Title

Supporting Big Data Research at the University of California, Berkeley: An Ithaka S+R Local Report

Permalink

https://escholarship.org/uc/item/4403c0f4

Authors

Foster, Erin D
Glusker, Ann
Quigley, Brian

Publication Date

2021-10-01

Oct. 1, 2021

Supporting Big Data Research at the University of California, Berkeley

An Ithaka S+R Local Report

Erin D. Foster
Research Data Management Program Service Lead
Research IT & University Library

Ann Glusker
Sociology, Demography, Public Policy &
Quantitative Research Librarian
University Library

Brian Quigley
Head, Engineering & Physical Sciences Division
Mathematics, Statistics & Computer Science Librarian
University Library

With contributions from Anthony Suen, Director of Programs, Data Science Education Program, Division of Computing, Data Science & Society as a consultant.

Berkeley Library
UNIVERSITY OF CALIFORNIA

doi:10.25350/B59G6F

BACKGROUND

In 2020, an Ithaka S+R project on “Supporting Big Data Research” brought together twenty-one U.S. academic institutions to conduct a suite of parallel studies aimed at understanding researcher practices and needs related to data science methodologies and big data research. Ithaka S+R is a not-for-profit research and consulting organization that “helps academic and cultural communities serve the public good and navigate economic, technological, and demographic change.” The University of California, Berkeley joined the project, and over the course of the 2020-2021 academic year, a team conducted and analyzed interviews with a group of researchers at the university. In addition to UC Berkeley, other institutions completing the same study included peer institutions such as the University of Illinois at Urbana-Champaign, University of Virginia, University of Wisconsin-Madison, and UC San Diego. This local report outlines the findings from the interviews with UC Berkeley researchers and makes recommendations for campus and library support for big data research. The Ithaka S+R capstone report will synthesize findings from all of the parallel studies to provide an overall perspective on evolving big data research practices and challenges to inform emerging services and support across the country.

There is a growing body of research and writing on the evolving relationship between libraries and data science research support. Burton & Lyon (2017) discuss efforts to address the skills gap and management gap in supporting data science in libraries through initiatives like the Data and Visualization Institute for Librarians[1] and Library Carpentry[2]. In a later report, they further outline recommended actions for fostering a data-savvy library workforce based on a two-day international workshop (Burton et al 2018). Extending this line of thought, others call for a holistic data science strategy in libraries, expanding existing support for data curation to also include support for data analysis, based on survey results showing demand for data analysis training among librarians (Maxwell, Norton & Wu 2018). An overview of the “data science revolution” in research libraries outlines additional progress on data science initiatives in libraries through case studies such as data literacy efforts at Georgia Tech and library collaborations with Moore-Sloan data science environments at UC Berkeley, New York University, and the University of Washington (Association of Research Libraries 2019). Oliver, Kollen, Hickson & Rios (2019) argue that the academic library’s role as interdisciplinary campus hub makes it particularly well-suited to provide data science support with a focus on computational literacy, geographic information systems, and reproducible science. While these studies address the role of libraries in supporting data science generally, not much has been written on library support for big data research specifically.

Focusing on the Berkeley context, the campus Research & Academic Engagement Benchmarking Project[3] efforts in 2014, 2016, and 2018 provided campus leaders with important context on research and data science support services at Berkeley and peer institutions to support their planning. Within the University Library, an in-depth survey of faculty conducted in partnership with Ithaka S+R in 2018 offered

[1] https://www.lib.ncsu.edu/data-science-and-visualization-institute
[2] https://librarycarpentry.org/
[3] https://rtl.berkeley.edu/rae-services-peer-benchmarking

insights into faculty perspectives and practices related to research data management. As that report states, “Results from questions about research data management agree with national results that showed that faculty value tools that allow them to maintain their own data themselves. One difference is that a somewhat greater proportion of UC Berkeley researchers utilize cloud storage than those at other institutions (58% at Berkeley vs. about half nationally)“ (Li et al 2019, p.42). However, nowhere in the report does the term big data appear, inviting investigation into the degree to which big data researchers engage in independent data management.

METHODOLOGY & PARTICIPANTS

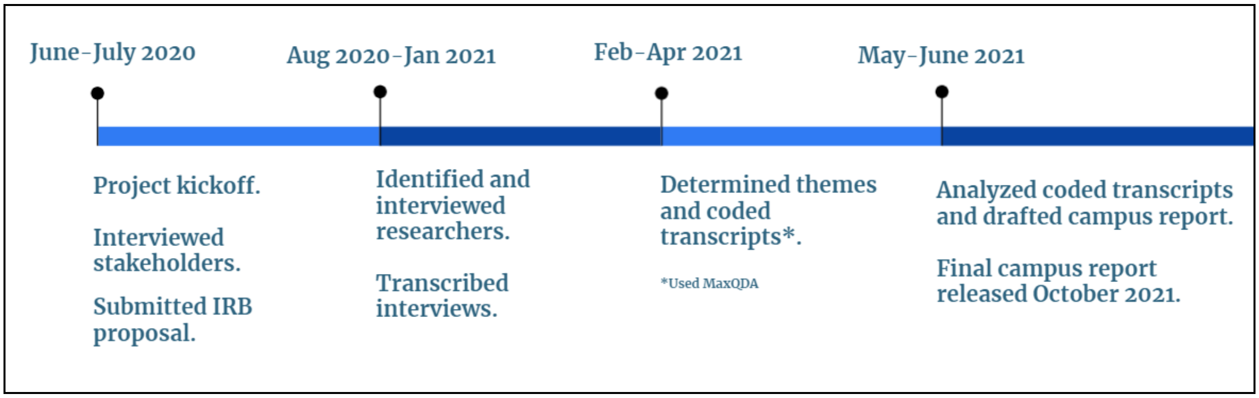
June-July 2020
Aug 2020-Jan 2021
Feb-Apr 2021
May-June 2021
Project kickoff.
Interviewed stakeholders.
Submitted IRB proposal.
Identified and interviewed researchers.
Transcribed interviews.
Determined themes and coded transcripts*.
*Used MaxQDA
Analyzed coded transcripts and drafted campus report.
Final campus report released October 2021.

Figure 1. Project Timeline

After obtaining IRB clearance from the Office for Protection of Human Subjects, the team began recruiting interviewees for the study. The team used a purposive sampling approach to recruit participants. Team members contacted potential interviewees who were identified as likely big data researchers based on personal knowledge, faculty websites, and recommendations from stakeholders and those who declined to be interviewed themselves. The team aimed to interview researchers from across ranks and disciplines, the main criterion being that their research involved big data, defined as data having at least two of the following: volume, variety, and velocity[4]. Of the 46 researchers invited to participate, 16 participants agreed to complete semi-structured interviews; this participation was within acceptable bounds for robust results (Guest, Bunce & Johnson 2006; Baker & Edwards 2012).

[4] For definitions of the 3 V’s, please see: https://bigdataldn.com/intelligence/big-data-the-3-vs-explained/

Rank	Total
Professor	6
Associate Professor	1
Assistant Professor	5
Postdocs/Research Staff	4

Table 1. Participants by Rank

Discipline	Total
Arts & humanities	2
Life & health sciences	4
Physical sciences & engineering	5
Social sciences	5

Table 2. Participants by Discipline

While participants were classified into disciplines based on their primary departmental affiliations, many researchers crossed disciplinary lines through joint appointments or multidisciplinary research projects. We did not ask participants to self-identify their gender or race/ethnicity, and we do not know the actual proportions of big data researchers on campus for any of those categories. However, based on online biographies, the team members determined that the 16 interviewees included 5 women in the sciences and social sciences, and the team also strove for racial/ethnic diversity among those interviewed.

