Introduction

The University of Illinois at Urbana-Champaign was one of twenty-one institutions (Appendix I) participating in the Supporting Big Data Research study (Big Data Study). This study was organized by Ithaka S+R, a not-for-profit research and consulting service for libraries and academia. This study specifically focused on exploring the role of big data and data science within the context of current research, collection, analysis, research communication, training and support. Based on the strength of the data service programs at the University of Illinois, the University Library was invited to participate in the Big Data Study.

Prior to interviewing participants, we sought to understand the definition of big data from the stakeholders’ and researchers’ perspectives. The overarching goal was to understand what big data means, whether its depiction is similar between various disciplines and what lessons can be learned from these answers. These statements below from the stakeholders we initially interviewed are their definitions and descriptions:

“Big data is a dataset where you can do multiple, longitudinal, qualitative analysis, changes in behavior.”

“Big Data is anything that you cannot do on your own personal machine and when you need more computing power than personal computer power.”

“I think everyone’s kind of exhausted by the term big data. There is a lot of different ways of defining big.”
These depictions are a small representative sample of stakeholders’ perspectives but were also used as a guiding framework. It also indicates that the research conducted within academic institutions is varied and dependent upon the research community and context. This study reveals the researchers’ perception of its use and application as well as lessons to be learned thematically.

The report is structured around four areas of focus: role of data science, working with big data, challenges, and opportunities to expand or create library services. During the project, the authors gained some valuable insights of research perspectives of collecting and analyzing data, data sharing, documentation, referrals, and training.

Methods

As part of the larger, multi-institutional Ithaka S+R Big Data project, this study utilized some materials provided centrally. Ithaka S+R provided an Institutional Review Board (IRB) guide that included information and language that the authors modified and adjusted to complete our IRB application for the University of Illinois at Urbana Champaign. Ithaka S+R also provided the semi-structured interview questions (Appendix II), so our questions were consistent with those asked at the other participating institutions.

Prior to conducting the study, we met with three of our university’s academic stakeholders. These individuals’ positionality and research provided a diverse perspective of big data, needs, challenges and support services campus wide. This study’s target population was researchers or faculty who use or work with big data in their current or previous research. Our academic stakeholders provided an initial starting point for potential research participants, and we were also familiar with the research conducted by possible participants. We identified thirty-eight potential interviewees and compiled in a spreadsheet their basic information, including name, email address, department, and affiliations.

Our main selection criteria included individuals that worked with or used big data in their research, established relationships with librarians, and researchers/faculty outside of the University Library. While the larger Ithaka S+R study did not require a sample with a mix of faculty ranks or gender, we contacted a variety of individuals to try to gather diverse perspectives. Fifteen faculty and researchers agreed to participate, which included representatives from science, social science and humanities backgrounds.

We started recruiting participants and scheduling interviews in September 2020. Initial response was slow due to most participants working remotely due to the Covid-19 pandemic. The interviews were conducted from fall 2020 to early spring in 2021 using Zoom, and they were recorded (audio only) to allow transcripts to be produced. An external transcription service provided the fifteen transcripts as Microsoft Word documents. Upon receipt, we each reviewed the transcripts for our interviews to redact any identifiable information (e.g., names, distinctive publication titles) and to fill any gaps. After anonymizing the transcripts, we read them again and coded this information thematically.

In the analysis of the transcripts, we followed a constructivist approach to Grounded Theory, mainly the principles of saturation and coding. We selected a small set of transcripts, read them individually, and came together to discuss significant words and terms. We repeated the same process several times until the key themes emerged. To mark codes and themes in the transcripts, we used conditional formatting and word search features in Microsoft Word and Excel to isolate comments in the transcripts which contained words
such as need, challenge, concern, struggle, battle, effort, easy/hard/difficult, gap, lack, issue, and problem. Similarly, we looked for words such as help(ful), useful, interest, support, learn, train, teach, educate, and solution. Finally, we searched for question marks in the responses in the transcripts, since these often signaled places where the researcher was providing a hypothetical example and in some cases revealed an area where Library support, or support of some sort, may be welcomed.

We looked for opportunities for the Library to support these researchers. Because many of the researchers we interviewed are highly experienced in working with large datasets, few of them mentioned areas of need for themselves which they felt could be addressed by library services. This may be because they do not recognize a part of their workflow as in need of improvement, because they are unaware of the potential for support from the Library, or because the semi-structured interview questions were limited in terms of collecting data about this aspect of big data research. To address this issue and to supplement information that could be “unsaid,” we also utilized deep reading and reflection as ways to identify themes.

