery of new knowledge or development of products or whatever. Moreover, to have metadata available outside of UVA such that anyone can tell what it is that's here if they wish to engage with people in our community towards work on that.*"

One engineering respondent cannot share social media data that they license from private companies, but they do share everything else that they can. A psychology respondent also noted that they are obligated to upload neuroimaging data to NIH's neuroscience data archive, and at that point the data are "doubly anonymized." The data are publicly available to anyone who wants it.

Many want to do even more than sharing data or code. **One theme was to work toward creating "workflow engines" that would tie data and code together with computation resources to perform analyses.**

One engineering respondent is working to create a website where users can submit images. The website would deliver results using the engineer's computer vision methods -- the idea being that the website processes the public's images with the research team's models with a very low barrier to entry.

A medical respondent pointed out that they cannot share sensitive PHI data, so instead they share analytics and software distributions to their collaborators. These are collaborators working on essentially identical projects at different research sites. The idea is to produce a secure virtual environment: labs will not have to send their data anywhere, the other labs do not consume UVA computing or storage resources, and the greatest benefit is that everyone's analysis is done identically to the others. Other labs can also submit code or packages to reflect what they want to see in the final analysis. The idea is to design the shared analytical environment to be compatible with any research lab.

A psychology respondent is an editor of a neuroscience journal with an emphasis on data science; at that journal it is expected that authors submit data (when possible) and source code, although the data or code can be made available through "GitHub or some other repository." Echoing what the engineering respondent said, the medical researcher commented:

> *"One of the things I'd love to be able to do is if we have a workflow engine, writing on top of our computational resources to be able to process these data. The file, the documents, the workflow description of what you did to the data is something that can also be shared and incorporated into supplemental materials for peer reviewed publications."*

They noted that the software tools are already open source, "but it's the assembly of them into a workflow that is a little bit of an art, in addition to the science underlying."

Community for Data Sharing

Many respondents mentioned disciplinary norms when discussing data sharing. Not all respondents felt that their disciplines really embraced data sharing. Even then, there's a growing awareness about open data: "*We haven't been, meaning our community, hasn't been really active in that. But recently people are starting to do that and we are also thinking about publishing our stuff online.*"

Other respondents pointed to disciplinary norms that fully encourage open data sharing. One said that they relied on disciplinary or journal data repositories, and fully embraced working with those. Another described that urban planning encourages data sharing in ethical terms, and avoiding coming across as a "free rider" which would give people an "unfair" advantage when everyone else shares their data.

Incentives for Data Sharing

To learn about incentives for data sharing, the interviewers asked the question: "Are there any incentives that exist either at UVA or within your field for sharing your data and code with others?" At UVA, there are 12 schools, and each can set different requirements and incentives for their faculty, so even at the one institution the responses can vary depending on which school the respondent is in. At one end of the spectrum, only one person said that, no, there were no incentives. On the other end is the School of Data Science; this new school is "recommending" that their faculty and research staff publish in open access journals and make data publicly available, as part of their promotion and tenure guidelines. Most respondents fell somewhere in the middle.

One mentioned that if you share your data and other people use your data, your citation count goes up, and that citation count is beneficial for tenure and promotion; the incentive there is indirect. Another mentioned grant funders (specifically, NSF) promoting open data and that some journals require open

data for publication, the implication being that if you want those grants or publications, you will need to share data and code.

Even the respondents who truly believed in open data recognized that there are not always incentives for getting people to publicly share their data. One respondent indicated that there is a "lag" in the benefit that comes back to researchers, and how they get credit. They did point to the new School of Data Science openness requirements, pointing out that that might be the thing that gets researchers' attention because it has such direct consequences. It will be interesting to see how the new guidelines at the School of Data Science are received within and outside of the school.