The qualitative study process was conducted in accordance with established, rigorous methodology. It used a grounded theory approach, as outlined in the literature and through instruction from the Ithaka S+R team (Corbin & Anselm 2015; Creswell 2002; Creswell 2007). Each participant was interviewed by one team member using a semi-structured interview protocol developed by Ithaka S+R. Interviews were recorded, transcribed, and redacted to remove identifying information. Recordings were deleted once the transcriptions were completed. Transcripts were coded in MaxQDA using an open coding process to identify and develop themes.

The team identified six themes that will be discussed in the following sections:

- Data Collection & Processing
- Analysis: Methods, Tools, Infrastructure
- Research Outputs
- Collaboration
- Training
- Balancing Domain vs Data Science Expertise

Research in the time of COVID-19

The interviews analyzed in this report took place from October 2020 through January 2021, at a time when the prevalence of COVID-19 was affecting research and teaching throughout the nation and world, and the UC Berkeley campus was virtually closed. Interviewees were offered the chance to report on their research either before or during COVID.

Among STEM researchers, the main reported impacts on their work were changed modes of communication; rather than in-person conferences and casual conversations, information was disseminated virtually, with several interviewees noting that it was easier to keep up with information and attend talks and webinars in the virtual environment. Social sciences and health sciences researchers reported pivoting in order to study COVID's societal effects directly, which meant having to quickly create research agendas and infrastructure that could accommodate big data where none existed before, as well as new pathways for research dissemination and publishing. Due to the extent and nature of the crisis, much secondary data that had previously been proprietary was being made available openly, and researchers, especially those with existing relationships with data providers, jumped at the opportunity to use them. This worked both ways, however, as some sources of data dried up due to COVID-related agency closures, and of course planned fieldwork had to be canceled as well. In every case, among those who mentioned COVID as a factor, the readiness to shift agendas and adjust to a new normal was of central importance.

DATA COLLECTION & PROCESSING

The Data Collection & Processing theme covers all aspects of gathering, acquiring, or capturing research data in order to answer research questions and undertake analysis of some kind. This theme also includes preprocessing needed to ensure research data is ready for analysis.

The researchers interviewed, all of whom used big data, fell along a wide spectrum -- whether dealing with satellite imagery that results in high volumes of data or air quality data that needs to be processed at a high velocity or a digital humanities project that involves a high variety of data types and sources. Often researchers were engaged with big data in several ways in a single project, for example, dealing with high volume and velocity with the air quality data project or high volume and variety in a research group that develops high bandwidth instrumentation.

Use of primary & secondary data

It was common for those interviewed to deal with both primary and secondary data in their research. Primary data is defined as data directly collected and/or gathered by a research team and secondary is defined as pre-existing data, such as census data, collected by an external entity and provided to the research team.

Regarding secondary data, there were a variety of sources mentioned from which researchers obtained data. The most commonly mentioned were federal agencies (e.g., National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, Environmental Protection Agency) followed by industry/private sources (e.g., Ancestry.com, Google). In general, data from federal or state agencies is “free” in cost (if not in time and efforts to wrangle) while data coming from industry/private sources often comes at a cost. In a nod to COVID-era research, an interviewee stated in regards to their use of secondary data:

> *...prior to COVID, these data were really hard to obtain because they’re held by private companies that are not interested in research. There’s nothing in it for them for some academic to go write a research paper on [research topic]. Now with COVID, all the companies are making all this data available. Google has all their mobility reports, I think Apple has some, [Platform X] is providing their data. [Social sciences researcher]*

Of the researchers interviewed, several also created their own datasets from secondary data, which made the distinction between primary/secondary less clear. A social sciences researcher noted: “We definitely spend a lot of time building new datasets... It’s not like we’re using prepackaged datasets from [Social media platform X], we have to create it from our own end.” This was similar to what one arts & humanities researcher experienced as well: “We have to actually produce the data from the texts...we’re producing something that is very different in structure, different in kind from what the primary documents are.”

In a similar spirit, a life & health sciences researcher noted that “...there’s a subgenre of data collection that’s not really a collection but it’s important to me and it’s making analog data digital. So we do some digitization of historical data (...) found on old maps or in archives.” This act of making analog data digital was raised with an arts & humanities researcher as well who dealt primarily with historical documents and, as a condition of using these historical documents from an archive, was asked to digitize these materials for future use. While this is akin to data/code sharing, it also demonstrates how the original datasets in these instances are as much products of research as some of the more traditional products (e.g., academic papers, derivative datasets).

Challenges

There were several challenges that emerged when discussing data collection and processing with those interviewed. Many discussed the challenges of “noisy” data collection -- as one researcher put it: “Generating the data is the easy part. Recognizing whether the data you’ve generated is trash or garbage or useful is the other time-sensitive challenge." Several researchers touched on this challenge, with a physical sciences & engineering researcher stating how it can make it difficult to evaluate results of others because "...you don't know if the data are meaningful outliers or if they're just outliers because they're reported wrongly." The challenge of evaluating data -- especially in this data collection and processing stage of research -- was a consistent thread that appeared among interviewees and seems to be a potential gap in education as well as training for researchers. In light of the focus on big data as well, the scale at which data is being generated and processed adds to the challenge of this stage of research.

When it came to secondary data in particular, a researcher in the life & health sciences noted the challenge of evaluating datasets for use, in addition to inconsistency in metadata in the datasets from both federal and industry/private entities. This researcher also noted "a proliferation of formats" when it came to these datasets. Cost was raised several times as well with no particular interviewee noting it as a unique challenge, more so just the reality of dealing with private/industry entities.

An additional element of dealing with data from an external entity is the need to establish, in many cases, a contract or agreement of some kind to use the data for research purposes. For example, one researcher in the physical sciences & engineering stated for their research:

> *You need to have a separate data mining agreement. Most of the time it specifically describes the terms of usage of this data, what you can share from this data. For some publishers we had to specify the machine we'd be using, provide the IP address, and we have to sign in the agreement that we will not be using a large speed of download in order to prevent an overload in the server. [Physical sciences & engineering researcher]*

In the case of another researcher in the social sciences, they found the process to be challenging, particularly as it came to working through University of California (UC) processes:

> *[A] typical kind of data agreement [is] where we have to have data in a secure set up and the access is restricted. People have to write an application, it has to be approved, it has to pass through the lawyers. I would say the biggest hurdle is UC bureaucracy. It seems much more work intensive for people at UC than it is at other institutions. [Social sciences researcher]*

Several of the researchers discussed COVID-specific issues as well including difficulty getting access to data collection instruments or trouble accessing areas where data collection instruments were deployed. There was also reference to existing challenges with obtaining current data from federal agencies -- an issue that was heightened during the pandemic.

