Big Data Research at the University of Colorado Boulder

Emily Dommermuth, Rebecca Kuglitsch, Andy Monaghan, and Nickoal Eichmann-Kalwara

Executive Summary

Background: Big data, defined as having high volume, complexity or velocity, have the potential to greatly accelerate research discovery. Such data can be challenging to work with and require research support and training to address technical and ethical challenges surrounding big data collection, analysis, and publication.

Methods: The present study was conducted via a series of semi-structured interviews to assess big data methodologies employed by CU Boulder researchers across a broad sample of disciplines, with the goal of illuminating how they conduct their research; identifying challenges and needs; and providing recommendations for addressing them.

Findings: Key results and conclusions from the study indicate: gaps in awareness of existing big data services provided by CU Boulder; open questions surrounding big data ethics, security and privacy issues; a need for clarity on how to attribute credit for big data research; and a preference for a variety of training options to support big data research.

Recommendations:

- Evaluation of current access to existing research infrastructure at CU Boulder across departments and disciplines, including recommendations for how inequities could be addressed.
- Development of big data training curriculum, particularly for big data ethics, privacy and security, through a variety of channels (e.g., documentation and context-specific consultations for specific big data services, more general course-based curriculum).
- Consider how to address the complexity and dynamic nature of big data in the IRB process in a manner that fully and reasonably considers the ethical, security and privacy implications of a given big data research project.

- Creation of CU Boulder guidelines for attributing credit to the myriad contributors in big data research projects, and considering the sometimes unconventional contributions, with the goal of helping departments develop clear policies and incentives for researchers performing big data research.
- Development of marketing and outreach strategies to increase awareness of existing and forthcoming big data research support services at CU Boulder, in a manner that promotes equitable access to services across disciplines.
- Assessment of staffing gaps and staff-training needs in order to support big data curriculum and services.
- Periodic evaluation of emerging trends and needs in big data research, in order to adjust strategies and services appropriately to ensure CU Boulder is providing state-of-the-science support and infrastructure.

An optimal way of addressing the complex questions above may be to establish a steering committee composed of a broad range of CU Boulder (and possibly external) stakeholders and decision makers.

Introduction

As technology advances, the amount, frequency and complexity of digital data is increasing rapidly (Vassakis et al. 2018). For example, telescopes and satellites now regularly produce tens of terabytes of data daily, the social media platform Twitter yields 500 million "tweets" per day, and the availability of complex genomic data doubles at timescales of less than a year (Langmead et al. 2018, Amani et al. 2020, Khalil et al. 2021, Whitney 2022). These and other big data have enormous potential to advance scientific discovery, and thus research methodologies are evolving to facilitate big data analysis (Guo 2015).

To date no systematic effort has been conducted to understand the scope and nature of big data research at the University of Colorado Boulder (CU Boulder), a large, decentralized Carnegie classified "R1" research institution. The present study explores big data methodologies employed by CU Boulder researchers across a broad sample of disciplines. This study is conducted in conjunction with a larger set of parallel studies conducted across numerous U.S. research institutions, coordinated by ITHAKA S+R. The objective is to better understand researcher knowledge, attitudes, practices, challenges, and needs in order to inform the development of resources, services, and support for big data at CU Boulder. Qualitative, semi-structured interview research is by nature exploratory and nuanced rather than generalizable. The findings highlight themes and observations seen across our limited sample size and invite further study. The study was conducted via a series of semi-structured interviews with researchers, described in the *Methods* section. Results are presented in the *Findings* section. In the *Summary and Recommendations* section, key themes and issues are synthesized, and recommendations for supporting big data research at CU Boulder are made.

Methods

Interview Design

A semi-structured interview guide (Appendix A) developed by ITHAKA S+R was conducted with each participant. Semi-structured interviews contain a list of predefined questions but allow for flexibility in exploring topics as they arise during the interview (Longhurst 2003). The interview questions were designed to gain insight on participants' knowledge, attitudes and practices with respect to big data research. For the purposes of this study, big data is defined as data that has high volume (i.e., “large” datasets as perceived by the researcher), complexity (i.e., numerous different types of data are involved), or velocity (data are created at a rapid pace e.g,. social media posts) (Lycett 2013). The study protocol was submitted to the CU Boulder Institutional Review Board and approved with “exempt” status (Protocol # 20-0344).

Conducting Interviews

A list of potential candidates was identified by study team members, and by suggestions from persons in leadership roles at the University, with the objective of selecting a broad representation of persons involved in big data research at CU Boulder. Research team members then emailed the candidates describing the study and interview process and requesting their participation. If they accepted, a follow-up email was sent to schedule a sixty-minute interview. A total of 11 candidates accepted the invitation to participate in interviews. Their disciplines included psychology and neuroscience; sociology; media, communication and information science; history; computer science; political science; aerospace engineering; and geosciences.

