9-28-2021

Supporting Big Data Research at Case Western Reserve University

Jen Green
Case Western Reserve University, jxg854@case.edu

Ben Gorham
Case Western Reserve University, rxg498@case.edu

Roger Zender
Case Western Reserve University, rez7@case.edu

Lee Zickel
Case Western Reserve University, lxz11@case.edu

Em Dragowsky
Case Western Reserve University, mrd20@case.edu

Author(s) ORCID Identifier:

Jennifer Green

Ben Gorham

Supporting Big Data Research at Case Western Reserve University: An Ithaka S+R Local Report

Jen Green, Team Leader, Freedman Center for Digital Scholarship
Kelvin Smith Library, jennifer.green2@case.edu

Ben Gorham, Research Data and GIS Specialist, Freedman Center for Digital Scholarship
Kelvin Smith Library, robert.gorham@case.edu

Roger Zender, Associate Director, Creation & Curation Services
Kelvin Smith Library, roger.zender@case.edu

Lee Zickel, Research Computing Technologist
Research Computing, lee.zickel@case.edu

Em Dragowsky, Research Computing Technologist
Research Computing, em.dragowsky@case.edu

Background

Ithaka S+R Research Study

This report is an investigation of the research practices of faculty and research staff who utilize or support data science or "big data" methodologies at Case Western Reserve University (CWRU). The study was conducted by librarians and library staff within the Kelvin Smith Library (KSL) in collaboration with staff within CWRU University Technology ([U]tech), and was part of national selection of parallel studies occuring at public and private academic institutions throughout North America.

The study was coordinated by Ithaka S+R with the goal "to understand researchers' processes in working with big data toward developing resources and services at [name of your institution] to support them in their work. The study contributes to the wider fields of library and information studies and data science, within the context of the evolving relationship between libraries and data science research support."[1] Participating institutions conducted local studies of researchers representing a broad range of disciplines to determine their big data use cases and needs in order to compile independent research results and recommendations for creating and enhancing big data support services locally. Additionally, each participating institution contributed their findings to a final capstone report written by Ithaka S+R. This is a cumulative summary of big data needs and recommendations representing public and private institutions more broadly. For more information on the methodology of this study, please see Appendix A and Appendix B.

Research at Case Western Reserve University (CWRU)

CWRU is an independent Doctoral University located in Cleveland, OH, and holds the Carnegie R1 Highest Research Activity classification. [2] With an enrollment of 5,430 undergraduate and 6,035 graduate students, the University ranks 42nd among 312 national universities (U.S. News and World Report).[3] CWRU comprises ten schools, which represent around 115 departments or programs.[4] A selection of those were represented in this study, and they include:

- Department of Materials Science and Engineering
- Department of Radiology
- Department of Biomedical Engineering
- Department of Genetics and Genome Sciences
- Department of Mathematics
- Department of Cognitive Science
- Population and Quantitative Health Sciences
- The School of Applied Social Science

[1] Ithaka S+R Big Data Ethics Review Instructions
[2] https://case.edu/ir/sites/case.edu.ir/files/2020-10/at%20a%20glance.pdf
[3] https://case.edu/ir/cwru-facts/university-rankings
[4] https://bulletin.case.edu/departments/

- The Comprehensive Cancer Center
- University Libraries
- University Technology Department

Limitations

Attempts were made to recruit a diverse group of up to fifteen faculty, researchers, or staff representing a broad range of disciplines and departments. The final pool of twelve interviewees represented those who responded to targeted interview invitations (Appendix C: Recruitment Email), with significant representation from STEM disciplines This may have impacted the study as it contributed to disciplinary consistencies in perspectives, approaches, and big data methodologies. In addition to participant demographics, the global pandemic of 2020 emerged contemporaneous to this research study. This situation presented limitations in faculty, researcher, and staff availability and dictated the need to conduct all interviews remotely via Zoom rather than in-person. The virtual nature of interviews may have impacted our findings for many reasons, but the most tangible being that the physical locations of our participants was out of the study's control and therefore varied greatly. During the interviews for this study, consistent big data themes arose across diverse research projects underway at CWRU.

Findings

This study seeks to examine researchers' practices in working with big data/data science methods in order to understand the resources and services that researchers at Case Western Reserve University need to be successful in their work. Big data is typically described as data used for research that is high in volume, velocity, and variety. Big data projects and project roles range broadly in scope and diversity at CWRU, and some of these projects and their data management needs are already supported in parts by CWRU's Research Computing department and the Kelvin Smith Library (main campus library). While CWRU's affiliation with University Hospitals and host of a thriving medical school positioned CWRU researchers well to define their research around the 2019-2021 COVID-19 pandemic, pre-COVID research within other disciplines continued to thrive. This study revealed diversity in big data projects and project roles (Appendix D: Project Types and Descriptions) and well as challenges related to big data project management and University policies around sharing data across institutions.
.