Campus services & support

When it came to data collection and processing, there were not many campus services explicitly mentioned in the interviews. The campus's high performance computing cluster, Savio, was mentioned several times[5] as well as the ZTRAX dataset which is jointly managed on campus by the D-Lab and the Office of Intellectual Property & Industry Research Alliances (IPIRA)[6]. One researcher noted how great particular campus services can be (Berkeley Open Computing Facility[7] in this case), yet they still encountered challenges when it came to storing data:

> *I tried to reach out and utilize campus resources for this, and the open computing facility on campus is an amazing resource. They're happy to host your website, but if you're… I was like 'I need a few hundred gigabytes for my data in your database,' and they were just like 'No. We can't help you, that's too much.' I also reached out to… I forget the other department on campus. They basically charged more money than Google was charging us. We're paying 300+ dollars a month and theirs would have been $450. I was like 'Okay I think I'm just gonna buy my own servers to get this working, can you help with that?' And they were like 'We should be able to help with that,' but it was crickets after. [Physical sciences & engineering researcher]*

By and large, mentions of campus services were positive throughout all of the interviews -- even when noting gaps in support and resources. As will be discussed further in later sections, lack of familiarity with available campus research support programs, resources, and systems was the most clear takeaway. For example, when it comes to data collection, there was no mention of campus-supported data collection tools such as Qualtrics. While that could entirely be because it is not relevant to those interviewed or not thought to be mentioned, there is much to be said for the need to continuously promote existing campus services to ensure the research community knows what is available to them.

Overall, it was consistent throughout the interviews that researchers and research groups often rely on themselves for data collection and processing support rather than campus services.

ANALYSIS: METHODS, TOOLS, INFRASTRUCTURE

Following data collection, the Analysis theme covers all aspects of the stage of research in which researchers are analyzing the data that they have collected or acquired. This includes analysis methods, software, tools, writing code, and computing infrastructure.

Methods

While many researchers continue to rely on traditional statistical analysis methods, researchers from across all disciplines mentioned using machine learning, natural language processing, and visualization techniques in their research. Individual researchers mentioned specific techniques such as predictive modelling, deep learning, regression analysis, clustering, entity recognition, and Bayesian statistics. As stated by one researcher, the choice of methods or models "really depends on your class of questions."

[5] For more information about Savio and high performance computing at UC Berkeley, please see: https://docs-research-it.berkeley.edu/services/high-performance-computing/overview/

[6] For more information about the Zillow ZTRAX dataset, please see: https://www.zillow.com/research/ztrax/

[7] For more information about the Open Computing Facility, please see: https://www.ocf.berkeley.edu/

The size of their datasets, which can easily get to be multiple petabytes[8], often drives researchers' decisions to use data science methods. One social science researcher notes that traditional statistical methods "break when we try to do things like look up patterns." However, that was not the only reason to use them. More than one researcher indicated the need to use algorithms to repeat the same processes and computations whenever new data was ingested. Several researchers also discussed the opportunities for exploratory data analysis and data mining that high-dimensional datasets allow. In interviews with several researchers, they discussed the role of hypothesis testing versus data mining in big data research. As one researcher stated, hypothesis testing still plays an important role:

> *There is a role for that exploratory data analysis sometimes. But at the same time, you want to be able to articulate some hypothesis and have it be theoretically driven so you know what's an interesting question worth answering in the first place. [Social sciences researcher]*

On the other hand, another researcher noted risks in starting with a hypothesis:

> *But I think things are changing a little bit because with these new technologies, there's no hypothesis because it's pre-imposing a strain on the data because you think this is gonna happen. You might actually end up hurting yourself. [Life & health sciences researcher]*

In a variety of ways, expertise remains a challenge in using these data science methodologies. Some researchers mentioned the need to understand the discipline and dataset to be able to recognize if algorithms are generating rational results.

> *It is always amazing to me to see how the clustering actually identifies things that I was expecting it to. It's not just overfitting, it's finding really important boundaries, so that's exciting. It requires some knowledge about how to get the right parameters in the settings, and that takes a little play and exploration. [Arts & humanities researcher]*

Others wanted more help from data scientists to implement machine learning techniques or to discover new methods that could advance their research.

> *I think it's a continuing challenge for us in the domain to understand and access new approaches in an experimental mode. Once we know that we need to take this approach or use this tool to tackle this problem, then we can go get the resources or access to resources or expertise and access to expertise to make it happen. [Physical sciences & engineering researcher]*

[8] 1 petabyte = 1 million gigabytes

Some researchers also expressed concerns about the potential for misusing machine learning approaches. With packages for machine learning readily available, people can unintentionally produce poor results or even biased results if they do not understand how the packages work or how to use training sets properly. As one researcher with a computer science background described it:

> *I really don't like when people use the tools without understanding how they work, especially now when it's related to machine learning when we have all these nice libraries which are used as the black boxes. [Physical sciences & engineering researcher]*

Or as another researcher put it succinctly: "It's pretty easy to do things nowadays. [The phrase/mindset of] 'I know programming' can be dangerous."

A life & health sciences researcher also raised the issue of algorithmic fairness:

> *If you base your algorithms on existing practice and the existing practice is biased, you reproduce the biases in the computer but with the veneer of unbiasedness to them, which can be really bad for us structurally … It has to do with everything from who's included in the training sets to how much you pressure-test implicit bias in what you're reporting. [Life & health sciences researcher]*

These discussions indicate a need to help researchers and students to think critically about choosing machine learning libraries, implementing machine learning methods, and constructing training sets.

Tools

Researchers continue to use a wide array of programming languages, open software, and proprietary software for their analysis. This includes pockets of researchers that continue to use proprietary software such as Matlab in the sciences and SPSS and Stata in the social sciences as well as others who are using the open source language R. However, the most popular tool mentioned in the interviews by far was Python. Here is a full list of the analytical tools mentioned by researchers in their interviews:

- ArcGIS
- C++
- CUDA
- Doc2Vec
- Excel
- Fortran
- Google Colab
- Google Sheets
- iTorch
- Julia
- Jupyter Notebooks
- MatLab
- Python
- R
- ROOT
- SPSS
- Stata
- TensorFlow
- Tesseract
- Word2Vec

The interviews clearly indicated a strong trend toward using Python and Jupyter Notebooks for research at UC Berkeley. This trend included researchers from each of the broad subject areas: arts & humanities, life & health sciences, physical sciences & engineering, and social sciences. This is not surprising since the co-founder of Project Jupyter is a Berkeley faculty member and the Data Science Education Program relies on Python and Jupyter Notebooks to teach data science to undergraduates, reaching thousands of students across every discipline each year. As an open-source tool, it has also allowed for the shared development of packages within research communities. We expect to see continued growth in Python use across all disciplines as a result.

> *Within the last 3 years or so, the community has transitioned. All of these tools have been incorporated into packages within Python. So Python is really the prime tool that everybody is using. [Physical sciences & engineering researcher]*

On the other hand, for some research projects with extensive legacy programming, the cost to transition to Python is just too much. As one life & health sciences researcher explained:

> *Switching is a huge hurdle because it's a full time job of 5 or 6 computer scientists to be able to switch over the code base that we developed over to Python and have it function exactly as expected. [Life & health sciences researcher]*

While there is a strong trend toward Python, several researchers described using different languages for different purposes related to speed and efficiency. A physical sciences & engineering researcher discussed using a framework "written partially in C++ because it's fast and then partially in Python which can be convenient for doing final data analysis." A life & health sciences researcher similarly stated that "when it's bread and butter in terms of processing, we go low level. So we have C, C++ or CUDA applications." Perhaps that researcher explained it best when they stated:

> *We basically go with whatever is the most robust, whatever is the most sustainable and whatever is the most expedient. It's a balance of those things. [Life & health sciences researcher]*

Infrastructure, storage & organization

In discussing data storage and computing infrastructure for research, participants described using a mix of local resources, departmental resources, campus resources, collaborators' resources, and cloud services. Researchers seem to make choices based on their specific circumstances, focusing on the simplest or most economical options in each case. However, it was not always clear if those interviewed knew what additional resources were available centrally. Some researchers described using different options for different projects, while others might use multiple options across one project:

> *Typically we run the first steps on the grid until we produce a dataset that's on the scale of less than a terabyte. That's typically the goal. Once we're on that scale, we bring it*

> *here to Berkeley, up to Cori, up to the NERSC [National Energy Research Scientific Computing Center] supercomputer. Then we run the next step of analysis. The final step, the one where we make the plots, that's something we run on our laptops. [Physical sciences & engineering researcher]*

In addition to their own computers, researchers made use of departmental servers, cloud services such as Amazon Web Services and Google Colab, and the campus high performance computing cluster Savio.