Interviews were conducted between October 2020 and January 2021 via Zoom (n=10) or Google Hangouts (n=1). Consent to participate was received either in writing by email ahead of the interview, or verbally at the beginning of the interview. After the interviews, team-members transcribed the interviews from video media to text documents that were then de-identified. Transcripts were reviewed by the respective interviewer for accuracy. Video files of the interviews were discarded to eliminate identifiable information.

Coding

The study employed grounded theory methodology for qualitative content analysis (Strauss and Corbin 1990), such that a team of researchers developed themes based on initial read-throughs of the 11 interviews by all team members. Each team member wrote down key themes from their perspective, then the team worked together to synthesize this information into key overarching themes and sub-themes. These themes were defined in a codebook (Appendix B). The four overarching themes identified include: 1) practices, 2) challenges and needs, 3) learning and development, and 4) unresolved questions, with sub-codes addressing topics around data access, curation, collaboration, and ethics.

Next, two team members were assigned to code each survey. This was a "semi-random" process by which each team member coded approximately the same amount of interview content. For interviews containing discipline-specific jargon, at least one team member with subject-matter knowledge was assigned. Each team member read their assigned transcripts and tagged relevant excerpts with their associated codes. After the initial coding was completed, a second team member reviewed the interview and codes, and the two team members discussed and reconciled the codes. The coding and analysis was conducted within the software package Dedoose, version 9.0.17 (Dedoose 2021). Qualitative, semi-structured interview research is by nature exploratory and nuanced rather than generalizable. The findings highlight themes and observations seen across our limited sample size and invite further study.

Findings

Practices and Challenges

Several themes emerged from research discussions of practices and challenges and needs: researchers use and obtain widely diverse data types, synthesizing data from disparate sources to answer new questions; researchers see emerging practices around credit, tenure, and publication; collaboration grounds work with big data, whether to access data or to bring new skills to a project; tools and infrastructure are vital to big data work but often present challenges; and big data researchers have concerns around privacy, security and ethics.

Using heterogeneous data to answer new questions

A consistent theme was getting different kinds of data in order to answer new questions. Researchers used a wide variety of data: primary and secondary data were often integrated, or heterogeneous data were used. Primary data were collected from sources ranging from imaging data, app data "that's being collected from a person's cell phone, things like GPS, accelerometer data, texts, calls, taps, swipes, those kinds of metrics," APIs, and web scraping. Many interviewees mentioned extremely heterogeneous data: "[The] unique thing is that data sets are very heterogeneous, heterogeneous in terms of type of data and also, also more in a sense [of] data format, locations and in times." Combining imaging data with survey and app based data, or field data with remote sensing, or multimedia data were all mentioned.

Interviewees reused data because it was expensive to gather new data, or beyond the scope of an individual researcher to collect. A researcher who primarily worked from existing data, said “It costs, you know, tens or hundreds of millions of dollars to get this data collected and so we use data sets that have already done that, and develop methods using that.” Data are often shared and occasionally purchased. Data may be cost-free but not necessarily publicly available, either requiring an application, for example “I do have to apply for access and go through some rigmarole and get IRB approval” or credibility within informal networks to receive data from colleagues. Others work with fully free and open data. Researchers create datasets using existing information sources. Web scraping text data from open sources was commonly mentioned, with a range of satisfaction in the results: modern US legal code as textual and tabular data was easy; historical documentation, less so. Ease of access to data can shape research programs: “I developed research programs in an area where I can have access to data. So for example, DOD [US Department of Defense] data sets that I'm interested in. But, [when] I don't have access to [data] I move away from that kind of results.” This suggests that data availability may be driving the nature of research questions.

For some researchers, especially those doing simulation research or building analysis tools, the content of the data is less important than having data that is similar to a *kind* of content:

> “Sometimes we have collaborators who come to us and are like we want a tool that's going to allow us to do X, Y and Z, but we can't give you the data yet …. So we can either look at existing open source solutions and stream that data in as example dummy data, or we can build solutions and deploy them with stakeholders and essentially cross our fingers and hope it works once things get out in the field.”

This highlights how code, tools, programs or similar may be the research output some big data researchers are working toward. For this type of researcher, the content of the data does not matter; they just need a data set that is structurally similar so that they can develop an application.

Emerging practices for publication, credit, and tenure

Although researcher’s traditional publication habits generally followed unsurprising disciplinary norms, some interviewees reported using these more established formats to report new kinds of work or that they struggled to merge new kinds of work with these publications. Some still feel obliged to center traditional forms as a way of making research official, e.g. using traditional article structures to argue for sharing project histories, or publishing in data journals. GitHub and Open Science Framework (OSF) were explicitly mentioned as supplements to traditional publishing and were widely trusted platforms for sharing research objects and documentation. In addition to data journals, researchers might simply post data freely to the web on personal sites, publish to GitHub, or archive in repositories (sometimes chosen because of grant requirements).