As a result, we found more themes than could be addressed in this initial, local report. We focused on major themes that each had significant sub-elements and that were more likely to have a connection to libraries and library services. We decided on four main themes: role of data science, working with data, challenges, and the opportunities to expand or create library services which are addressed in the next sections.

Themes

Role of Big Data or Data Science in Research
Participants were asked about the role of big data or data science in their research. This study establishes that the term big data is used frequently, and its interpretation and application is overarching and wide. The following comments are representative:

“...The data that we use we’ll typically refer to as cross-section libraries and these describe the probability of certain nuclear interactions over the full set of thousands of radio isotopes and many axes.”

“...we’re looking at tools and mechanisms by which people access and synthesize information ...

“...traditional way of doing research was common for about a century...in the last 20 years, there are these sensors and digital tools that are becoming more and more available. Many of us are trying to figure out what is important, what is not, how to store and how to manage the data.”

Traditionally, the researcher is known as the primary investigator and oversees all facets of the research process. All participants in this study conducted collaborative research projects (i.e., multi-institutional, partners from various research units, departments, and disciplines). This study revealed researchers have various roles in conducting research. For example, in collaborative projects, there are researchers who work on the formation of research questions, disseminating the research and other individuals focus on analysis.

Researchers also discussed their experiences with analytics and applications. R, a statistical programming language, and the computer programming language Python were the most commonly used tools by the faculty and researchers we interviewed. That might have contributed to the fact that RStudio and Jupyter
Notebooks (a popular tool for Python programming) were also popular applications among the participants. Few participants mentioned other programming languages/applications (e.g., Perl, Stata, Excel), specialized software for working with spatial data or software and applications developed specifically for their research.

**Working with Data**

Participants generated data and used secondary datasets interchangeably or dependent upon the type of research project. Participants worked with data available via different types of media. This was particularly important to participants who were in the humanities. Scanning sets of print publication or documents, running object recognition and structuring the machine-readable data were expensive, but a foundation to some participants’ research. There were participants who harvested data via APIs or data scraped from the web. It is worth noting there that while relational database technologies were a common choice among the participants, some participants articulated the use of high performance technologies such as HDF5 (Hierarchical Data Format) or MongoDB and Cassandra (NoSQL database programs).

When it came to storage options, the participants described their practices largely centered on volume, velocity and variety (3 Vs of big data). Almost uniformly, participants utilized common cloud storage (e.g., Amazon Web Services, Box or Dropbox) or university supercomputing services (e.g., National Center for Supercomputing Application, NCSA). Some participants still stored the data, or a version of data that they were working with, on their own desktops or laptops. We also observed a fairly uniform practice with regards to storing code. Most participants store and manage code where they store data or common platforms such as GitHub or GitLab.

Most participants stored data for their own research needs, and they preferred storage options with easy-to-use APIs for them as well as their collaborators or research groups to access and manage data. It was evident that sharing data or code with people beyond collaborators or research groups was not a main interest of the participants. It should not be generalized as if the participants are unwilling to share their data/code. Only a few participants specially talked about depositing data and code in repositories (e.g., Illinois Data Bank) or other open platforms (e.g., Zenodo). Some participants mentioned that their collaborators/research group members were the ones to manage the data/code sharing aspect of their research.

**Challenges**

This study reveals researchers use the word “access” differently when describing challenges they face working with and using data. “Access” was defined in the context of acquiring, finding and permission. For example, a participant noted the problem of accessing data that is not widely available. Whereas other participants noted problems accessing data in large groups due to file access restrictions sensitive data, and security. Historically, sensitive data refers to healthcare data, employees, student records and data with identifiable information. This study expands its meaning because of its potential application of use, licensing restrictions, copyright, and privatization. Each of these issues impacts sharing, transparency, and openness. The following comments are representative:

“...a lot of the data we use within nuclear engineering can be used to simulate nuclear weapons; this creates sharing barriers in relationship to open source science and security compliance of sensitive data.”