Savio & Condos

The high-performance computing cluster at UC Berkeley is known as Savio, named for Free Speech Movement activist Mario Savio. "As of April 2020, the system consists of 600 compute nodes in various configurations to support a wide diversity of research applications, with a total of over 15,300 cores and a peak performance of 540 teraFLOPS (CPU) and 1 petaFLOPS (GPU), and offers approximately 3.0 Petabytes of high-speed parallel storage."

The cluster consists of nodes purchased by Research IT as well as nodes contributed by researchers as part of the Condo program. Researchers purchase their own servers that are connected to the cluster, allowing other researchers to use any idle compute cycles. In exchange, these condo owners get access to Savio's high-speed interconnect and parallel storage. "By becoming a Condo Partner, through purchasing and contributing compute nodes to the Savio cluster via the Condo Cluster Service, researchers and their groups obtain priority -- and hence nearly unlimited -- access to resources equivalent to their contribution."

Source:
https://docs-research-it.berkeley.edu/services/high-performance-computing/

Five researchers in the life & health sciences, physical sciences & engineering, and social sciences explicitly mentioned using high performance computing in their own research: three using Savio, one using a National Energy Research Scientific Computing Center (NERSC) supercomputer, and one using a mini compute cluster that they built for their own lab:

> *We built a mini compute cluster with half a petabyte of storage in 2 tiers. One is a fast all Flash array and the second tier is slower but larger capacity. And then we have 600 compute cores over 24 nodes. We have a dozen GPU's for all the machine learning and CUDA processing. [Life & health sciences researcher]*

Two other researchers also described interest in using high performance computing but noted challenges in getting the necessary software or codes to work on the campus cluster, even with support from Research IT, as described by this arts & humanities researcher:

> *[A Research IT consultant] initially helped with the OCR stages in the initial kind of way, with [the high performance computing cluster]. It didn't end up working so well because it was too cumbersome at the time and I couldn't figure it out entirely. There's always great intentions from people around. The consultants. And the technical support is there. It just has to match the time and place with schedules and my readiness for it, it's such a big project. [Arts & humanities researcher]*

Even when not using the campus cluster, researchers indicated that they were increasingly relying on shared computing power like departmental servers and cloud computing services when their own computers were not sufficient. As a physical sciences & engineering researcher explained:

> *AWS [Amazon Web Services] and various others like Google, they make various cloud resources freely available to researchers up to some sort of limit. Several of our researchers have made use of that. For big machine learning or big simulation processes. They can use a bunch of servers for an afternoon. [Physical sciences & engineering researcher]*

Several researchers also described the trend toward increased use of graphics processing unit (GPU) nodes and tensor processing unit (TPU) nodes in their research. Much of their computationally-intensive research requires these new architectures for machine learning and natural language processing. Savio and other high-performance computing clusters incorporate GPU nodes into their architectures in addition to CPU nodes, and TPU nodes are available through the Google Cloud Platform and Google Colab for use with the TensorFlow framework. These are used across disciplines:

> *We've been moving fast toward significantly increased parallelism, and even new architectures, so things made of GPUs, that sort of thing. [Physical sciences & engineering researcher]*
>
> *The challenges there are the fact that a lot of work in NLP [natural language processing] right now demands a specific kind of computing device. I mentioned some GPUs as one example of this, but there's another one called TPU, which I think is a tensor processing unit, that you can only get from Google Cloud. They're really great for working with very specific algorithms, but these algorithms also happen to be the ones on which we run every part of our analysis. [Social sciences researcher]*

But this social sciences researcher also described uneven access to these technologies and the impact that computing power has on their approach to algorithms:

> *It also shapes the way that we think about developing algorithms. I work a lot with people in the humanities and social sciences who don't always have access to the server with the GPU. And we need to write code that can work on their own laptops. Trying to negotiate that balance between an algorithm that can run faster on those devices, but also one that can run locally at any machine can also be challenging. [Social sciences researcher]*

In addition to their own hard drives, the researchers interviewed use a range of options for data storage. Only one researcher mentioned using the campus data servers. The most popular options for data storage were local/departmental servers and Google Drive, while others mentioned using Box and AWS. One large-scale collaboration stores data in their own distributed network of computers but also utilizes tape storage to archive data that is not immediately needed for use.

Researchers also discussed the importance of backups and version control for their data, making use of more than one mode of data storage for redundancy. A physical sciences & engineering researcher described meeting with library consultants who encouraged them "about adhering to the 3-2-1 rule about having your data in 3 locations, one of them offsite." However, at least two researchers noted that the challenge of sufficient storage persists. With more and more data being produced all the time, researchers have to consistently upgrade their storage. Costs play a major role in decisions around data storage. As one life & health sciences researcher framed it, "Is it cheaper for me to redo the experiment? Or is it cheaper for me to buy new storage servers to have some of these things?" This leads to a preference for Box and Google Drive for many researchers since they offer free unlimited storage for Berkeley researchers, though that same researcher notes, "even though it's unlimited storage, we understand what the actual limits are because we have tried to push those." As some researchers recognized, however, there will be a need to rethink or reframe their usage of Box and Google Drive because of changes to our UC Berkeley contracts with those services.

Box & Google Drive Service Changes

"Google and Box have announced they are changing their service and pricing models for educational institutions. These new models end unlimited storage for all accounts. UC Berkeley will be transitioning to these new pricing and storage models between now and April 2023."

Source:
https://bconnected.berkeley.edu/projects/storage-changes

Managing data and code becomes an important task within these big data research projects. Some researchers discussed protocols that they have developed to keep things organized and maintain the integrity of their datasets and codebases. Long-term, large-scale collaborations have developed very mature workflows over time including mechanisms for controlling access and validating code changes.

The most popular tools for keeping things organized were Jupyter Notebooks and GitHub, with half of the researchers noting that they rely on GitHub in particular. In addition, two researchers discussed the role that GitHub plays as a portfolio for graduate students going into industry:

> *Half of our students go into academia and half go into industry. All of those that go into industry… part of getting the job is being able to point your potential employer to your GitHub repository, this is what I've come to understand. So they are very motivated to be creating this GitHub repository during their PhD so that they have that resource. [Physical sciences & engineering researcher]*

While most researchers did not have security concerns regarding their data, issues around security and privacy were raised by researchers working with geospatial data, health data, and administrative data. For example, two researchers described "fuzzing" addresses from GPS data so exact locations cannot be identified. A social sciences researcher described a project where confidential data limited the tools that they could use, and another mentioned that data providers can also impose security requirements even for data that is not sensitive or confidential. Another social sciences researcher described the risk of deductive disclosure from large and detailed datasets about people. This indicates that researchers need to think carefully about how they store and share such datasets, even when the data has been anonymized. Some researchers highlighted their need for secure research environments, which Research IT's new Secure Research Data and Compute (SRDC) platform will hopefully satisfy.[9]