Researchers published data and code for a variety of reasons, and at least one felt they were “more inclined than the median researcher in my discipline to share code and data.” Some shared data based on personal values, including reciprocity, in recognition of public funding, and openness as a foundation for science. Academic and commercial values could conflict, however, as when sharing academic research prevented filing for a patent. Interviewees also mentioned disciplinary values and ethos.

When researchers did *not* publish data, the range of explanations included:

- A perception that the data was of internal interest only.
- Research conducted “on the fly” without enough structure to share.
- External constraints from funders or research bodies, short or long term.
- A sense that data could not be shared fully with existing technology and projects.

However, these reasons were often presented as exceptions to normal practices of sharing.

Publishing code occurred, but does not yet seem to be as widespread a norm. One researcher, working on an NSF-funded software tool said “[I]t's unusual in our community, in my discipline, it’s at least new for our community, to share software. [I]t takes time to develop such a resource, and then people tend to keep it to themselves.” Both newness of habit and the effort sunk into developing code were a barrier. Still, one researcher mentioned valuing the reciprocity of working off others’ code as a spur to publish code and at least one researcher simply published code as a matter of course.

The rewards, incentives, and credit for sharing data and code was not consistent for researchers. Some felt there was a clear reputational reward for tenure and promotion: “I think it gave me a reputation of being willing to share my data and help other scholars, made a difference, really, but I would say that's on the incentives," although the interviewee thought that not sharing data would have no effect on a case, either. Another humanities researcher felt best practices were developing, but were not yet codified and effort did not match reward for tenure, promotion, or annual evaluation. Formal credit for projects and publications is also an active discussion. These issues are not always new or exclusive to new types of research and publishing, with one interviewer likening the problem to authorship dilemmas in traditional academic publishing. Published articles on project histories were mentioned as a solution, but there was still some resistance to formalized agreements about credit, and concern about how credit would be apportioned when datasets were reused.

In addition to career rewards, some researchers felt there were research-community-wide rewards to sharing data, where the data publishing ecosystem as a whole improved. Others, however, saw a conversation coalescing but could identify no impact other than grant requirements yet. One researcher expressed concern about punitive attitudes developing around publishing data, with fewer reward incentives and more punishment for not publishing data, whether by researchers refusing to review or use research that did not publish data.

Interviewees highlighted a number of ethical awarenesses and concerns related to data sharing. Overall, interviewees valued openly sharing their data, especially for research funded by the public. Still, they expressed concerns around open data policies, including inconsistent policies between different funders or data providers, inconsistent data sharing policies impacting who can work with certain data, and data publishing policies excluding those who work with industry partners and private data. While there was generally a willingness to share data, some also considered situations where data sharing would not be as strongly of interest, such as a high financial or time cost undertaken to get data, or concern about getting credit for data if they are shared.

Collaborating is foundational to big data research

Collaboration recurred in researcher discussions. It might be to facilitate access to data, or to bring knowledge and skills to a project. For some researchers, publication and collaboration can blur, when formally published data or code resulted in collaboration or in publishing open source software. Sharing data informally sometimes bled into collaboration naturally.

For others, collaboration is a way around restrictions on access to data; one researcher who worked with pediatric mental health data described collaboration as a way of sharing data that cannot be published: “[A]s data get bigger and more complicated, it's easier to instead of trying to send data to another institution to basically import the investigator virtually to where they are.” Data that might not be widely published for commercial reasons was also accessed this way. Collaboration to access data might be secondhand, where graduate students take internships to access company data, or a postdoc might be valued for access to data from another university related to their PhD. This type of collaboration is valued but fraught, as it adds authors and carries the risk of losing access to data should a collaboration end.

Collaboration also plays a significant role in fostering new knowledge or paradigms in a project. When expertise is needed from another field, interviewees reach out to graduate students, faculty, and other experts in their networks. One interviewee found that a librarian had the expertise needed to “debug some really deep down in the weeds issues. So it was valuable having another expert to bounce ideas off of and collaborate on working through a bug or an issue. So I really value having those kinds of opportunities to engage with other experts.” Just as interviewees relied on other data experts for support, they were willing to sit down and share their expertise with others who had questions. In one case, an interviewee needed support for hiring programming expertise:

> “that's where I'm handcuffed right, because I have the data, I understand, I can see how it operates but I don't...have those specific skills. So the dependency on programmers, [has] become a challenge because you're into these multi-disciplinary collaborations, where it may be my main project, but it certainly won't be theirs.”

Collaboration is clearly vital to learning and new developments, however one interviewee shared that learning strategies and best practices for collaboration would be of interest.

Collaboration features heavily in interviewees’ research, whether with individuals or large inter-institutional projects. One interviewee thought that “about half the work I do… actually maybe even more… is in collaboration with other institutions.” Others frequently cited work with postdocs and graduate students. Collaboration was not always close, however:

> “Just the students that I work with right now all pursue their own individual projects that I support them in and it's less like, here's a project I'm working on that you're gonna support me on. So it's more of a model [where] they go

> out and find and collect data sets themselves and there's very little in the way of trying to collaboratively share data and code and things like that.”