Challenges to Sharing Data

Respondents faced several challenges to sharing data, including:

Data Sharing is Prohibited

One obvious challenge to sharing is that not all data are allowed to be shared. This is especially true for proprietary and sensitive data. Some researchers have licenses with third-party companies in order to access data. Medical researchers often must deal with restrictions for sharing PHI data, as well. One research lab was a Data Coordination Center for a multi-centered National Institutes of Health (NIH) project. Even though the size of data was manageable, UVA would not sign off on any sensitive data coming in (from the other labs) or going out (to the other labs). This of course is an example of sharing data between labs on the same multi-site project, not sharing data publicly, and it goes to show exactly how much concern there is about sharing sensitive data.

Size of Data Makes Sharing a Challenge

Some of that sensitive data can be anonymized and shared. One medical researcher was ready to share results data, and to put it in LibraData. The original dataset was around 10 terabytes, while the results dataset was in the multi-gigabytes range. There were no files larger than 5 gigabytes (the limit to what LibraData can accept), but there were a lot of files, and the import process took a while; it was not a seamless process for the researcher.

Recreating the compute environment

There are some significant challenges to sharing data and code. This is made more apparent in disciplines that are not obviously heavy in data scientists, computer scientists, or statisticians. One engineering respondent compared their work to the practices in a different discipline. In the other discipline, replication of a study might be simply getting existing R code to run. But the engineering respondent pointed out that their computing environment is much more complicated to set up, and it would be too much to ask a reviewer to do that kind of work. Having reviewers recreate the computing environment successfully is a non-trivial challenge.

Tension between Sharing and Competitive Advantage

Intellectual property concerns and "competitive advantage" can also be a challenge. An engineering respondent noted that a challenge with sharing data or code might come up if there are possible financial gains. They noted that some research is funded by specific companies or organizations, and if the organization deems the data or research to have value, they will be less likely to share them. Any gains would have to go back to the company first.

A medical researcher noted similar constraints when working with the health system at UVA, since the health system might potentially patent or license the intellectual property associated with a study. The researcher indicated that even if there could be a benefit to the public, that "goes out the window [if you can] make money off of it."

Similarly, a business respondent who often works with corporate partners pointed out that in business, data are proprietary and a "competitive advantage," but from a pro-social and ethical perspective it is important to share knowledge, and from a scientific perspective, if you cannot share your data, nobody can replicate your results. The same respondent also noted that some journals now require data, and sometimes the researcher can share a small sample or aggregated data with consent from the company.

Collaboration and Community

Collaboration was a theme that came up frequently in responses to various questions. Collaborations happen across UVA, as well as with labs outside of UVA. One respondent strongly stated that their students, both undergraduates and graduates, are collaborators, too.

The business respondent was especially interested in interdisciplinary community and collaboration at UVA. They work with researchers from different schools at UVA. Different disciplines are working on their own cutting-edge ideas, but most researchers are not aware of what is going on in other fields. This respondent was interested in having more interdisciplinary workshops focused on emerging methods. In addition to fostering local interdisciplinary community and data sharing across Grounds, they saw UVA as extremely well positioned to encourage this kind of engagement because of the new School of Data Science. The new school provides momentum around data science across Grounds, and it is very well resourced. If this respondent's reply is any indication of enthusiasm for what the School of Data Science could bring to the rest of UVA, potentially more researchers would embrace cross-school, data-based communities. This respondent went on to say that the library could be a good host organization for creating this kind of local community.

Strategies for and Challenges to Collaboration and Community

Respondents described various strategies for creating and maintaining successful collaborations. Some strategies include:

- Documentation – Setting up requirements documents that specify the computing environment; describing expected inputs and outputs of functions they have written; specifying where the data are, the current state of the data, and where the executables are; naming files properly and structuring folders so that the organizational logic is clear.
- Cloud Services– Using cloud document services, like Google Docs and Dropbox Paper, for shared documents; using cloud storage, such as Google Drive, Box, or Dropbox, so that all collaborators have access.
- Version Control – Using version control such as Git/GitHub,
- Common IRB – Using a common IRB when working with researchers at different institutions, and
- Using UVA HPC – Relying on UVA's high-performance computing when non-UVA collaborators do not have access to that kind of resource at their home institutions.