Campus services & support

Many researchers expressed their appreciation for the research resources and services provided by campus through units such as Research IT, D-Lab, BIDS, and the Library. Several praised campus consultants from these units, describing experiences that helped them to explore options or implement new methods. Others indicated that they rarely rely on campus, instead building the necessary resources and technical support within their own research groups or departments, or leaning on resources that are available through their collaborations. Even when researchers relied heavily on campus tools and services, some expressed disappointment with the robustness of those resources and their costs:

> *One of the most frustrating things for me when I started at Berkeley was recognizing that we're at Berkeley in the Bay Area and being a little embarrassed by the fact that we're at Berkeley with the level of research resources that we do have. What I mean by that is obviously the Bay Area is the tech capital, the innovation capital, of the world. Berkeley has been known for pioneering open source standards and pioneering some of the most advanced network capital tools. But the infrastructure at Berkeley in the building where I'm at was, at least when I started, [low network bandwidth for this research]. And there were plans at the time to upgrade to [higher bandwidth]. The plans at that time were already obsolete for the plans that we had." [Life & health sciences researcher]*

[9] For more information about the SRDC platform, please see: https://docs-research-it.berkeley.edu/services/srdc/

One physical sciences & engineering researcher indicated a need for more technical support for research computing: “That’s one of the challenges that we have at Berkeley is we have been bleeding staff for a long time and the burden falls on the faculty and the graduate students.” They indicated the support does not need to be within the department, but having access to technical staff would be beneficial. This raises the question whether researchers are unaware of the technical consulting available from IT Client Services, Research IT, and others, or if there is a need for more or different kinds of technical support. In reality, it could quite possibly be both.

Other researchers also commented on campus network reliability and bandwidth, costly storage options, and insufficient large memory nodes in Savio. There was a sense that campus computing infrastructure may not be evolving fast enough to support data science and emerging research methodologies:

> *The idea is that we want to be trailblazers. At the same time we understand the limitations of what is feasible in reasonable amounts of time. My frustration always stems from the fact that I don’t have the context of what is a reasonable amount of time.*
> *[Life & health sciences researcher]*

In summary, researchers across disciplines are analyzing big datasets using machine learning and other data science methodologies. While there is a growing consensus around the programming tools to use for analysis, researchers exhibit varying preferences related to local and cloud computing options. Challenges persist in terms of storage, computing power, and technical support.

RESEARCH OUTPUTS

The Research Outputs theme refers to a dissemination, publication, communication, or sharing of research (including data/code) to people other than those participating in the research project/study.[10]

Channels

By and large those interviewed disseminated their research findings in traditional ways -- namely, papers, conference presentations, and posters. Multiple researchers mentioned social media (in particular, Twitter) as an increasing method of sharing, promoting, and discovering research in their given discipline. Some researchers were more engaged and excited about social media use than others -- as one researcher said: “social media is a hit or miss for me.” Several researchers also published findings and outputs from their research as blog posts or built websites as products of their research. This was particularly common among those interviewed in arts & humanities and social sciences.

[10] Definition inspired by The University of Auckland’s *Research Definition & Research Outputs: System Categorisation Guidelines*: https://www.auckland.ac.nz/en/about/the-university/how-university-works/policy-and-administration/research/output-system-and-reports/research-outputs--definition-and-categories.html

For those researchers interviewed in the areas of life & health sciences and social sciences, there was also significant discussion about the level to which the initial dissemination of research findings was prioritized to external partners. Whether this came in the form of workshops or policy development, the prioritization of disseminating to an external research partner before publishing in more traditional ways was a trend in responses. As one life & health sciences researcher put it:

> *...I'm typically doing the research in partnership with the government or with large organizations of one sort. And the dissemination process is usually first with them. Our partner/client. Then in the academic literature or at a minimum simultaneously, but always with advanced warning. That's the number one. I see that the dissemination of that work is… These tend to be very large organizations, so there's significant dissemination within the organization in an attempt to create policy for that organization. It is then disseminated also to the academic literature, but also through associations of organizations like theirs. [Life & health sciences researcher]*

A final mode of dissemination that was unique is that of public safety updates. One researcher developed a mobile app that the general public can use to be informed when a given event -- or set of events -- occurs. In addition to the more traditional academic methods of disseminating findings, this particular instance provided a good example of how research outputs and methods of communication of research findings can be quite varied with the potential to have a more immediate impact on the lives of everyday citizens.

Data & code sharing practices

Data and code sharing were consistently supported across all those interviewed, due to ethos or due to requirements. Various federal agencies (such as the National Institutes of Health and the National Endowment for the Humanities) were cited as having requirements; however, many researchers mentioned that data and code sharing (along with open access) were tenets of their research communities. A physical sciences & engineering researcher stated that their research consortium has an actual data policy published on their website that formally communicates how they will share the data and code from their research. The same researcher noted:

> *We have a principle of open access, so everything we do is also put on the arXiv and we even have our own website where we put all our results. That's where every single paper is. That's one of the things that was good because of the scale of these, we were actually stating that and to put some pressure on the journals because they wanted to publish our papers to do it. [Physical sciences & engineering researcher]*

Another researcher stated:

> *Typically we are in the business of making our data open sourced. That way people when they are looking to develop new computational tools, they have examples or*

> *representations of datasets that are generated by next generation instruments. [Life & health sciences researcher]*

Not many researchers spoke to the role of data and code sharing and reproducibility though one researcher noted in regards to their research involving natural language processing (NLP):

> *This is where reproducibility is really important, and the fact that having open sourced data is really important. A lot of the ways that NLP works is by having some validation, per task or a validation dataset, where you can take some code that someone else has written, run it on that data, and see how it performs. [Social sciences researcher]*

This same researcher noted the effect of data and code sharing when it came to building community as well:

> *...by sharing the code and sharing the data, you build a scholarly community and a network. And like you said, there's a kind of reciprocity. In terms of if you're generous with what you're doing, then it generates that in others. Success is a very powerful thing, it's not just theoretical. [Social sciences researcher]*

There were a couple of comments acknowledging how the COVID pandemic affected data sharing with a social sciences researcher noting:

> *In some ways, COVID has encouraged more data sharing, but I don't know how long that will last or what that will look like. Initially all of these sorts of data were through relationships with people. [Social sciences researcher]*

Similar to observations made about obtaining secondary data, another social sciences researcher observed:

> *For COVID, they've created this bridge between public health agencies and [Platform Y companies, which] provide daily data to the government. Not just on how the population is moving around, but also information on [their individual customers] so they can help identify where the high risk clusters might be. They're sharing all sorts of data that I think prior to COVID would have never been shared. I think if it remains to be seen worthy, these looser standards will last, but I think overall it's been really helpful. [Social sciences researcher]*

While both researchers did not necessarily see this expansion of data sharing as a permanent change, it provides additional insight into how data sharing can benefit the public at large.

Data & code sharing tools

When it came to the tools used to facilitate data and code sharing, GitHub was by far the most used platform. Several discipline-specific repositories were mentioned (i.e., GEO, HEPData, Harvard Dataverse) as well as preprint servers (i.e., arXiv, bioRxiv). Perhaps most surprisingly, no mention was made of Dryad as a platform used for data sharing among those interviewed. This is particularly noteworthy because Dryad is the data repository for the University of California (UC) system and is free of charge to UC researchers. There could be an issue of size when it comes to sharing "big data" via Dryad since uploads are limited to 300GB for self-deposited items; however, it again highlighted an issue of awareness about campus resources available to support researchers in their work.

Due to the sheer size of data being shared, several researchers mentioned setting up standalone websites where interested people could download chunks of data through file transfer protocol (FTP) or other mechanisms. Globus was also mentioned by one researcher as a method used for sharing data as well as ingesting data during data collection.