The nature of interviewee’s collaborations varied across a spectrum of different practices.

Collaboration can shape how data are processed, as one researcher recounted training a postdoc to work on their lab systems and described their analysis systems: “those big systems that I have that are, they're running on MATLAB that I inherited from somebody, they're running on C that I inherited from somebody else, all the way out to IDL and R at various levels of complexity.” This highlights how a patchwork of collaborations over the years may result in a patchwork of systems used by a research group.

Collaboration might answer new questions by bringing disparate disciplines and skills together. Collaborators might include historians and coders, social scientists and geneticists, modelers and field researchers. One interviewee described themselves as “a bridge in many ways to people whose science either requires big data or needs to be set up to produce big data. Kind of both directions,” later saying “so with each one of those categories of peers I overlap with them, but I am different from them.” Interviewees saw collaboration bringing disciplines together as a way to “actually find new ways to address long-standing questions we’ve had in our respective disciplines…really start asking you new questions… to actually tell a stronger social scientific study story if you will, so I harness all this.” This shows how many interviewees view collaboration as key to moving their research forward.

Interviewees also found access to dedicated programmers challenging. Whether interviewees relied on graduate students for programming, campus resources, or hiring programmers with grant funding, they recognized that programmers typically had many projects and priorities, which were not always aligned with the researchers own sense of urgency. Additionally, hiring programmers via grants was challenging because grants come with their own restrictions--for example, one researcher was challenged by having hired an excellent international collaborator they were unsure they could pay with grant funding. Interviewees acknowledged the sustainability issues of dedicated programmers in the university environment, and recognized cost challenges. Disciplinary disparities in access to programmers are likely to exist.

Several collaborations were highlighted as particularly successful. In one case, the collaboration relied on other institutions structures:

> “some institutions have very clear policies, and then it’s easier if [we] make sure that we follow their guidelines, implement that as a part of software that the users, have an easier time to follow their data providers’ guideline and policies so we can integrate...one project is really designed to facilitate collaborations. ...so they try to remove almost all obstacles for collaborators to use our software tools and then make sure that collaborators can use data and then analyze data using our software tools. So that was sort of in a, it's a, sharing or collaborative data sciences is a part of their design of project. That's one extreme case, good practice case.”

This example highlights the value of clear policies, as well as tools that facilitate both the policy and collaboration.

Interviewees used a range of tools to foster collaboration. Git and GitHub were essential tools for many: “We are largely a GitHub shop. And so I've got an account set up for our lab, and that's where we keep all our version control rolling for our tools. And if we have external collaborators, we will add them on. For another interviewee, Jupyter notebook is invaluable:

> “So for me, that notebook is really valuable as an artifact to both revisit, load new data into, share code with collaborators, have all the results live together with the code.”

One interviewee mentioned Globus for getting data to collaborators and an FTP site (set up by Research Computing) for outward facing sharing, though there was a desire for Globus to connect to Google Drive. Both Google Drive and Dropbox were referenced by another interviewee. Interviewees expressed a desire to utilize open source tools but have a reliancy on proprietary platforms that can create problems for collaborations, sharing and access, and research integrity.

Interviewees shared perspectives that reflected conflicting ethos and practices related to the nature of collaboration in big data research. One researcher’s conflict was in the fact that “I don't do any of the research anymore myself. So it's always people in my lab that are actually doing it.” This interviewee described this as a challenge, indicating that collaboration has led to the distribution of labor and disconnection from the processes and daily practices in their research team’s work. Additional job duties, such as accepting an administrative position, may also lead to disconnection from or delay in completing research, such as one interviewee who has not been able to publish a data set since becoming an administrator.

The questions and discourses of who is a collaborator and who is a service provider is interesting and one that could benefit from more study. For example, one interviewee reports relying on technical experts:

> “the imaging center has a data scientist who's on staff, and a data analyst who's on staff and so they both have been incredibly helpful,... they are the ones who are doing that process and troubleshooting. And that's great because then it frees up me and my lab members to be able to focus on the analysis of those data.”

This interviewee clearly valued the contributions others could make to their work, because it allowed them to focus on their area of expertise. Another interviewee also valued the contributions of others, but expressed concerns about how some work with others is seen as a service, rather than a true collaboration:

> “That the people that have the skills to do that, to be able to actually organize, wrangle that data if you will, get it towards useful, don't typically get the props that they deserve. They don't get that respect -- it’s seen as almost sort of like, oh some sort of skilled labor thing. And to me I'm like no, this is - we have to stop thinking that way. This is critical to us doing this science correctly.”

This highlights how various experts may be providing critical contributions to research, but since they are seen as service providers rather than collaborators, they are not getting appropriate credit for their contributions. Issues related to receiving credit in collaboration were noted by another interviewee, who was specifically concerned about trust and transparency in co-authorship. Finally, one interviewee who conducted multidisciplinary work shared a challenge stemming from working with collaborators in different fields with multiple projects, and how this collaborator’s priorities did not align with the interviewees.