At the same time, not all collaborations are easy or straightforward. One medical researcher discussed challenges they face. They cannot send data back and forth between research collaborators at other

institutions. If the UVA researchers want to see a particular analysis at another institution, they must describe the idea for analysis or send code for the other institution to run (and vice versa).

Another specific collaboration challenge for this researcher was the data source. Among the collaborating institutions, they all had data from one of two major vendors of electronic health record (EHR) systems. But the systems are different versions with different names for things. Nationally, there are groups working to solve that problem.

A second specific problem this researcher mentioned was that the medical data they use are continuous data, "which is a very bulky thing. Not many senders are undertaking that kind of investment to store and analyze that particular form of data, which we think is central to the problem." The specific form of data this researcher uses adds to the complexities of collaboration.

Strategies for and Challenges to Disseminating Research

Each participant was asked how they disseminate research results; the participants listed both traditional and nontraditional platforms for dissemination. Reponses included traditional academic platforms, social media or public platforms, and industry platforms.

Using Traditional Academic Platforms

All respondents said that they disseminate their research in some sort of academic platform. This typically means publishing in peer-reviewed journals and presenting at conferences. Several respondents mentioned posting preprints to disciplinary community repositories, such as Social Science Research Network (SSRN), arXiv, or bioRxiv. The humanist respondent intends to create a journal-encyclopedia, beyond what a traditional journal looks like, to publish interpretive results, and it would also have the tools and data available for anyone who wanted to explore an analysis more deeply.

The School of Data Science respondent works with a School of Education researcher, and that respondent outlined an intentional strategy for disseminating their work. They have been trying to be more "progressive" in the conferences and journals they submit to. Their research team is moving away from presenting at traditional education conferences, and instead are presenting at conferences that are more oriented towards data science, such as at the Academic Data Science Alliance conference, and Tom Tom Applied Machine Learning conference. Similarly, they are beginning to target journals that are focused on where education and data science meet, such as the Journal of Educational Data Mining.

Using Public Platforms

Almost all respondents also mentioned a strategy for public dissemination (outside of academically oriented journal articles and conferences). A few respondents mentioned social media as a main way to publicly disseminate their research. A business respondent mentioned that a research center they are associated with at UVA has a social media presence, so they will just have the center post to those accounts. Sometimes this respondent will also reach out to UVA Today, a website and newsletter that highlights interesting news about UVA, to publicize their work.

Only one respondent said that they simply do not communicate their research findings to the public. Most other respondents, even if they were not currently successful in communicating results to the public, did seem to intend to do so, or felt like they *should* be doing so. An engineering respondent said, "We have actually been thinking about getting more active in terms of like online presence and things

like that." The School of Data Science respondent said, "I think that that's a fair one is that we don't do enough of that. We should, but not really." Another engineer echoed the sentiment that they "should" promote themselves more: "I don't do as much as I should around that space. I don't have a Twitter account for example. It's becoming more of a thing where you've got to promote it a bit more." And a psychology researcher said, their team does not focus on it, but "we can always do better on that." It is interesting to hear that so many respondents felt like they should be doing more to promote their work more publicly.

Another big outlet for public engagement is invited talks. The social science respondent discussed being invited to more public talks and colloquia, especially with everything being virtual during the pandemic. An engineer also mentioned that they get media requests sometimes, to discuss their research, but that they do not seek those out necessarily. The urban planner mentioned being invited to similar public talks.

Some respondents were specifically interested in setting up interactive websites that would let the public engage directly with their research. An engineer wanted to set up websites for their research lab to share data and models. The urban planner was also interested in having a web portal that was public facing.

Finally, one engineer described the policy implications of their work. They chair a committee on the state level that is about coastal flooding and what should be done about it. It's not research but is more about providing big picture recommendations.