Challenges

The challenges mentioned around data and code sharing were fairly consistent with challenges raised across all research areas -- big data or not. There were concerns from several researchers about "sharing hesitancy" based on how it might impact their research group by allowing external, perhaps more resourced groups, to move ahead more quickly. To this point, a physical sciences & engineering researcher stated:

> *We do not share all the records, we only share some records and the attributes which will not hurt our research, so we limit the information if we share it if we think that this is too competitive and some other group can use it and make something faster than we will do. [Physical sciences & engineering researcher]*

Another physical sciences & engineering researcher echoed this concern while also voicing another challenge about analyzing shared data raised by other interviewees as well:

> *And there's certain people who think that all our data should be public. But then there are others who are concerned about the fact that it's a huge amount of work to make the data and if you make it public in that way then you disencourage people to do that work. The second worry is that it's tricky to analyze the data. If you don't do it very carefully, you can find all sorts of things that turn out to be complete nonsense. The concern is that we would spend a huge fraction of our time dealing with that. When people do this and say "I found this new thing, I found that new thing." So that's one of the concerns. I don't know if I've made up my mind here. I see it as an issue and I see both sides. [Physical sciences & engineering researcher]*

There was a concern among some of the interviewees about reuse of shared data when users might not have the background knowledge necessary to properly interpret the data.

A particular challenge related to big data most notably came in regards to the issue of size. As one researcher noted: “There are repositories for imaging data, but they limit to a few gigabytes, or tens of gigabytes. When you’re in the 10’s of terabytes or hundreds of terabytes, what do we do?” Additionally, this researcher discussed how this impacts sharing via journal-supported platforms as well:

> *Obviously the scale of our data is not worth the journal’s time. We’re happy to share the data, but they don’t want it. We make a statement in our publication saying that the data is available upon request provided that the requester provides a means for transfer, whether it be drives or a FTP site that we can push the data to. [Life & health sciences researcher]*

Other specific challenges around data sharing included one researcher mentioning data sharing being influenced by the relationship with the data provider -- in this case, the research group does not share data because it had taken a while to establish a relationship with the data provider, and the research group do not want to share anything that might jeopardize the relationship. Another researcher discussed not sharing human subjects data because they had explicitly written into the consent form that data would not be shared. Lastly, a physical sciences & engineering researcher talked about how certain research outputs (preprints in this case) were not viewed in the same way as other research outputs, such as papers in peer reviewed journals:

> *I do know that people don’t get credit, in my field, for publishing in an archive. When we’re doing research this year and we’re looking at people’s records, the question is what kind of journals are they publishing in? Is it that people want to see papers in Science and Nature still? People want to see publications in what are considered the top journals in [this discipline]. The fact that you’re publishing in a reputable journal means a lot when it comes to evaluating the quality of somebody’s work and somebody’s product. So I don’t see preprint archives taking over from journals anytime soon. But some people… What you can do is you can put it in an archive while it’s being reviewed, then the archive record points back to that. So people are doing that, but I haven’t seen people not proceeding to publish it in a journal. [Physical sciences & engineering researcher]*

Campus services & support

Not many campus services were explicitly mentioned in regards to research outputs. The Library was mentioned by a life & health sciences researcher around data sharing: “We went to the library a couple of times and we’ve chatted about ways in which one could better share their data and some strategies to back it up. I’m a huge advocate of the work that they do and their efforts.”

Another researcher noted when it came to campus services:

> *I definitely do reach out. I don't like to work or operate in a vacuum. But at the same time, I want to be realistic in terms of expectations because I don't expect everybody to solve everybody else's problems. Rather, these are tools and resources that can help point to the directions. For the low level stuff, for sure campus resources are the first place to start. For anything the next level, it would be amazing if the campus had the resources to train someone or develop a practicum or curriculums based around some of these real world things as opposed to actual courses. [Life & health sciences researcher]*

In general, as mentioned in previous sections, it seems that continuing to promote and showcase available resources to campus researchers is the most helpful action.

COLLABORATION

The Collaboration theme refers to the researcher's participation in and use of teams and networks in carrying out the research. The collaboration can be with other researchers in the field, across institutions or disciplines, and/or it can be with post-docs, graduate students, undergraduates, and other research staff.

Within & between institutions

Interview for interview, every single respondent had research collaborators. It was almost never "I do this" but "we do this," "my team does this," "my students do this," and so on. Where big data is concerned, the research scope is generally too large to be carried out by the mythic lone figure toiling away in their study. Given that collaboration is the norm, the nature and extent of it is of interest.

Outside of their immediate research teams and outside of Berkeley, there were many accounts from those interviewed of projects that are national and international in scope; one project involved 3,000 researchers (about 50 of them at Berkeley) in 200 countries. Several of the national-level collaborations have come about because the Berkeley researcher was studying a phenomenon of scale and significance to their field, so that pooling resources, both financial and intellectual, made sense. Often a collaboration arose in the process of finding, acquiring, and using open and/or secondary data. Nevertheless, more than one respondent mentioned that it would advance Berkeley's research agenda for researchers to have more connections with other faculty and researchers on campus, and questioned how communication channels might be enhanced to achieve that.

> *I'm just not seeing the level of faculty collaboration across departments that you would think would be happening in this day and age… The whole movement, to me, is led by the students, which is great and they're wonderful, but I think there needs to be more transformation at the faculty level, together with more interdisciplinary collaboration. I have a better relationship with folks across the world than with faculty who work on our own campus. [Social sciences researcher]*

Interdisciplinarity

Recent years have seen an increase in interdisciplinarity in research, with funders increasingly looking for cross-disciplinary projects to fund. The evidence for this kind of activity at Berkeley was mixed among those interviewed. Berkeley researchers reported that there are some natural partnerships for their activities in outside departments, generally within the same division, which arise around the use of common datasets, or curricular/training offerings in a partner department that can be of benefit. More centrally, many researchers must cross disciplines in order to draw on the big data methods needed to answer their questions; more than one interviewee noted that using those methods was the only way to visualize their phenomenon of interest. The advantages of sharing data, training, and methods are part of the drive towards interdisciplinarity, and technology and software advances available outside a researcher's home department may also draw them across disciplines. While some noted that their research was interdisciplinary at its core, it was more common that the crossing of disciplines happened in order to use the approaches and tools of data science, from a position which is solidly grounded in single-discipline intellectual concerns.

Researchers showed interest in the possibility of interdisciplinary collaboration, but also expressed concern at having to learn and keep abreast of an entirely new discipline, and were not sure how the current campus environment (let alone research funding structures) would support interdisciplinarity.

> *Data science should be by definition about interdisciplinarity, and where that interdisciplinarity will come from on our campuses I think is a question that nobody has resolved well. Even Berkeley, I think, was on the verge of national leadership with its idea of a cross-cutting division that would have dotted lines to [hard sciences] and also bring in faculty from across campus. But now it looks like it's going to be more of a college which then will make another silo on this. That's the question that I would love to get the answer to: how do we get successful interdisciplinary collaborations? [Social sciences researcher]*

Team composition & division of roles

An important part of the academic mission of Berkeley is to train the next generation of researchers, which meant that most of the researchers interviewed had teams that included faculty, postdocs, and graduate students. These teams often also included full-time research staff and/or undergraduate students. In addition, several researchers reported directing multiple projects. Not surprisingly, the composition of a team and the varying skills of its members often determined the division of roles.