Tools and infrastructure

Researchers expressed a variety of challenges that they face when working with big data tools and infrastructure. Infrastructure needs encompass both technical and social infrastructure. Identifying and implementing secure storage and analysis environments is challenging. Another area of technological infrastructure challenge is datasets that can only be used within a service provider’s application or that prevent download unless the dataset is purchased. These fees can be relatively high (e.g. $5K--an amount one researcher noted as ‘nominal’ but which is far from nominal for researchers in disciplines with less financial research support).

Tools are a necessary and valuable aspect of big data research, however they present challenges for researchers. One interviewee in the social sciences has found that collaborators they bring in to work with are not familiar with even the environment that tools they use for their big data research run on, and that a great deal of training and time is needed to bring folks up to speed. Researchers may also work with tools for which they have limited familiarity, and thus they become reliant on the expertise of a collaborator. Another researcher struggles with multiple research projects that all use different core languages:

> “I don't have the luxury of becoming, or deepening, my expertise in any one of them. It's the MATLAB project and I have to figure out how to reset that, and so I dip into MATLAB just enough to do that thing. And then I have to go work on the Python project.”

They go on to describe this as a heavy cognitive strain, with a great deal of re-learning to be done. This interviewee assigned some fault with this issue to funding, and how each project comes with heritage and that there is not funding available to update that heritage aspect of the project to the preferred language.

Another issue related to the tools researchers work with is that the tools often do not integrate with each other. For one interviewee, who works with image based data and whose research output is a website, they would love for the tool that they use to modify their image data could be integrated with the user interface, so that modifications appear on the webpage instantaneously. Even in instances where web-based data tools can be integrated, it generally requires more work and customization, which can make tools more difficult to work with. Highly customized tools can also require ongoing maintenance for healthy functioning.

Even if tools exist that could make working with data easier, data literacy and critical competencies continue to play an imperative role in addressing bias and power in data creation and collection. One interviewee discussed how OCR technology is improving, including for reading historic handwritten textual data, however the difficulty of reading handwriting is not the only issue the interviewee working with this kind of data experiences - the textual data they are working with involves the transliteration by the original authors who were writing down words and names from non-written languages into written English or Spanish of a different time period. So, even when tools might be available to support working with a data source, researchers may be unaware of how their tools create and perpetuate loss of meaning and context due to the nature of the data.

Interviewees shared a mix of awareness of existing social infrastructure, services, and expertise on campus to support data management planning and assistance with data publishing and preservation. Several interviewees discussed instances of being unaware of campus support resources, or were aware and appreciative of some services but didn't know that additional services were available: "So if I had data that I wanted to make available, it would be up to me to figure out how to do that. It would be nice, if there were folks who were there to help us with server space, or the ability for things to exist in perpetuity on a platform somewhere." One researcher requested a service to help publish and host data and another indicated willingness to contribute funds for these services. Maintaining awareness of these existing resources is challenging.

Second, while big data projects and grant awards represent prestige for researchers and the university, the sharing of limited infrastructural resources can slow down their activities and timely completion. This includes lack of project management support and a desire for priority in support queues. "The more projects you take on the less time you have for each individual one. And so the more success you have actually becomes a detriment in the long run because it takes too long to make changes for individual projects because you have to fill out a ticket and wait your turn." Thus there is a need to streamline workflows to ensure researchers can access infrastructural resources efficiently.

Some challenges with infrastructure deal with external data sources and restrictions that do not align with researcher practices and current technologies. "So it's not even very computationally intense but I have to do that. I have to use those betas or slopes from external source[s] with my own internal data and believe it or not, that simple moving of that file.... I could do it right now in two seconds with a very simple command for my laptop but it would potentially violate every risk that we have so to do it properly the administrative paperwork is a nightmare." This indicates there may be a disconnect between infrastructure workflows/operations, and keeping data secure.

Ethics, Privacy and Security

Researchers had challenges related to data privacy, and keeping their data secure. Those most concerned with data privacy and security were the interviewees working with personal health information (PHI) and personally identifiable information (PII), though some of those working with data not considered to be PHI or PII were concerned that there could be unforeseen privacy issues. In addition, some interviewees reported no issues with privacy or security.

The challenges faced by interviewees with private data that need to be kept secure included difficulties in getting their data and being unable to share data, or they believe data sharing takes a great deal of time. Additionally, interviewees reported greater challenges in manipulating and analyzing data while still keeping data secure or confidential, including not being able to do all possible or desired analyses because the infrastructure they need does not have the security required for that data. One interviewee expressed concern that the free tools their graduate students prefer to use to work with data might have security issues, even though the proprietary tools have the same issues and with less transparency. It was also reported that securely storing data is not compatible with replicability. Another challenge interviewees identified includes differing privacy and security policies across the different agencies the researcher might get their data from or work with. Conflicting policies may require greater time and resource inputs, as new workflows and tools are implemented and learned in order to adhere to those requirements and restrictions.