Big Data Research Project Roles and Themes of Working with Big Data at CWRU

Big Data Research Project Roles
The role of big data within the diverse project types listed above emerged in consistent themes. The following section will provide a summary of these findings, with more specific details to follow in the "Big Data Methodology" section. Generally, data formats range from images (e.g.,

scans and MRIs), environmental measurements, text, audio, and video used with machine learning applications, and biomedical or genomic data used to improve medical imaging or develop health diagnostic tools. Scanning and imaging data were represented more frequently across big data projects included in this study, and their file sizes (opposed to file quantities) as the primary characteristic for defining these as big data.

Themes of Working with Big Data at CWRU

Consistent themes also emerged regarding the methods by which data are captured for projects represented in this study. Most of the projects involved imaging of some type to capture data (e.g., scanning, photography, MRI). Spatial views are also commonly deployed within CWRU projects, and allow researchers to capture environmental data. GPS data is also used frequently for GIS mapping. One outlying example is represented in the work of one researcher who indicated that environment data "can be better for research than satellite data because it is real time data." Another scholar runs a worldwide consortium of researchers across hundreds of national and international institutions (see Red Hen Lab redhenlab.org). This group relies heavily on existing datasets within academic and private sectors to teach students coding for the construction of big data projects. In this case, they are using big data partly to bring together people from diverse backgrounds and fields when they usually don't have the opportunity to talk with each other or even know how to talk to each other, and some members within the consortium are considered "interdisciplinary translators." Data is leveraged to identify linguistic patterns and to be the connection and the equalizer for scholarly collaboration and multimodal communication across diverse research communities.

Researchers in this study indicated two primary challenges in their data methodologies: collecting data and transferring data. In terms of collecting, the issue is that mechanisms to collect data change frequently and it can be difficult to keep ahead of those developments. One researcher noted that "Sometimes the technology is ahead and sometimes behind, so we're always taking time to figure that out." When it comes to transferring big data, researchers tend to face challenges as they navigate big data sets through web portals. This process can be slow and sometimes unstable. These challenges in data collection and transfer could potentially be reframed as opportunities in CWRU's big data discussions and innovations.

Storage

Research that requires large scale data sets is common at institutions like CWRU, and this study reveals various needs and approaches to storing data at high magnitudes. Reliable methods for downloading and transferring large data quantities from one storage location to another are essential. Most projects required hundreds of terabytes of storage for their data projects and needed the data to be accessible across different projects supported locally or beyond. One researcher states, "we have about 200 terabytes of storage for our data projects and we draw from that data to use collaboratively across different projects we support." In pursuit of large-scale, adaptive, and high velocity datasets, for CWRU researchers big data storage needs to support "time-series powered data" and "data the power plants produce." To date, CWRU has supported cloud computing as big data storage solutions (e.g., AWS, Google, and Box), however participants of this study identified challenges in that support model. While AWS is reliable, it is costly and an expense that falls on the department, unit, or organization that supports the

project. And, when large data sets become inactive, it is not always cost effective to continue supporting their storage on AWS. Google and Box are alternatives, but one information technologist participating in the study noted that "services like Google are eliminating storage models, forcing big data transfers, sometimes abruptly." This unreliability of service posed challenges with project sustainability, data permanency, and collaboration.

Privacy, Ownership, and Permission

CWRU has a strong medical program and is affiliated with the university hospital system, which means that patient data privacy emerged frequently in conversations with researchers and information technologists as an important issue that impacted big data use and storage. This is true when CWRU researchers collect their own project data, and there are also examples where researchers work directly with companies who provide them with identified data so that it can be "de-identified, merged, and studied by our center[s] to report back to the company."

Challenges that emerge around privacy, ownership, and permission are primarily concerned with the availability of HIPAA compliance on cloud storage solutions. For example, CWRU has relied on Box as a HIPAA compliant platform for storage of large sets of private data, but the future of Box's status as HIPAA-compliant is uncertain. If Box loses its compliant status, projects that have been approved by the IRB to use Box for private data, will need to change their approach and potentially their previously approved IRB application. Collaboration with multiple institutions can also be challenging from this perspective as each partnering institution may enforce different data privacy compliance guidelines and support different data management and storage practices and tools. In regards to permissions, University restrictions on access make it difficult to collaborate with other institutions. Researchers expressed frustration with the notion that collaboration is encouraged, but difficult from a permissions perspective. It is the nature of many research collaborations that participants from varying institutions are beholden to varying policies around data sharing, storage solutions, and privacy limitations. In such cases, there is a need for data storage and sharing policies that transcend individual institutional mandates while not compromising the security of the data or the project. In some instances, researchers working with big data found their data storage systems compromised by external, unauthorized intrusions by malicious actors and noted the need for secure systems to prevent such incidents.