Working with Industry

A different aspect of non-academic dissemination is working with industry. Three of our respondents answered this question with an industry response. One engineer is working to commercialize their research. The social scientist and the medical researcher also replied with examples of working with private companies as another method of disseminating their work.

Challenges to Disseminating Research

The most common **challenge related to sharing data/code/research related to time and speed**. Peer-reviewed journals can take eight or more months to provide researchers with feedback on their manuscripts. Even getting published in the University newsletter, *UVA Today*, can take several months. The participants are concerned that the information is old news by the time it reaches the public.

One respondent had a particular challenge regarding media outreach. They specifically had larger media outlets in mind, such as the *Washington Post*, CNN, and *New York Times*. This faculty member had recently been contacted by a major news organization to comment on a story but ended up feeling uncomfortable with how the situation went. The faculty expressed concern about both directions of contact: either the faculty reaching out to media to highlight recent work, or news media reaching out to faculty to have them comment on the news. The suggestion was to **have the University help support interactions with the media, and to give faculty a media coach to help them figure out how to answer questions and assess any legal constraints.**

Teaching and Learning

Big Data Training

The participants were asked about the training that they had received for handling Big Data. Over half of the participants described their training as self-taught; a third described taking formal courses or degree programs to learn about tools for Big Data; and one participant had received no training, instead he passed the computational work to "other people." Also, a third of all participants said that they had attended workshops offered through the Library or Research Computing.

Of those participants who had attended workshops offered through the Library or Research Computing, 75% had suggestions for improving the workshops:

- The workshops are focused on beginner skills only – there should be workshops with more in-depth information,
- The workshops should be offered more frequently, and
- The workshops are too short to provide both information and hands-on activities – they should be extended to at least half-day sessions.

Future Trends

Finally, the participants were asked about how future trends in Big Data would impact training within their discipline. An over-arching response was that the prevalence of data collection and analysis will lead to the need for training in the culture, ethics, or social implication of handling Big Data. In addition, the participants believed that more interdisciplinary and in-depth workshops should be available. They specifically mentioned the graduate-level course "Computation as a Research Tool," stating that it could provide more formal training in analysis techniques and programming. Finally, the participants suggested that as more third-party analysis tools become available, students will need training on understanding the limitations of the software packages.

Advising and Training

A first piece of advice to trainees was to go online for initial training or to search for answers to specific questions. YouTube, Kaggle, GitHub, Coursera, and DataCamp were all mentioned. Additionally, respondents pointed to specific online training, such as Amazon Web Services (AWS) training.

Regarding the Library workshop series, one respondent said that they send students to check out the series. (The series most often focuses on R and Python, but will sometimes cover different topics, like QGIS, Designing for Data Visualization, and Reproducibility. The series was recently expanded to include Research Computing's workshops as well, which brings more advanced topics, such as accessing the local HPC and building containers. Depending on when the respondent participated in a workshop, the workshop might have been offered only by the Library or could have been the combined Library-Research Computing series. People register for the series on the library website.) One respondent attended a workshop event but was not impressed and would not return or recommend it to trainees.

One participant described the need for training students in local knowledge – information about what the research team is trying to do with data, and how the research team has made decisions regarding how to process that data. This knowledge can only be passed down from the researcher or from graduate students who have been on the project long enough to have historical knowledge.

Conference workshops and symposiums were also mentioned as another useful source for education. One respondent pointed to a two-day long educational program specific to the discipline and how to apply data science; that research team tries to send their trainees there.

A social scientist would advise the trainee to (1) take the formalized statistics program in their department; (2) if they need more in terms of analytic models that the department does not teach, go to the Statistics department or Psychology department; or (3) if the trainee has something novel in terms of data creation and collection in mind, go to Data Science. That respondent commented frankly that they had a preference to send the student to another department in the same school. (The School of Data Science is a separate school from the respondent.) Other respondents said they would also send students to Computer Science, in addition to the School of Data Science. Similarly, a Data Science respondent said they would point trainees to specific faculty in their department or to Research Computing.