Interviewees also discussed ethical concerns surrounding their big data research practices. While many interviewees shared ethical considerations, one interviewee did not have ethical concerns and felt that IRB takes care of these questions. Frequently however, researchers struggled with IRB requirements, which rarely seemed to fit big data projects well. Some felt IRB approval for their research was unnecessary or outdated with the way research was performed, particularly around sharing data with collaborators or working with data in the cloud. This points to a need for data privacy education on campus on how IRB, open-source and proprietary tools, and cloud computing intersect.

Other ethical concerns were related to bias, accessibility and technology. For example, one interviewee shared concerns related to the nature of a population being studied, people with intellectual or developmental disabilities. This interviewee shared concerns around equitable access to data and research outputs for their subject communities:

> “we're working with people with intellectual and developmental disabilit[ies]. And so … one of our challenges … is how do we understand the ways that they interact with data that are not a function of the technology. And how do we make it such that they can access, and they can really work with the kinds of technologies that we're trying to build and the kinds of guidelines we’re trying to set up for a more equitable engagement with data.”

This indicates a challenge with accessing and presenting data and other research output in a universally accessible manner, and points to how data engagement requires equitability. This challenge may not be unique to big data researchers, but it may be made more complicated by the volume of subjects. Another interviewee working in statistical genetics described the issue being wrestled with in their discipline:

> “There are concerns that the data is way oversampled for Europeans and the discoveries that we make have a Eurocentric bias. And that's true, and that is something that we're not going to change, because you know we're not collecting the data, but it is something that's on almost everybody's radar at this point, and there are huge initiatives right now to change that. So hopefully that won't be as much of a case in the future.”

This interviewee recognizes an ethical issue that their discipline faces and also sees work being done to make progress on fixing the problem of bias.

Interviewees demonstrated careful consideration of ethics in their research. For instance, several interviewees shared examples of data that were not technically private, but which they remained careful in using and thoughtfully considered potential ethical issues. One interviewee collects GPS location data and treats it as sensitive data even though it is not PHI or PII. Another conducts machine learning research, using real people’s musical talent to generate musical robots, and has thought through a number of potential questions related to the attribution of the musical talent. As a final example, one interviewee shared that the data they use is available, but the manner in which they got their data was technically in conflict with the terms of use of platform from which they got it, highlighting again how infrastructure, tools, and policy may be putting up roadblocks to big data research.

Several interviewees discussed the importance of ethics to data education, especially for graduate students in data science programs. Specifically, interviewees felt that graduate students should explore the implications of the data being collected on people. Additionally, students should be encouraged to consider unforeseen consequences of data being collected, as well as implicit bias in coding.

Learning through connections and collaborations

Interviewees were asked about staying abreast of new developments in their field, and training they have received or would recommend. The findings related to these questions also highlight how collaboration is central to big data researcher’s practices and success, allowing them to leverage the expertise of others and learn about new developments. For example, one interviewee stated “since some of the things that we're doing are brand new, it's more useful to have access to an expert or team of experts who can help with the creating part as opposed to teaching me how to do stuff.” Talking with collaborators about new ideas, attending conferences and seminars, and using social media (such as Twitter) were all ways interviewees learn about and contribute to new developments and ideas.

The student researchers that interviewees work with are another important source of expertise. Researchers may seek out students with certain expertise to join their lab or collaborate with, or they may guide their students to training (via courses, instituties, workshops, and more) so that the student can become proficient in needed skills, and then share their knowledge with the faculty researcher. Interviewees discussed how new technical developments are difficult for them to stay abreast of, and for one interviewee this led to their students deciding the architectural, platform, and coding choices for the lab because the students can more effectively stay tuned into technological developments. However, another interviewee noted a flaw with this strategy, sharing “the way academe works is, you know, you're reliant on your graduate students, and they, they don't stick around forever. So you have to, your expertise comes and goes.” This indicates that faculty may feel there are gaps in the knowledge or skill set of their team, and the faculty researchers may not be able to fill those gaps themselves because of the difficulty of keeping up with technical developments.

In contrast, learning-by-doing and self-directed learning are also important strategies. For example, interviewees stated “I think all the training is just, training that you get from doing the research” and that they prefer to learn via “applied personal projects, rather than generic kinds of introductory projects.” This highlights how researchers focus on learning exactly what they need for their project. When learning what is specifically needed for their research, they also have a preference for when they learn something. Interviewees shared that they prefer “just-in-time learning,” and another interviewee detailed what this just-in-time self-learning looks like: “a lot of it has really again been self taught through other people's documentation and tutorials.” Faculty researchers may seek out asynchronous learning materials to consult while learning-by-doing.