Analysis

Projects in this study used a variety of approaches to analyze data. Common examples within this study include analysis of medical data and health institution administrative data In the case of biomedical research, researchers employ statistical analysis, regressions, covariant structures, elastic net regression, penalized regression, and linear regression models to understand, among many other things, "distinct genome variables in humans." Time series recordings are used to observe multiple nerve cells and activation over time. One project employs a data analytics team to work with local and statewide hospital medical data in order to manage intake and do the analysis. For another project, biologists tabulate data as well as ensure that it is annotated and heterogenous. Then software is used to understand the annotations and parse it. One researcher states "big data helps us do computational simulations. We can build a network of data to represent the kinds of things observed and measured in a lab." Another explains, "we construct a mathematical model to analyze behaviors from a mathematical perspective."

Data analysis varies across disciplines and project needs; however, consistencies emerged within the study as they pertain to data analysis. Most research investigators are presented with the challenge of "messy" or "problematic" data, which can be time consuming and even difficult for humans to interpret. The amount of data also poses challenges when it comes to analysis, and projects will often employ students to work on cleaning up large quantities of "noisy" data. One problem that emerges here is that students often need time to acquire the domain specific skills required for the project to run efficiently and smoothly for all collaborators. As one researcher puts it, "For example, a student may be good at algorithms, but he doesn't understand how to apply that to the system we are working with. In one situation where we didn't address this well, the student was frustrated and the collaboration was frustrated."

The Big Data Lifecycle at CWRU

Integral for an understanding of the landscape surrounding big data at CWRU is an engagement with how members of our research community are involved in data collection and generation. The concept of big data means many different things to different members of the academy, and thus we anticipated a diverse set of responses when asking our participants about the processes by which they came to work with their data. Throughout this study, each participant was asked to explain whether they primarily generate their own data for their research – via instruments, surveys, codes, etc. – or whether they found and collected data from outside sources – secondary datasets.

Though the small sample size covered in our survey precludes authoritatively quantitative inferences, it is worth noting that the majority of subjects engage both in the production of new data through various methods and in the acquisition of extant data from external sources. Among those who generate their own datasets were research labs producing terabytes of audio/video files, extensive human genome sequencing producing hundreds of millions of variables per subject, time-series datasets being produced by solar arrays and capture stations, image sets from MRI scanners or automated stochastic models, and many more. The sheer variety of types of big data being produced via these processes is daunting, and poses challenges in assessing the ways in which our researchers then process and analyze their data. Many of these initiatives producing their own large datasets also source similar datasets from external collaborators, repositories, and databases, augmenting the volume of data produced by their own labor with external comparanda. Those participants who focus their efforts on secondary data derived from external sources tended to be more oriented towards the humanities and social sciences, with mathematics, physics, and engineering adopting a more hybrid approach to data production and collection. Regarding the collection of data from secondary sources, most participants adopted a multimodal approach, taking advantage of public, curated databases, web scraping, and APIs to locate and access data. Many participants noted the importance of external collaborators for the acquisition of such secondary data; knowing someone in the hospital system, local government, or specific industry that was connected to desired datasets is an excellent facilitator for many researchers seeking to acquire data from external sources. Other avenues of secondary data collection included research labs who monitor and collect television, radio, and internet broadcast

signals to gather publicly available data on patterns of speech, gesture, and language.

It should come as no surprise that the methods by which subjects in the present study undertake the analysis and modeling of their datasets is as varied as the subjects themselves. In many instances, big data analysis was performed via statistical software packages such as STATA, SPSS, and R. There was a pronounced utilization of Python observed for many of our interviewees, as well as a reliance on command line interfaces more generally to access, query, and analyze the datasets. More specialized platforms such as ArcGIS and QGIS play a role in many projects with geospatial or imagery-rich datasets. Two interconnected themes are visible in participant responses to how data was accessed and processed. With respect to access, many participants noted the difficulty in knowing where to look. The plethora of data that is available is overwhelming yet disparate, with countless unknown repositories and resources of which the researcher was unaware. Concomitant with this big data diaspora is the heterogeneity and diversity of the data itself. In some cases resulting from diverse environments in which it is stored or conflicting requirements on who can access it and how it can be used, many participants noted the difficulty in finding uniform coding solutions. The absence of authoritative schema that fit a particular lab or research agenda's needs often presents an obstacle in the timely processing and analysis of the data. Before researchers could reach the analysis stage, much labor was spent in the cleaning, recoding, reformatting, and general idiosyncratic processing of raw datasets. These challenges prompted many of our participants to express a desire for better exploratory interfaces for big data; more singular destinations or one-stop-shops for their secondary data accessing needs, and ideally more widely adopted and better communicated authoritative standards and schema for such data.