Most researchers reported managing and directing the research studies, and making sure their team members receive the experiences and professional development needed to be competitive in the job market. Mentoring the next generation of researchers, and creating a vibrant learning community was a central concern, along with conducting the research itself. Readiness for the academic job market is particularly important for postdocs and graduate students, which means that, in addition to doing their research activities, they are simultaneously learning software and methods, and disseminating their work as widely as possible, with social media being an important venue. Postdocs are considered desirable because they arrive at Berkeley with skills to do some of the data science methods needed for big data projects, and do not have competing classes or other responsibilities. However, graduate students are often considered the lynch pins -- they are expected to attend training sessions, bring the latest content back to the team, teach the PI/head researcher, and also often to teach undergraduates on the team. They are teaching both up and down the team hierarchy, and as such play an integral role on the research team.

> *That would be the way to create a successful project, is to create a learning community that ramps up on its own and has exponential growth in terms of skill and sophistication because everybody learns from each other. The smarter the people and more knowledgeable around you people become, the more knowledgeable you become. [Social sciences researcher]*

Professional networking & keeping up with trends

Reported networking strategies and avenues were fairly uniform among respondents, whether formal or informal. Almost all mentioned at least two of the following activities, with much overlap across disciplines: conferences and meetings, journals and other publications, email discussion lists and association newsletters, professional training, and Twitter.

Knowing what others in their discipline are doing through publications and data and code sharing can lead to collaboration, but it can also lead to tension between wanting to share vs. wanting to protect huge investments in years of work. As one researcher noted of others in their discipline outside of Berkeley, “they are our friends, and our competition.” Another networking and collaboration avenue for postdocs (and sometimes graduate students) is maintaining connections with their previous institution; in fact, for some postdocs the boundaries between the two regarding collaboration and use of resources such as computing can be undefined. In general, the networking reported by researchers is discipline-specific.

The same mechanisms were mentioned in terms of keeping up with the literature and latest developments in the researchers’ fields; in this case, however, collaborations on projects can expand the resources that a researcher accesses, because there are multiple colleagues checking all the different possible sources. The pace and scope of scholarly publishing has become so vast that it is hard for any one person to keep up, and this is especially true if a researcher is trying to follow both domain-based content and data science content. Some groups have regular meetings with mini-presentations, using tools such as Google Docs, project management tools (e.g., Asana), and Python notebooks to make sharing easier.

Pressures of time and volume of reading combine to make Twitter an often-mentioned and attractive option for researchers. When colleagues tweet the main points of their work, researchers feel they have gotten the essence of an article without taking time that they do not have to read and digest it. It maximizes literature perusal efficiency. Google Scholar is another heavily used tool (although no one mentioned setting up search alerts to notify them of updates of interest, either in Google Scholar or library-provided databases). Blogs and word of mouth (despite there being no in-person water cooler chats during COVID) were also occasionally mentioned.

> *Keeping up is hard. So I don't know. I constantly feel behind. So that's part of it, is acknowledging that I'm constantly behind. But I do try to… What's new? What have you found out there? What's relevant? But honestly COVID has been so weird that some things have fallen by the wayside. [Life & health sciences researcher]*

Challenges

The big data researchers interviewed mainly experience the problems of any researcher, but the sheer size and complexity of their data and analysis needs can add layers of responsibilities and challenges. Perhaps the most concerning for Berkeley as a whole is that researchers do not want to be vulnerable to resource problems, the central of which is computing resources. The impact of the reported lack of campus computing infrastructure and support can lead groups to acquire and use their own dedicated servers and solutions. This in turn leads to many small computing silos and less collaboration. At the same time, researchers expressed the sense that if resources were pooled, more support and technological solutions could be within reach, and hoped that Berkeley finds campus-wide solutions that will rival competitor universities. It was also noted that Berkeley's process can be cumbersome and licensing requirements strict compared to other universities.

The environment of scholarly communication offers its own challenges. There was a mix of opinions among the interviewees about the question of data and code sharing. To be a trailblazer in their field, a researcher may have to shoulder a big burden and construct and/or find data no one has before. It may be that the data are in plain sight and they must find creative ways to access that needed data—because others do not want to share. They in turn may want to protect their own investment, yet make their work discoverable and accessible—it can be difficult to find the balance among these competing perspectives and concerns.

Furthermore, a number of respondents noted that they are directing multiple projects, meaning that for them, project management competency is an additional issue. In fact, some researchers expressed interest in campus assistance with provision of and training in project management tools such as Trello and Asana. One respondent also noted that research communities can be prey to turning inward in such a way that it is difficult to ensure equity and inclusion for a range of research projects and researchers. Geographic scope also creates its own challenges for those researchers who collaborate outside Berkeley and the United States; variations in communication tools, governmental oversight, ethical and cultural concerns, and even time zones can wreak havoc with the smooth running of a project.

> *There are also challenges due to the fact that we're a distributed collaboration. So we have all these different people in different places all around the world. So we have*

> *something called the grid which is a distributed network of computers all across the world. [Physical sciences & engineering researcher]*

TRAINING

The Training theme encompasses a broad range of activity that leads to researchers and their team members gaining the competencies needed to carry out their research, and potentially to keep up with trends in their field. Training can take place in a formal setting (courses, workshops, self-paced learning) or informally (picking it up as you go along, learning from colleagues or in intermittent bursts). Related to training is the question of how a researcher can gain a needed competency or get support and consultation for an immediate issue.

Training of faculty & researchers

A substantial number of researchers reported that they have no formal training in big data methods. This is especially true if they have been doing research for a long time, since their doctoral training, which would be a natural avenue for formal training, happened many years ago. The recency of the availability of computing power and concomitant development of big data methods means that these early big data adopters had to learn as they went along. For some this happened incrementally: “it’s hard to know when you transition from data to big data.”

The training options which researchers did mention align closely with their professional networking activities and the ways in which they keep up to date in their fields. Domain knowledge is central, but those interviewed expressed that there is an increasing need for training about big data tools. They mentioned webinars and online courses, which they find through meetings and conferences, journals, and other publications; Coursera was mentioned for learning the fundamentals of an unfamiliar tool or method. Needing to consult about or solve a particular problem or challenge may necessitate training, while teaching a class may provide motivation to learn new content (more than one interviewee mentioned being “a week ahead of the students”). Similarly, writing a paper which requires background reading was another way to pick up new material; and, “of course Google is my best friend.”

> *One of the luxuries of being a faculty member that is tenured at a research institution like Berkeley, and having undergraduates, master’s students, and doctoral students, is that they help you stay abreast of the literature. Working with them to do systematic reviews. That is a great way of staying abreast of the literature. Similarly if you update your course readings to make sure that you’re accessing the recent literature for the classes you’re teaching. [Life & health sciences researcher]*

One issue is that researchers may need training about the research, but they also must become at least conversant with tools such as GitHub, the variety of data repositories, and other issues related to data sharing, reproducibility, preservation, etc. A challenge is that “you become out of date really quickly.” Also, a busy researcher may not be able to take on learning a new software package. Often they are supervising and managing at the project level (for one or more projects), for which they need to know the concepts but less so the details of the tools used.