Challenges with Advising and Teaching

A concern regarding trainees is teaching them to understand that a lot of time is spent cleaning and understanding the data, and a much smaller portion is spent analyzing it. One respondent, a research scientist, pointed out that they spend a lot of time getting the students to understand that point. It is unclear whether the students are taught about the data cleaning and how time intensive it is during their training.

One respondent at the School of Data Science reported that students need cloud-based infrastructure to "get to 21st century Data Science education." As they mentioned, once an SQL database or a machine learning model gets to a certain size, you cannot run it locally on a laptop. Teachers need to access to "steady state infrastructure to be able to teach certain things." One example solution mentioned was to have students purchase small pre-paid credit cards and use those to protect them from any accidental purchases above the free tier of AWS. Another suggested solution was to have an institutional purchase that would cover all instructional compute time and to have all access to the compute infrastructure be though a web browser. The respondent was not aware that UVA does offer comparable systems that can handle large data, use machine learning models and be accessed through a web browser.

Staying Abreast of New Developments

Big data and data science researchers have a wide variety of ways to keep up with technological developments in and outside of academia that inform their research.

Community

Many respondents mentioned community or social interactions as part of their strategy for keeping up with their field. One medical respondent described a journal club run by the research center they are a part of: one week, someone will discuss an outside paper related to their area of research, and the following week the same person will then present the research they are working on. A humanist respondent intentionally helped create an international community of scholars that shared the same research interest, and at the same time, worked to build analysis tools for this community.

A medical respondent said they keep up with other research centers doing work in the same research fields and attend their Zoom calls and presentations. Another medical researcher mentioned a weekly

meeting with their research team and described it as “another venue for keeping up with what's going on.”

Others rely on direct connections to people. A business respondent said they stay in touch with a lot of executives to keep up-to-date on computational trends. An engineer described how they are not on social media, but their project manager is; so, the manager will relay news items or postings of a few close friends in the field who will share their research results.

The urban planner mentioned keeping up with industry as part of their strategy to stay up to date with new developments. This respondent monitors conversations happening between cities and mobility industries. Much of the work they do is evaluating the mobility industry and seeing what kind of effect it has on cities. In this case, it is necessary for the researcher to stay abreast of what is happening outside of academia.

Conferences and Journals

Several mentioned attending conferences and keeping up with the literature. Most respondents mentioned conferences as a general response to this question about keeping up with the field. But one respondent noted that the pandemic illuminated part of what is so special about conferences: the random interactions with researchers in the same field. It is difficult to replicate that kind of serendipitous interaction in the current virtual-only environment.

Most of the respondents mentioned keeping up to date by reading journal literature. Respondents monitor the work of other researchers and other labs. This is tied up in the community aspect that was specifically mentioned by a few respondents – some of that community is maintained by monitoring journal literature.

News and Social Media

Both the business respondent and an engineer mentioned news as sources of information. They intentionally follow the news in their field to stay on top of new developments.

Interestingly, social media was not mentioned as a particularly strong theme for how to stay abreast of new developments. No one said in a very strong way that they use social media for staying up-to-date, even though some mentioned Twitter when disseminating their research.

As Needed

Many mentioned that they only keep up with new technology developments on an “as-needed” basis. One respondent commented, “If I ever encountered a data set where I really needed something like [PySpark], then I'd go out and start kind of looking for it. But I don't spend a whole lot of time doing that unless I have a need for it.” The specific, immediate needs of the research team often drive any search for new developments. Another respondent said, “Keeping up with the literature is a really tough problem and the way that I do it and the way that my group does it, is that the projects that we're working on drive what we're reading.”