Numerous challenges are faced by researchers working with big data around issues of privacy, ethics, and computing environments. There is a pronounced concern around the de-identification of any data from or respecting human subjects, and researchers noted often that such concerns were generally addressed by their IRBs and administrative or funding requirements. Thus, the process of de-identification is standard enough practice to be less of an inconvenience and more of a procedural expectation. Researchers are careful to utilize storage solutions that comply with HIPAA standards whenever necessary, using only Box as a HIPAA compliant cloud storage solution and working locally within CWRU's Secure Research Environment (SRE) to ensure the security of their data. The SRE was frequently named as a collaborator for many of our big data researchers for whom data privacy and security was a noteworthy concern.

Parallel to the SRE, our participants frequently collaborate with the High Performance Computing (HPC) division of our campus IT department to access multi core processing and distributed computing for their data analysis needs. Frequently, researchers turn to members of the HPC for assistance in establishing custom databases or computing environments to facilitate data processing. Many of our participants have turned to cloud computing for some or all of their data processing, recognizing it as a means by which a geographically and temporally distributed research team can most efficiently handle large amounts of data in a standardized computing environment. Perhaps unsurprisingly, those same researchers who reported utilization of both cloud-based and local computing environments for their data storage, process, and analysis also raised the issue of transfer speeds. Efficient, reliable transfer of large quantities of data to and

from the cloud is a pronounced concern for many of our researchers, and there is a notable need for improvements in intelligent compression to help mitigate the timing and workload challenges that arise from using both cloud and local environments.

Research Communication

Big data projects produce communication deliverables that include the data sets, academic publications, and social media. This study addresses all forms researchers use to communicate their own research, share their project results, and stay current with developments within academia and beyond.

Sharing research findings

Participants in this study value sharing their work and do so in diverse ways including via social media, journal publications, reports, conference presentations, preprint servers, institutional repositories, workshops, science carnivals, data science colloquia, virtual workshops, and grant writing. Top venues for sharing included conference presentation (9 out of 12), journal publication (8 out of 12), and social media platforms, especially Twitter (5 out of 12). Only one researcher indicated that they publish in open access journals. Despite the frequency of subscription journal publication, more than one researcher noted that authors within their field prefer open source publishers or preprint servers (as a professional best practice) because their work is easily accessible and therefore generates more attention and discussion than journals hidden behind paywalls. It was noted by many participants, however, that there are no particular departmental incentives for sharing data and findings in open access avenues; researchers recognize the benefits themselves but are not specifically encouraged to pursue open access opportunities..

Aside from publishing, virtual workshops and conferences replaced in-person events this year. There were mixed feelings expressed about the success and future of these virtual opportunities, but nearly all of our study respondents indicated that conference presentation is a standard way for them to share findings on at least an annual basis.

Regardless of where they publish or present, time poses a major challenge for most researchers. Typical time constraints, such as teaching and leading lab students were prevalent amongst participants, but more interestingly were time challenges imposed by publishers and the publication process itself. One researcher said, “I no longer share my work through research publications because they are too slow and very few people have access to them. It can take years to write an article after the work is complete, and then another 18 months to publish it. I work in multimodal communication.” Another researcher stated, “the publication process has been a challenge. Our monograph was supposed to have come out five years ago. Then four, then three, and it’s still in the publication process. All of our data is part of that monograph process and it’s hidden until it’s published, so nobody accesses it.” Researchers within this study also commonly shared research findings in preprint servers such as BioArxiv. Preprint servers and social media allow most of our participants the opportunities to share their findings before the formal publication process, which serves them well for summaries of their research statistics and results.

These formats also allow for immediate feedback before engaging in the peer review process of formal publication. However, despite these challenges, one participant observed that, "data has more credibility after it's published via a scholarly publication outlet" and "I don't want to share scripts in detail before they are published."

Sharing data

Overwhelmingly, participants agreed that they believe sharing their data openly is important and they make attempts to share both data and code in a data repository as well as use data that is openly available. Several researchers indicated that data acquired from social media platforms, for example, can be incredibly informative as well as inexpensive. However, it was also frequently expressed that intentions to share data are often intercepted by the obstacle of time, funding, and privacy. This suggests that more researchers are using big data than are sharing it, an obvious obstacle to fostering a more sustainable open research environment

Issues of faculty and researcher time have been well-documented, and this won't be considered in great detail for the purpose of this report, however, the prevailing sentiment is that once a researcher finds time to write, they tend to focus on that and would share more data if there were easier (more automated) ways to do it and the writing process reaches completion (i.e., from the publisher's platform).