Researchers and their graduate students employ numerous other strategies for formally learning and developing their skills. Learning from courses, institutes, workshops and the like was generally done by graduate students, rather than interviewees. Interviewees sent their students to these learning opportunities to grow an understanding of disciplinary research practices, learn a specific tool, and more, including “it's also part of the collaboration and making new connections with different players that are out there.” Several interviewees talked about developing curricula that would attract and support graduate students in big data research, for example one said: “I think … we want to …attract graduate students … who have these skills, or want to acquire these skills. And the way to do that is to have exciting, innovative, and high profile programs.” This quote highlights how the interviewee views curriculum and learning opportunities as entwined with having graduate students who can help conduct big data research. This connects to our finding that faculty rely on their graduate students to learn new developments. It is likely that educational opportunities for graduate students will lead to some faculty learning, as graduate students share with their own students, advisors, and collaborators. Curriculum needs to be supported by technical and social infrastructure, as highlighted by one interviewee who discussed the challenge of creating opportunities for a class to gain hands-on experience with data analysis. The interviewee did not know of resources available for a class of 50-100 students to download and/or manipulate a database for an in-class experience, although that experience was central to the course goals.

Collaboration and learning are deeply intertwined. Collaborators and connections often point researchers to new developments, or they use their connections with other experts to learn or do something new. On the flip side, attending various trainings is a great way for graduate students to develop new connections and collaborations.

Summary and Recommendations

In creating the Center for Research Data and Digital Scholarship (CRDDS) and coordinating the data services of University Libraries and Research Computing (RC), CU Boulder has taken a tangible step in supporting big data research. CRDDS provides a central location where researchers can receive help with a broad range of topics encompassing data research and pedagogy. Expanding upon this model to more fully and formally support big data issues may be a logical next step for the University (see recommendations below).

Interviewees indicated uncertainty and lack of knowledge around presently available services on campus. This indicates a need for enhanced marketing and outreach to increase awareness around data support and education. Consideration should be given to outreach strategies used. For example, the findings within suggest that graduate students are an important source of knowledge and expertise for their faculty advisors. Thus graduate students may be a particularly important audience to reach.

The findings of this study were inconclusive about the disciplinary reach of research data services at CU Boulder, and how these services might be accessed more widely. Our sample size and focus makes it difficult to identify specific inequities in disciplinary access, but additional outreach is needed to identify and address disparities that may hamper interdisciplinary and domain-specific research.

The study results indicate a clear need for additional support around data ethics, security and privacy. Interviewees have varied questions and concerns around many stages of the big data lifecycle, suggesting that ethics support and education could be integrated into most big data services (e.g., via documentation and consultations that address the specific areas in which researchers work), and more generally (e.g., through development of curriculum). Additionally, at least one interviewee alluded to challenges posed by the IRB process — which traditionally has dealt with questions of ethics, security and privacy in research — in the context of big data research. The complexity, “messiness,” and dynamic nature of big data may be difficult to address via IRB assessments, and questions may also arise regarding how to keep IRB protocols up-to-date over the course of a project given the constantly-evolving nature of big data.

Another need is for guidelines and ideas for attributing credit to all of the contributors for big data research that speaks across disciplines and across industries. While different disciplines have different norms for credit, success, and promotion, data and information experts can provide support by facilitating conversations and collaboration, and offering guidance that aligns with open and ethical data practices.

Finally, this study has highlighted that big data research will continue to grow, and that research practices, including how data are accessed and used, will continue to change. Users want interfaces that facilitate convenient and automated ingest or use of data, so that they can more efficiently analyze these collections for greater amounts of information. This common need opens up opportunities for critical big data education around IRB and privacy, unintended bias with tools, and demystifying the data management process across disciplines.

Recommendations

We recommend the following actions:

- Development of big data training curricula, particularly for big data ethics, privacy and security, through a variety of channels (e.g., documentation and context-specific consultations for specific big data services, more general course-based curriculum).
- Consider how to address the complexity and dynamic nature of big data in the IRB process in a manner that fully and reasonably considers the ethical, security and privacy implications of a given big data research project.
- Explore CU Boulder guidelines for attributing credit to the myriad contributors in big data research projects, and considering the sometimes unconventional contributions, with the goal of helping departments develop clear policies and incentives for researchers performing big data research.
- Development of promotion and outreach strategies to increase awareness of existing and forthcoming big data research support services at CU Boulder, in a manner that promotes equitable access to services across disciplines, colleges, and centers.
- Assessment of staffing gaps and staff-training needs in order to support big data curricula and services in University Libraries, CRDDS, and RC.
- Periodic evaluation of emerging trends and environmental scan of needs in big data research, in order to adjust strategies and services appropriately to ensure CU Boulder is providing state-of-the-science support and infrastructure.

An optimal way of addressing the questions above, most of which are complex, may be to establish a “Big Data Steering Committee” composed of a broad range of CU Boulder (and possibly external) stakeholders and decision makers.

Acknowledgements

Our thanks to Jonathon Anderson, Cindy Edgar and Ithaka S+R.