Challenges with Staying Abreast of New Developments

A few people mentioned challenges, or what did not work for them. Information overload was one concern. One respondent noted that they get a lot of emails inviting them to webinars to learn about a

certain topic, but "you don't have time for everything." Another noted that during the pandemic, they had less time to read and stay up to date.

Another challenge with staying up to date with technology is that it can be tough to invest time and effort into learning a new tool, only to see technological changes advancing faster than their learning. This respondent who was comfortable with using Stata shared this brief anecdote:

> *"And it's funny, I spent some time last summer learning R, and now the folks I'm working with out of Stats or elsewhere are saying, 'We don't use R anymore. We do all this in Python.' And I'm like, 'Oh, come on. For the love of …'"*

Another respondent had something similar to say but re-framed the issue. Instead of seeing it as a challenge to stay abreast of new technology, they saw keeping up to date with cutting-edge tools as not necessarily optimal. The perception is that there is churn to technology, and a concern with the technology's longevity, making researchers a little suspicious about taking the time to adopt it. At least for this respondent, who was more senior in the field, there is a calculus about whether learning a new technology is worth the time and effort.

A data science researcher described that a lot of what large tech companies are developing are very cutting edge and more advanced than they would ever need for their research project. Instead of focusing on industry, they spend more time looking at academic research like theirs. Ultimately, they said they tend not to look to industry to keep informed of their field. Overall, most respondents, except the urban planner, did not mention industry when considering how they stay abreast of technological developments.

Advancing the Discipline

A theme that emerged from the respondents was that the use of Big Data and Data Science methods can "advance the discipline." **Data Science tools can bring a new focus to many disciplines, even creating an overall shift in disciplines and providing innovative approaches to answer old questions.** Respondents view data science as becoming embedded in their disciplines. They perceive the use of data science as analogous to using statistics or having statisticians on the team.

One respondent described civil engineering as a discipline that is moving away from new design and construction. There is less of a focus on building the next Golden Gate Bridge. It is moving towards maintenance of our national infrastructure, in this case via sensors and deep learning: "[There are] so many people who are doing this kind of stuff these days in civil engineering, so I think it's a trend." Big Data and Data Science methods can drive the shift in the discipline.

Some respondents discussed how Big Data and Data Science methods provide the ability to answer questions in a new way. For example, new sensor data changes the kinds of questions that can be answered – questions that could previously only be answered indirectly, in an aggregate way.

> *"These [small-scale, sensor-based] approaches allow us to understand social and spatial phenomenon at a much finer scale than we were able to in the past and allow us to potentially address longstanding planning objectives, planning questions, in a way that we just simply haven't been able to do in the past because of limitations to*

the data available and limitations to the computing, and analytic methods we had available to us, period."

This sentiment was echoed by a humanities respondent, who explained that digitizing and linking massive amounts of text allows researchers to do what they always would have done, but in a much more efficacious way. They can get things done in a brief period what would have taken weeks of intensive study and analysis.

Not all our respondents would call themselves computer scientists (far from it). Some respondents see Data Science and Big Data as ways to answer questions in their field. For many researchers, their discipline-based research is the main driver for them, and they will either learn data science methods to advance their research agenda, or partner with data scientists who can help. In the future, domain experts may want to turn to data scientists in the same way they turn to statisticians now.

Recommendations

Although the research environment at UVA is diverse, one challenge that was repeatedly mentioned was the handling and storage of data that is sensitive, private, or proprietary.

For research that involves sensitive or proprietary data, we recommend that:

- The Library and Research Computing provide more training for the handling of sensitive data, including what the different types of sensitivity are, where sensitive data can be stored, and where it can be processed.
- Research Computing plan for additional infrastructure where sensitive data can be stored and processed, to handle the future increase in sensitive data.

In addition, there are opportunities to help researchers in general.