If researchers breach time hurdles, platforms for sharing are numerous, well-known, and openly available to most. However, some participants complained that there is a lack of consistency in data sharing environments, which makes data interoperability and user navigation a challenge. At CWRU, the Open Science Framework (OSF) is a well-known platform by which to share large sets of data and documents openly within and beyond the institution. At least one researcher is working with the library's digital scholarship center to use OSF in order to share both content and data there. Google Drive and Box were also noted as popular choices for sharing content internally and externally. Many participants indicated that they publish scripts and share portions of code on Github or Docker Hub so that "people can test their own data with the prototype." One researcher states, "we use Docker containers to help us set up software environments, run some analysis and publicly disseminate software and code."

Researchers participating in this study are generally more willing and/or able to share their code than they are their data. This seems typically due to issues of privacy and/or intellectual property concerns. For example, researchers in clinical fields indicated that sharing data is "rare in our field" because it needs to be anonymized in order to avoid HIPAA violations, and that "releasing patient data can bankrupt an organization." Some projects outside medical disciplines were noted to simply be confidential and impossible to share pre-publication. At least two researchers identified a need for a code repository (similar to preprint repositories), and they have attempted to initiate that. For those investigating this kind of initiative, it was mentioned that the resource COPE (Committee on Publication Ethics) provides helpful guides and standards. In the meantime, some researchers are using Overleaf to share data with other national labs and institutions who are provided permission to login to CWRU clusters and high-performance computing environments. A concern tangential to privacy is that of intellectual property. More than one researcher expressed resistance to sharing data, which stems from a desire, as described by one

researcher, to "hold data for as long as possible and not let anyone look at it." This prevents the risk of giving away one's claim to publish the work.

At times, funding is a barrier in establishing methods by which, and platforms on which, to share data. One researcher said, "we collaborate with other funded institutions or centers to manage, store, and disseminate data, and this serves as a 'hub' for a lot of the data analysis we do." Some researchers expressed an interest in developing "homegrown" solutions for data sharing, but funding and resources is cited as a barrier for that. Everything from ensuring robust transfer speeds, leverage enough cloud computing space, staff to code systems that support multiple geoportals so that students or faculty can contribute data across institutions, and hiring curators to manage the data are all essential components that require funding that is not obviously available. Comments on the financial aspects of data sharing revealed that researchers are informed and practical about the work involved in building and maintaining something like a data warehouse. For example, one researcher noted that "The biggest challenge we often face is we'll have a good idea [for sharing data], but unless you've got a way to keep something sustainable, what's the point?" and:

> "Curation of data is essential, but challenging to coordinate from the lab perspective. There is not enough time to plan for something like this. We'd need somebody to spend at least 20 hours per week acquiring the data and curating. A warehouse of data that is continually populated by others and curated within our institution would be very useful." And that larger institutions are collaborating with commercial partners (e.g., Facebook) who can contribute "hundreds of thousands of dollars to de-identify and secure data. It's rare to do because a lot of organizations/institutions don't have the resources to support that."

Some participants indicated that publishers are "incentivizing" sharing by enforcing a requirement that all authors who include computational models as part of their published work must make their data available either through GitHUb or ModelDB, a model database hosted by Yale University. However, all participants indicated something to the effect that CWRU does not provide concrete incentives for sharing, but that they attempt it anyway as it is the "right thing to do." Another comments, "People aren't sharing their data openly enough, and that is a big issue. There is more sharing when it is a requirement of a grant, but if it's not a requirement, people don't share enough."

Staying Current

Study participants are staying abreast of their field by referring to the same venues to which they publish their findings and their data. In addition to that, Google Scholar was noted by some as a great resource for accessing pre-prints, conference proceedings, and some publications. Most participants are reading disciplinary blogs, subscribing to vendor listservs for updates, and using Hadoop and Spark to keep up with new developments in the field. One researcher favorably described participation in science carnivals that offered "birds of a feather" opportunities that offered ways to network and engage in more focused discussions with colleagues. The most significant challenge to staying abreast of one's field is time management. Nearly all researchers indicated something similar to this participant's observation: "I feel like I have my foot in four

different worlds, and it's difficult to stay up to date on everything." Another indicated that there are "not enough hours in the day, so I take weekend time to experiment with new tools and resources."

Training and Support

Part of the survey conducted at CWRU involved conversations around training and support for big data, both what has been made available as well as what researchers hope to see in the future. The results of our study here are perhaps somewhat surprising, as the vast majority of our participants reported having received little to no formal training around issues of big data. Respondents with proficiency in big data analytics, processing, acquisition, etc. seem to have acquired such proficiency gradually through industry experience, through practical applications on site in lab work, or via informal knowledge acquisition processes involving publicly available online resources. Very little in the way of structured coursework or workshop series seem to have gone into the development of skills around big data; researchers at CWRU noted the general absence of such offerings in their formal educational opportunities. Such lacunae seem to have been compounded by challenges around the training opportunities which researchers had encountered: generally such instruction was either far too broad (e.g.: "What is big data?") or far too specific, not tailored to the sorts of research questions they themselves were concerned with. Researchers expressed a desire for more a-la-carte training opportunities whereby they could select exactly the components relevant to their own work rather than being overwhelmed with tangentially related seminars.