“...the unstructured text is copyrighted, and sensitive data...we can’t just share/show the raw text. This creates a challenge navigating through open science standards.”
“The harder thing is the contract or licensing law. On our campus, research talks with scholars from our campus and beyond is a common occurrence. Yet licensing, contract law and ethics restrict us from allowing others to view our data and analyze it.”

“...The private sector has a huge impact within our work. Yet non-disclosure agreements, privacy agreements within our work creates barriers in advancing research.”

**Data Documentation**

The need for data documentation training and assistance was the most common need expressed which relates to data management. The following comments are representative:

“...I wish they offered a documentation course for research or coding documentation at the undergraduate level.”

“Coding is a process that evolves incrementally and in some cases we don’t document every single step within the process. Because of this reality it becomes hard to document our research processes.”

“We have to be able to find the data, so organization is extremely important.”

“...I would like to learn how to efficiently manage and clean data.”

**Data Discovery**

Researchers generally talked about locating and acquiring the data they need for their research through the assistance of collaborators or other partners; through open data repositories or sites on the web; by exploring Library collections; or by purchasing the data themselves. This was not surprising, as these researchers are all experts in their disciplines and many are well-established in their areas of research, so they are highly practiced in locating data for their research.

However, there were multiple comments which indicated that researchers would appreciate better discovery tools, to understand what datasets are available to them, either through the Library or through open repositories. This knowledge would impact their approach to a research project by helping them to understand which datasets may be readily available to them, reducing the time to acquire the data and begin analysis and eliminating the need to fund a purchase of data. Knowledge of readily available datasets could also give researchers options for testing code under development or to generate and test ideas before investing more time and resources in a full research project. Representative comments in this area follow:

“But it would be great to have something...where it’s just one place you could go to find all the data sets that are available to people at the university in a one-stop shop, rather than having to go and search around and find them in a piecemeal manner.”

“And another thing that I’d like to see is actually data discovery, having some kind of hubs or indexes to go and find data for, like I have idea to do this project about, I don’t know, air quality. How do I find that data? Who do I need to contact? Where does that data live? And all that...If there was something like a service, maybe not a central hub but a collection of hubs that can provide data or a sample of data so people can look at those, play with those, and then develop ideas for leveraging that data for their projects, that would be great.”
“...it’d be really great for a lot of faculty if they could have help from librarians about the, if they have some or if they do have some instruction sessions of how to deal with this, the data website in order to create data table or something like that. That would be really helpful I think.”

“I think that definitely anything to cut down on that huge amount of investment you have to do just to find the data set and clean up the data set would be valuable in the sense that after you have those steps done, you're actually doing the interesting scholarly work.”

**Training**

One of the most common themes across all the interviews was the need for training, in a handful of cases for the researchers themselves but more often for new members of their research teams. The researchers who are seeking solutions for this challenge presented the need in these types of comments:

“It’d be really awesome to have a semi-regular boot camp-type training opportunity with an individual on campus that is available as a resource perhaps all the time. Because the thing about learning these methods with analyzing big data, it’s hard to learn those things unless you have a problem at hand on which to apply them. But if you have a problem on hand with which to apply those tools, it’s already kind of too late to learn the tools.”

“So I notice that in many workshops they will provide very simple tutorials and everybody was happy with that, but when we really start working with big data, then there will be all kinds of discrepancies and errors. So everybody has to learn it again from trials and errors. So if we can have more in-depth tutorials that would be helpful.”

“I think that would be really helpful if there was, yeah your school or other units on campus could offer humanists workshops, basic workshops on thinking about big data dimensions to their research projects. That would be really good.”

Some of the training areas most commonly mentioned by these and other researchers included AWS, Google Cloud, R, Python, and GitHub.

Other researchers provided some advice for training methods and approaches which worked for them. Some researchers mentioned the training they received through Software Carpentry workshops and alluded to other training they had received on campus or through campus resources. Approaches for self-teaching, for both the researcher and for the onboarding of new members of the research team, included Coursera and Udemy courses, YouTube videos, and O’Reilly guides. Speaking of the O’Reilly guides in particular, one researcher commented “in a way they’re better than any training because if you get something you can move on, right. Or if there's something specific you want to focus on you can just go straight to it, instead of learning things that have no relevance to you.”