For data acquisition, we recommend that:

- The Library helps acquire proprietary datasets or datasets from industry (e.g., social media companies). The Library has the infrastructure in place to license and store proprietary datasets but would require funding from other units for data acquisition (for example, funding from faculty, Provost, or Vice President for Research).[1]

For data and code preservation, we recommend that:

- The Library provides better support for LibraData to accept "big data," or market LibraData as being able to point to another site to find the files.
- The Library markets LibraData as able to accept code. The respondents seemed familiar with LibraData as a place to preserve data, but LibraData was never mentioned as a place to store code. GitHub and Zenodo came up most frequently as sites to share and store code.

For data processing, we recommend that:

[1] Additionally, the Library only licenses datasets that it can distribute to all university affiliates. See the UVA Library Data Collection Development Policy.

- The Library reviews the proportion of course time dedicated to data cleaning in the various departments and schools. This is potentially a type of training the Library and Research Computing could offer more intensively. Continue to market the statistical consultation services we already provide.
- The Library and Research Computing explore tools that provide "workflow engines" that can tie together data, code, and computing resources.

For development of community and collaboration, we recommend that:

- The Library and Research Computing provide more interdisciplinary methods and data-focused workshops or community building events across Grounds. The Library would be the host organization.
- The Library and Research Computing provide workshops and outreach materials highlighting the strategies that big data researchers and data scientists use for successful collaborations, for example documentation, file naming, organization, version control, and cloud storage and document platforms.
- The Library invites the local IRBs to collaborate on a workshop about designing IRB protocols for collaborations across multiple institutions.

For general knowledge, we recommend that:

- The Library highlights journals and conferences that are disciplinary in nature but that embrace Big Data or Data Science methods.
- The Library and Research Computing continue to provide workshops for programming languages, such as R, Python, and MATLAB, but also include workshops and consultation service that show how to efficiently handle Big Data.

Conclusion

Big Data and Data Science methods are used extensively across the schools at UVA. Furthermore, Data Science methods are changing how research is done in many disciplines.

However, researchers face challenges when dealing with Big Data. Some researchers must license proprietary data sets, but more often they are spending significant amounts of time downloading large data sets, verifying the quality of the data, and de-identifying the data to ensure compliance with IRB or HIPAA regulations.

The researchers take the ethical implications of their work seriously. They have clearly thought a lot about the sensitivity of their data but expressed concerns about unintended policy outcomes based on their research, and how the advancement of machine learning algorithms could re-identify personal information in anonymized datasets.

The researchers use a variety of tools for their analyses, with a focus on machine and deep learning, spatial analysis, natural language processing, and image analysis, using languages such as Python, R, and MATLAB. A variety of platforms are used for storage and computing, with most using the UVA systems (e.g., high-performance or secure) or a commercial system, such as Amazon Web Services (AWS). Although a commercial system can be extremely costly, some researchers still use these systems.

All researchers were excited about sharing their results from Big Data and Data Science Methods, but their major concern was that traditional outlets, such as peer-reviewed journals, take too long to publish their information. They would like to receive more support in finding other outlets for their research.

Most researchers seemed engaged with the conversation about open data but often faced limits of what they could share based on the proprietary or sensitive nature of their data. Even so, many researchers share their codes or other research outputs to places like LibraData, Zenodo, GitHub, or disciplinary or grant-funded repository communities.

Incentives (or perception of incentives) to share their data and code vary from feeling like there were basically no incentives to share, all the way to the new School of Data Science that is building in recommendations for data and code sharing in its tenure and promotion policies. Most researchers fell somewhere in the middle, feeling that incentives to share were indirect, such as expecting citation counts to go up when sharing, which in turn would impact tenure prospects.

The researchers came to big data and data science with a variety of experience. Some were formally trained in these methods, while others went out of their way after their graduate training to pick up skills. All do what they need to stay up to date in the field. They also had very practical ways of staying organized in collaborations, which sometimes span multiple institutions.

Although not asked directly, many researchers mentioned that Big Data and Data Science Methods are impacting their disciplines. These methods have provided innovative approaches to answer old questions and are shaping the future direction of their discipline.