When asked how participants would recommend new students or collaborators acquire big data expertise, the general attitude was that of institutional knowledge, passed down from senior members of a research team to those more junior. This approach seems in part to ameliorate the difficulties in finding appropriate training via external methods: since the PI and senior researchers on a project have become especially well-versed in the skills, languages, and workflows necessary for the operation of their research, they are able to then provide the very a-la-carte training opportunities that they were unable to find themselves. Common to many of our participants' responses was a recognition that, when confronted with a need for expertise that they did not already have, the best option was to locate a collaborator who already had the required expertise. Rather than reinventing the wheel, researchers who deal with big data seem inclined to leverage existing networks of specialists to tackle niche problems and offer training to junior members tailored to the precise needs of their research.

Looking toward the future, respondents noted that training opportunities on machine learning, AI, statistics, programming, and cloud computing would be especially valuable. Underpinning many of these responses was the idea that, more so than training in specific programs or tools, researchers need to be primed to take advantage of the vast network of resources to which participation in academia grants them access. Being made aware of who is out there, what is available, and how to integrate it with their own projects seems to have been the most pressing need for future researchers in the arena of big data.

Conclusion

This study revealed that big data projects at CWRU are already numerous and diverse, qualities that present both strengths and challenges for the campus. Currently, there is no dedicated, centralized, and structured support system specifically designed to support big data research, therefore researchers are designing disparate and sometimes inadequately supported approaches to their work. In light of these factors, one primary conclusion is that CWRU would benefit from establishing a centralized knowledge base to support research throughout the life of a big data project (much in the way that the library and information technologists have supported researcher life cycles to date).

Throughout interviews, it also became clear that in the absence of a centralized big data support structure or knowledge base, researchers look in different directions to find the resources they currently need. Even when they find the resources needed to support big data, researchers are still looking for ways to identify and connect with other local big data experts in order to share information, expertise, training, resources, and opportunities for collaboration.

Therefore, a secondary conclusion is that CWRU would benefit from first determining where current big data researchers seek those resources and then identify and establish a centralized virtual and/or in-person space to serve diverse projects and project goals. This could be approached by establishing a steering committee composed of information and technology experts as well as faculty engaged in big data research in order to determine the best ways to harness big data work and design centralized structures and guidelines that will support the long term big data needs at CWRU.

The landscape of big data research at CWRU is already expansive and diverse, but continues to grow. Researchers interviewed for this project depicted a situation in which there exists an abundance of interest in and utility of big data approaches, but wherein many faculty would benefit from a structured armature of big data support offerings organized around specific tasks or workflows. Finally, it is becoming increasingly apparent that research into and around big data, as with many academic arenas today, might be best situated in geographically distributed context, not constrained to the limits of any single lab or university, but instead designed and implemented in such a way as to draw on resources – including experts from other universities, countries, and research groups – from around the world. Reducing the barriers to such collaboration, especially around data sharing, access, and transfer, would go a long way towards improving the efficacy and impact of big data research at CWRU.

In lieu of providing recommendations, this investigative study and resulting report offers CWRU important information regarding the need, the feasibility, and the next steps for supporting big data research and scholars at CWRU. This information should be shared with extant committees and groups around the university positioned to provide recommendations and solutions upon which CWRU leaders can decide and/or act. The CWRU Research Data Management Group, consisting of members from the University Libraries, U[Tech], University Compliance Program, the Office of Research and Technology Management, and the School of Medicine would be the primary example of one such group. This report is designed to inform the RDM Group and/or

others like it who have the authority to discuss and decide on a path forward as it relates to strategizing, leading, planning, launching, and supporting big data resources for CWRU. Authors of this report encourage those groups to read this report and reach out to discuss potential priorities and agendas for this work.

Appendix A: Research Protocol

Identifying Information

Project title: Supporting Big Data Research

Source of funding: Internal

Research site: Case Western Reserve University

Research Purpose

This study is an exploratory examination of the research practices of faculty and research staff in a variety of humanities, social science, and STEM fields who utilize data science or "big data" methodologies. The goal of the study is to understand researchers' processes in working with big data toward developing resources and services at [name of your institution] to support them in their work. The study contributes to the wider fields of library and information studies and data science, within the context of the evolving relationship between libraries and data science research support.[5]

Research Design

Participants will engage in a one-on-one semi-structured interview with an investigator listed in this protocol. The interviews will be approximately sixty minutes in length and will be conducted either in person or remotely [*modify as appropriate to your IRB's COVID-19 requirements*] via telephone or Zoom [*if your campus or IRB has a preferred video conferencing app, you may substitute it for Zoom here and below*], adhering to the [*name of your university*] guidance on in-person data collection at the time of the interviews. If interviews are conducted in person, they will take place in a private space such as the participant's or interviewer's office on [*name of your institution*] campus.