References

Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., Moghaddam, S.H.A., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q., & Brisco, B. (2020). Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5326-5350. doi: 10.1109/JSTARS.2020.3021052.

Burton, M., & Lyon, L. (2017). Data science in libraries. *Bulletin of the Association for Information Science and Technology*, *43*(4), 33-35. doi: 10.1002/bul2.2017.1720430409

Dedoose Version 9.0.17, web application for managing, analyzing, and presenting qualitative and mixed method research data **(**2021). Los Angeles, CA: SocioCultural Research Consultants, LLC www.dedoose.com.

Guo, H. (2015). Big data for scientific research and discovery. *International Journal of Digital Earth*, *8*(1), 1-2. doi: 10.1080/17538947.2015.1015942

Khalil, M., Said, M., Osman, H., Ahmed, B., Ahmed, D., Younis, N., ... & Ashmawy, M. (2021). Big data in astronomy: from evolution to revolution. doi: 10.14419/ijaa.v7i1.18029

Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, *19*(4), 208-219. doi: 10.1038/nrg.2017.113

Longhurst, R. (2003). Semi-structured interviews and focus groups. *Key methods in geography*, *3*(2), 143-156.

Lycett, M. (2013). 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381-386.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Sage publications.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: applications, prospects and challenges. In *Mobile big data* (pp. 3-20). Springer, Cham.

Whitney, M., (2022). “39 Twitter statistics marketers need to know in 2022.” https://www.wordstream.com/blog/ws/2020/04/14/twitter-statistics. Accessed April 6, 2022.

Appendix A

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 B

Codebook

Parent Code	**Child Code**	**Definition / Criteria / Examples**
Practices		(if they say, e.g. "this is how we do it")
	getting data	when interviewees discuss how they get the data they work with, any data, any methods
	processing / analyzing	when interviewees describe using and interpreting the data content
	data curation	when interviewees describe philosophies of gathering and processing of specific data types/topics AND automated metadata processing/description
	publication	when interviewees discuss what they do to publish and share their work - data, articles, presentations, etc
	collaboration	when interviewees describe the strategies (eg communication practices) they use to collaborate with others, broadly
	planning	when interviewees describe the processes that they use to determine what their data requirements will be
	tools	when interviewees talk about software they they use or build
	students	when interviewees talk about involving students
Challenges and Needs		(if they say, e.g., "this is what we need") (or "we have trouble with this aspect")

	access	when interviewees address issues with obtaining data and resources OR for others to acquire their data
	curation	when interviewees describe issues with gathering and processing data, philosophically and technically (e.g. metadata)
	tools	when interviewees point to issues with software, instruments, and sometimes OER to complete their research
	infrastructure	when interviewees refer to needing experts (social infrastructure, data as a service) and technical resources that help manage, store, promote and consult on their data/research
	planning	when interviewees refer to needing support with data management plans for grant funders OR when they address long-term preservation and access to their data/projects
	collaboration	when interviewees refer to issues with working with other researchers / institutions
	getting data	when interviewees discuss challenges with how they get the data they work with, any data, any methods
Unresolved Questions		("we don't know how to deal with this")
	privacy	when researchers address caring for sensitive data
	ethics	when interviewees express ethical concerns regarding their data use or research that can't more specifically be categorized as an issue of privacy, security, or bias OR when interviewees express concern about the nature of their data and broader implications to society, individuals, and others (researcher hesitation)
	security	when interviewees describe issues or practices related to securing their data, or the affects their data or research may have on the security of others
	bias	when interviewees express ongoing concern regarding existing biases in themselves, their field, their data or their tools; at this time social bias belongs in ethics
	personal vs institutional responsibility	when interviewees describe research issues where it is not clear if the researcher themself needs to resolve the issue or if it is something the institution should resolve
Learning and Development		("how we build skills") (or "we have these specific training needs"), potentially co-code with practices or challenges and

		needs
	students	when interviewees describe relying on their students to learn or develop new skills / knowledge in order for the group to do big data research, OR when interviewees describe learning from the students, OR when interviewees talk about learning opportunities they encourage their students to take advantage of, OR learning opportunities they wish existed for their students
	collaboration	when interviewees describe working with collaborators and learning via those collaborations, or when they describe seeking out collaborators with new skills / knowledge, irrespective of institutional boundaries
	tools	when interviewees discuss how they learn to work with new tools, including applications, software, coding languages, etc.
	self training	when interviewees describe learning new things or developing skills via just figuring it out for themselves, or seeking out instructional materials that help them teach themselves
	internal	when interviewee describes learning happening with the research team/group, or wanting learning opportunities at this level
	classroom/ curriculum	when interviewees discuss their students learning big data research skills and knowledge via their coursework, or need for learning at the level of credit courses
	conferences/ workshops	when interviewee mentions learning at a conference, workshop, or similar event, or a desire for a conference, workshop, or event to be available so they can learn something
	Institution	when interviewee mentions learning (or wanting learning opportunities) at a CU workshop, training, or other event that is not credit coursework