The training of graduate students is tied to Big Data and Data Science methods as well. The researchers, speaking as teachers and advisors, indicated that the university as a community needs to provide better training options to give students the most modern learning experience.

Overall, Big Data and Data Science methods are seen as an integral part of quantitative research. By providing training and resources for Big Data, the Library and Research Computing can contribute to the success of many research projects.

References

Ithaka S+R. (2020, May 14). Launching Two Projects on Supporting Data Work. *Ithaka S+R*. https://sr.ithaka.org/blog/launching-two-projects-on-supporting-data-work/.

UVA Library. Data Collection Development Policy. https://www.library.virginia.edu/policies/data-collection-development-policy/.

Acknowledgments

Thank you to our supervisors and work teams who encouraged us with this work and understood when it was taking so much of our time. Thank you to our interviewees – they graciously provided us with their time and insight, and without that this report would not have been possible. Finally, thank you to the Ithaka S+R team who provided opportunities for our training and showed patience with our questions.

Appendix I: Semi-Structured Interview Guide

Following the training and directives provided by Ithaka S+R, the local research team performed semi-structured interviews. We relied on the interview guide provided to us by Ithaka S+R, included below. The interview guide ensured consistency across interviews, and at the same time allows for probing and follow up questions to get a better understanding of each interviewee's perspective and experience. This allows us to balance breadth and depth. Questions may be re-arranged or slightly re-worded at the time of the interview.

Supporting Big Data Research

UVA IRB-SBS #3811

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) incorporating data science methods that you are currently working on. If you have multiple projects in process and if you prefer, you can refer to a single project when answering questions.

- 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?

- Does your data require special handling due to its sensitivity (e.g., does it need to meet HIPAA, FERPA, or other requirements)?

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?

- Could you describe briefly the primary analytic methods or modeling approaches you use with data? E.g., causal inference, machine learning, neural networks, image processing, NLP, inferential statistics, etc.
- What software or computing infrastructure do you use? Examples: programming languages, high-performance computing, cloud computing
- Do you encounter any challenges with short term storage of this data?
    - If not already mentioned: What resources do you use to store your data? E.g., cloud, Box, Google, hard drives, UVA's value storage, etc.
- What challenges do you face in analyzing or modeling data?
- If you work with a research group or collaborators, how do you organize and document 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 data or code 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?
- Do you encounter any challenges with long term storage or preservation of your data?
- 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 or data science methods? 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 or data science methods?

Wrapping Up

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

Appendix II: Invitation Email

Supporting Big Data Research

UVA IRB-SBS #3811

Sample Recruitment Email

Subject: UVA study on supporting big data research

Dear [first name of researcher],

In collaboration, UVA Library and Research Computing are conducting a study on the practices of researchers who use big data or data science methods in order to improve support services for their work. Would you be willing to participate in a one-hour interview, via Zoom, phone, or in-person in accordance with UVA COVID guidelines, to share your unique experiences and perspective?

Our local UVA study is part of a suite of parallel studies at 20 other institutions of higher education in the US, coordinated by Ithaka S+R, a not-for-profit research and consulting service. The information gathered at UVA will also be included in a landmark capstone report by Ithaka S+R and will be essential for UVA to further understand how the support needs of big data/data science researchers are evolving more broadly.

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

Sincerely,

Jennifer Huck, UVA Library

Jacalyn Huband, Research Computing

Sample Recruitment Follow-Up Email

Dear [first name of researcher],

Thank you for expressing your interest in participating in this study. We would love to set up a time to interview you at your convenience. Please advise me of your availability within the days and times listed below: .

[time frame]

Before the interview begins we will ask you to provide verbal consent in order to ensure that you understand the study and are willing to participate in it. We are attaching the Study Information Sheet and Oral Consent Template to this email in case you'd like to look over it now.

Sincerely,

[investigator]

On behalf of

Jennifer Huck, UVA Library

Jacalyn Huband, Research Computing