The collected data will be analyzed using grounded theory methodology, as per Strauss and Corbin.[6] As such, there will be no pre-existing codes; rather, a coding structure will be developed by investigators listed on this protocol in the process of reading through the data. During coding and analysis, attention will be focused on what the informants identify as their research support needs in order to develop ideas for improving library services.

---

[5] M. Burton and L. Lyon, "Data Science in Libraries," *Bulletin of the Association for Information Science and Technology*, 43:4 (April/May 2017), 33-35; M. Burton, L. Lyon, C. Erdmann, and B. Tijerina, "Shifting to Data Savvy: The Future of Data Science In Libraries," 2018, http://d-scholarship.pitt.edu/id/eprint/33891; D. Maxell, H. Norton and J. Wu, "The Data Science Opportunity: Crafting a Holistic Strategy," *Journal of Library Administration*, 58:2 (2018), 111-127; "Research Library Issues, no. 298 (2019): The Data Science Revolution," https://doi.org/10.29242/rli.298; Jeffrey Oliver, "Data Science Support at the Academic Library," *Journal of Library Information* 59:3 (2019), 241-57.

[6] A. Strauss and J. Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (Los Angeles, 2014).

The study at [name of your institution] is connected to a suite of parallel studies being developed locally at other higher education institutions. Ithaka S+R, a not-for-profit research and consulting organization that helps the academic, cultural, and publishing communities, has been hired by the researchers to coordinate this parallel effort and to provide guidance on research methodology and data analysis. The research project as outlined here will be implemented exclusively by the investigators listed on this protocol. The anonymized aggregated data and analysis will also be used towards a comprehensive report written and made publically available by Ithaka S+R. Ithaka S+R will have no access to the research subjects or their personal information. Ithaka S+R will only have access to de-identified interview transcripts and de-identified metadata about the transcripts, not the audio recordings.

Participant Selection

The subject population will consist of approximately fifteen researchers (aged at least 21 years old) who conduct data science research at [*name of your institution*], including tenured and tenure-track faculty, postdoctoral scholars, and staff researchers. Recruitment will consist of personalized email invitations sent directly by the investigators listed on this protocol to researchers at [*name of your institution*]. See [*appendix name and number*] for the text of the recruitment email and recruitment follow-up email. Participants will be selected purposively in order to capture the breadth in data science research at [*name of your institution*].

Baker and Edwards highlight that qualitative researchers should consider both methodology (purpose of the research) and practical issues (time available, intended audience) when determining the sample size of an interview-based study.[7] Because the goal of the project is to generate insights that can be used to inform and improve library services at [*name of your institution*], the project is designed to be exploratory, small-scale and grounded in approach.[8] This study does not purport to be statistically representative nor are the recommendations meant to be prescriptive; rather, the report and its recommendations are intended to be suggestive of areas for further investigation. The exact number of interviews for the sample was informed by Guest's, Bunce's and Johnson's research demonstrating that data saturation can be achieved at the point of about twelve qualitative interviews, as well as Creswell's suggestion that fifteen to twenty interviews be conducted when utilizing a grounded theory approach to qualitative analysis.[9]

Risks and Benefits

There are no known risks associated with participating in this study. Subjects may experience benefits in the form of increased insight and awareness into their own research practices and needs.

---

[7] S.E. Baker and R. Edwards, "How Many Qualitative Interviews Is Enough?" *National Center for Research Methods*, discussion paper, 2012, accessed Mar. 11, 2019, http://eprints.ncrm.ac.uk/2273/.

[8] Strauss and Corbin, *Basics of Qualitative Research*.

[9] G. Guest, A. Bunce and L. Johnson, "How Many Interviews Are Enough? An Experiment with Data Saturation and Variability," *Field Methods* 18 (2006): 59-82; J.W. Creswell, *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research* (Upper Saddle River, NJ, 2002); J.W. Creswell, *Qualitative Inquiry and Research Method: Choosing among Five Approaches*, 2nd edn. (Thousand Oaks, CA, 2007).

Compensation

Subjects will not be offered compensation for participating in the study.

Confidentiality

Interviews will be recorded and stored as digital audio files by the principal investigator(s) in a non-networked folder on a password protected computer. Interviews recorded using the Zoom audio recording feature will be immediately downloaded, stored as specified above, and deleted from any cloud-based accounts. Audio recordings will be transcribed by the investigator(s) listed on this protocol and/or a third party transcription vendor bound by a non-disclosure agreement. Audio recording files will be destroyed immediately following transcription. Pseudonyms will be immediately applied to the interview transcripts and the metadata associated with the transcripts. Public reports of the research findings will invoke the participants by pseudonym and not provide demographic or contextual information that could be used to re-identify the participants.