Participants also noted other challenges which include but are not limited to the time it takes to work with big data, the need for documentation, the lack of data standardization, the need for organization protocols when working in large groups. Comparatively, some of these challenges are similar to previous studies highlighting issues researchers and faculty experience when managing data. While managing research data is not the sole focus of this study, it is important to note the commonalities and areas of opportunity for subject specialists, librarians and researchers to provide solutions, services or training.
Opportunities to Expand or Create Library Services

Many of the researchers did mention approaches or services that they thought might be helpful or useful for someone, if not personally for themselves, whether provided through the Library or another unit. At times, these suggestions seemed to be things that the researchers may have been long aware of, and in other cases, it seemed that the question itself prompted the researcher to realize an area of need. In only a few cases did researchers identify potential areas of need for themselves as individuals; most often, the needs pertained to support for their graduate students or for new faculty in their department or research group. Some researchers spoke more generally of improvements that might be helpful to researchers like themselves, without expressing it as a need they themselves have.

Data Sharing and Documentation

There was evidence that data sharing cultures have been established or are emerging in the disciplines of many of the researchers to whom we spoke, and we view this as an encouraging sign. Several researchers spoke about their personal and professional motivations to share the data they collect, when it is possible to do so, and some spoke about how they feel that this perspective draws like-minded graduate students and faculty to their research group. However, this data sharing mindset was not present among all the researchers to whom we spoke, for reasons that related to their personal choices as well as to the circumstances under which they collect their data. Some of the decision factors were the way the data were obtained and whether the researcher had the right (legally or morally) to share it. Other reasons included the fear of unattributed use by other researchers, and the decision to fully utilize the data for their own research before making it available to others.

While the Research Data Service (RDS) is an established service within the Library and is supported by the OVCR, and similar services are available elsewhere in the Library, the visibility of this type of Library support still may not be high among long-time faculty members and researchers, who in turn would not be able to refer the new members of the research team to the Library if they themselves do not know about this type of service. A few interviewees did mention interactions with the RDS in the past, as well as with other research data services in the University Library, but most did not. We consider this to be an opportunity for the University Library to expand its marketing efforts to enhance the visibility of these research data services among populations working with large datasets, particularly new arrivals to the University.

Comments revealed a more specific need in which university library services could partner to create training that provides instruction to researchers and research groups about the best practices of data documentation. Although workshops and training in these areas currently exist, statements by the researchers indicate this knowledge is not widely known and partnering with other units within the library could be helpful. This addition may create relationships to bolster conversations and practices around data sharing.

Data Discovery

Researchers’ comments represent an opportunity for the Library to expand upon past and current efforts to provide better and more comprehensive data discovery wayfinders and tools for researchers across disciplines. Some initiatives in the Library have attempted to address this need. As a basic starting point, it has long been possible for researchers to search within the Library catalog to locate datasets. There are also a wide variety of LibGuides and webpages to guide researchers to our dataset subscription sources and, in
some cases, to owned datasets in a variety of media formats. The Illinois Data Bank enables users to search for and access datasets created by UIUC affiliates. Finally, the advanced search feature for Datasets in Easy Search searches across eleven science-based dataset repositories from around the world. So, the need may not be for more wayfinders, but perhaps some coordination between or compilation of the existing wayfinders so that this information is made more discoverable. We recommend that efforts toward wider discovery of datasets begin with wider promotion and visibility for the discovery tools we are currently maintaining.

Referrals to Training and Learning Resources
We are not proposing that the University Library should develop the capacity to provide the type of training the researchers mentioned. However, providing visibility for and access to learning opportunities on campus and training tools through the University Library is very compatible with existing services. Several Library units, in particular the Scholarly Commons, have long-standing partnerships with units on campus which offer the type of training needed, and by providing increased visibility for those services the University Library can support researchers across campus, particularly those researchers new to UIUC. In addition, the University Library has long provided access to authoritative manuals, including the O’Reilly guides, and there is an opportunity to bring more visibility to those parts of our collection so that researchers who would like to learn new skills or improve existing skills can do so on their own time and with a degree of anonymity.

Conclusions
In all of these areas—data documentation and sharing, data discovery, and support for training—the University Library has already established activities upon which new or expanded service efforts could be built. As such, the investment would be relatively minimal, not requiring new personnel, resources, or funding, but rather requiring a strategic use of our time, energy and attention, to create a positive impact on big data users across many disciplines at UIUC.
Appendix I

Supporting Big Data Research Project Partners

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
University of Virginia
University of Wisconsin-Madison
Appendix II

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?