[*If using a written informed consent form, include:* Participants will sign informed consent forms, either in person or remotely via email [*modify as appropriate to your IRB's COVID-19 requirements*], but these forms will in no way be linked to the collected data because there will be no key that corresponds the participants to their pseudonyms. Informed consent forms will be stored as paper copies in a locked file cabinet only accessible to the investigator(s) and/or as digital files by the investigator(s) in a non-networked folder on a password protected computer. The informed consent forms will be destroyed [*insert the time period required by your institution for destroying these records*] following the completion of the research project.]

[*If using a verbal consent process, include:* Verbal consent will be obtained in lieu of written consent to decrease the risk of breach of confidentiality. In order to document consent, [*insert the procedure, as outlined by your institution's IRB, for documenting verbal consent processes and ensuring that this documentation will conform to the confidentiality expectations at your institution*]. Documentation pertaining to this process will be destroyed [*insert the time period required by your institution for destroying these records*] following the completion of the research project.]

Informed Consent

Informed consent for the project will be sought in [*verbal or written form*]. See [*appendix name and number*] for the documentation that will be provided to participants as part of this process.

Dissemination

The results of the research will be publicly disseminated, such as through conference presentations, scholarly articles and as part of publicly available reports published online through [*insert name of the institutional website or repository where you will be uploading your local report*] and the Ithaka S+R

website. The Ithaka S+R report will be issued using a creative commons license which also enable it to be deposited in [*name of your institution's institutional repository*] as long as Ithaka S+R can be attributed.

Appendix B: Ithaka S+R Semi-Structured Interview Guide

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

Introduction

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

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

Working with Data

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

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

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

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

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

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

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

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

Research Communication

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

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

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

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

Helllo [name here],

The [Study Institution] is 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. We are reaching out to staff and faculty who have big data needs or support those using and managing big data so that we can gain an understanding of [Study Institution's] big data needs. Would you be willing to participate in a brief, 30-minute interview to share your unique experiences and perspective? If so, I will let [Study Team Member #1] know so that he/she can reach out to schedule time with you during the month of February or early March. [Study Team Member #1] will conduct these interviews in collaboration with [Study Team Member #2].

Our local [Study Institution] 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 [Study Institution] will also be included in a landmark capstone report by Ithaka S+R and will be essential for [Study Institution] 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,

[Study Sponsor]

Appendix D: Project types and descriptions

- Information technologists (IT) and developers working on COVID-19 surveillance for local hospitals in order to assist researchers with how they store big data coming from surveillance methods. This work includes assisting researchers with GPS camera setups to capture spatial views, spatial databases setup for continued COVID-19 instance monitoring, developing tools that correspond spatial views with narratives. IT also supported researchers in developing clustering methodologies to identify COVID-19 “hot spots” and with their ongoing data processing needs.
- Faculty capturing data around lifetime degradation of outdoor exposed technology, primarily solar panels by constructing a sun farm and gathering data from 122 power plants.
- Researchers performing big data analytics on employee wellness programs to improve biomarkers for those programs.
- Faculty and researchers working with power plant owners and utility companies to make modules or make the things that go into modules.
- Researchers studying genome sequencing in order to understand genetic polymorphism or base sphere changes. Big data results from performing multiple laboratory measurements of patient populations.
- Information technologists establish individualized compute power for researchers’ big data projects and supporting client applications and cloud computing for big data projects.
- Information technologists ensure that researchers have the technology resources needed to support their big data projects especially in the areas of high performance computing, hardware and software.
- Information technologists addressing security needs of big data projects at CWRU and consulting on their use of data, how they distort it, as well as who may or should access it.
- Information technologists advising on storage and access solutions with both internal and external partners.
- Researchers collect raw data from sensors in MRI scanners, taking raw scanner data and converting them quickly to high resolution images (10GB per second, 32 million pixels per data set, thousands of images per hour, 20 milliseconds per minute).
- Researchers generate and collect clinical data in order to capture images of the brain and use the images as patient diagnostic tools.
- Researchers working dealing with deep learning and machine learning.
- Researchers working with large medical transport data sets and big data science methods to study helicopter and transportation between hospitals across the US
- Researchers developing cancer data science research in order to look for differences in genomes that will lead to understanding of how different regions of a genome organize

and execute and their function in health and disease.

- Information technologists supporting and consulting research centers and departments to use and maintain a software application containing large quantities of clinical trial data (5,000 to 6,000 clinical trials in the system) that is able to download multiple formats of the data being collected.
- Researchers working in computational neuroscience and mathematical cell biology to conduct laboratory experiments and record video data sets.
- A coder working with international partners on big data projects in the areas of linguistics, cognitive and multimodal communication as well as neuroscience.