> *I do think these data science tools are becoming more important to social scientists... You can get the basics of how the tools work in a D-Lab workshop, take a couple sessions on text analysis. But I think to really understand what to do with it, you have to spend a little more time reading and thinking through and talking through examples in ways you use it, the way you would in a normal semester course… Machine learning is the same thing. We could have the neuroscientists and the social scientists all in the same course, but people don't learn that way, right? They need to learn A. starting from what they already know and adding this on, and B. they need to learn it through the types of examples and problems that they're going to work with. [Social sciences researcher]*

> *If I were to take a course or learn about something, I think I would rather learn something about how groups use Slack, or something like that, than some great algorithm for doing something. It's the nexus between people and technology to give people superpowers with the technology. It gives groups superpowers with the technology. [Social sciences researcher]*

Training the teams

Several researchers noted that they learn a great deal from their graduate students, and in fact expect them to take on the role of keeping abreast of the latest tools and methods. Many interviewees reported sending their graduate students to training sessions and workshops before they go themselves.

> *My training is my graduate students, it's really that straight forward. Graduate students do go and do a variety of different training. I send them to… workshops on accessing and using data... So that's how I handle this, I get my students to go to these various things… I get alerted about these things through our email lists, then I send them off and we see what they learn and then I learn what seems to be useful. That's basically the approach. [Physical sciences & engineering researcher]*

Several acknowledged that this can create disconnection between PI-level researchers and the "next generation." On the other hand, mentoring and training postdocs and graduate students, and launching them in their careers, is seen as a central activity of the research process, so that researchers invest heavily in their students' professional growth even while knowing they will eventually leave.

Not surprisingly, most of the researchers interviewed need to consider the composition of their teams and division of roles regarding prior education, interests, skills, and career goals, and then train accordingly. At the same time, the demands of the research itself must be satisfied, and it may be that a researcher will avoid, for example, hiring undergraduates for complex tasks that may take them up to a year to learn given no prior experience. One researcher made the observation that we might do better to teach data literacy and research-related skills in high school so undergraduates could arrive better prepared for the research experience.

Training & other resource needs

Researchers are acutely aware of skills and competencies needed by their various team members in the rapidly changing big data landscape. The need for training that focuses on relevant big data tools, skills, practices and subjects was often mentioned by interviewees, and for several, the starting place is the data science curriculum.

> *I have a bit of a love-hate relationship with the data science curriculum in that I think it's trying to reach a really large audience but at the same time, there are some skills that, at least from the data science students I've worked with, seem to be jumped past. I would almost maybe start with something a bit more core instead of just 'Hey we're just gonna write a code.' I get it, that is much more exciting to do and it's a lot easier to teach. Whereas if you're starting to really give each student the independence of setting stuff up on their own… That requires a lot more faculty hours and hands-on experience that doesn't lend itself as scalable as some of these courses are looking to be. [Physical sciences & engineering researcher]*
>
> *The other piece that's missing in our curriculum is data visualization. There's a great undergrad class on data visualization and there are a number of different courses on data visualization on campus. But I find that students, for some reason, are still putting together bar charts that don't make sense or don't have labels on the y-axis, these fundamental things. And I see this all the time with our data science students. They're writing in Python and their Jupyter notebooks, they're producing all these charts, but nobody truly taught them about how to communicate that data and how to get rid of the chart junk. [Social sciences researcher]*

In addition to needing an enhanced practical curriculum for the students who will be assisting in the research, a central concern is keeping up to date with the shifting needs for training in methods, techniques, software, and hardware with which to analyze big data. Interviewees suggested a broad and deep list of training topics, offered through a variety of channels:

Training Topics

- Data basics such as finding and cleaning data, data organization, and security and encryption
- Foundational statistical concepts (to understand the statistics needed for big data)
- Project management techniques and tools
- General coding concepts (but coding alone is not enough) and working in virtual environments
- R, Python, Jupyter Notebooks
- Managing very large datasets (using computing clusters, etc.)
- Machine learning and natural language processing

- Reproducibility
- Finding good materials online
- Predictable pitfalls
- Data visualization

Training Channels

- D-Lab, Library, GIS, and Research IT/RDM Program courses and consulting
- Statistics consulting
- Classes, boot camps, online tutorials, workshops, community learning
- Campus and disciplinary networks for solving various issues
- Practicums and shadowing

Challenges

Beyond a shadow of a doubt, researchers reported that the main challenges for training are time and money. As one researcher noted, “I would love to avail myself of such training, but I don’t fool myself into thinking that I’m likely to have the time to do that anytime soon.” Furthermore, trends in tools and methods are changing so quickly it is hard to know how to keep up, so it is challenging to know where and in whom to invest resources.

Additionally, campus communication channels may be too distributed, and researchers may not know about training and support opportunities that could help them -- several times the project interviewers informed interviewees about resources that they could use in their research. Researchers may be trying to balance the tension between domain knowledge and data science techniques. This may involve training needs that are interdisciplinary and cannot be met with campus resources. The material itself is extremely complex; as one interviewee noted, how can you get your understanding around data which is expressed in 5 dimensions and has a size of 5 terabytes? These challenges may lead researchers to hire team members who already have data science training and experience rather than those without it. Not only does the researcher not have to train those team members on methods, but they also have less worry about privacy and security breaches that may take place with inexperienced students.

Finally, not everything can be trained. Researchers need team members who can think critically about data and analysis -- what can and do the data actually show, and what they can’t and don’t show. This level of judgment is highly prized.

> *Part of what’s going on is that the faculty has such different computing backgrounds. It’s such rapid generational change. [I was planning to] teach a graduate class for next semester and I thought one of the things that they might do was to do one of these replication projects where we take a paper and replicate it. But even in something like R, there’s 2 different dialects now. And I use an old dialect instead of the new dialect. So in order for me to teach them, I’m not going to use the new dialect because I think that I should teach them what I use and what I find that works...And a lot of the great minds at*

> *the university are still thinking in old ways. It doesn't make them any less great minds, but they shouldn't be wasting their energy learning some modern computer language, and that's okay. [Social sciences researcher]*

> *In our department, we worked with all of the faculty that teach quantitative courses and had them all change the curriculum to Python… There's a whole team at the data science division that goes in and builds modules for classes… To integrate across campus, you have to be speaking the same languages. [Moving to Python has] really helped and now our students can take the basic methods classes where they're learning Python and are able to take classes across campus. It's a really transformative time. [Social sciences researcher]*

BALANCING DOMAIN VS DATA SCIENCE EXPERTISE

As mentioned above, big data researchers have many of the same experiences as their disciplinary colleagues doing research at different data scales. However, one notable difference is that due to the methods needed to analyze datasets of enormous complexity and size, big data researchers must add data science methods and tools to their armamentarium. This is not a trivial task, and a number of interviewees from a range of disciplines reported that they and their fields are constantly seeking a balance between domain knowledge and data science expertise in their work. In fact, they may be seeking tools that do not exist yet, which would allow them to seamlessly navigate the analysis and computing needs of big data from within their disciplinary norms rather than, as they must do now, hiring people who can help them navigate.

> *How do you even ask questions from the data? We very often need those data scientists: for example, there's a stand-alone package that allows me to run [certain domain-specific analyses], and there's a different package I can run that [supports the machine-learning analyses]. But I don't think there's a stand-alone package that will allow me to do both together at the same time. In order to do this, I'm going to need a statistician to help me code this analytic complexity. [Social sciences researcher]*

For some disciplines, data science methods are more built-in than for others. Researchers noted gaps in providing data science methods instruction in social sciences and particularly arts and humanities curricula, while at the same time stating that it makes little sense to expect all data science instruction to be handled solely within the disciplinary curricula. Coursework alone is not enough to prepare students to join research teams and be effective big data analysts; hands on working through real examples, with knowledge building cumulatively, is suggested as the best way, but is also time and labor intensive for the researchers (and postdocs, and even graduate students) training the inexperienced team members. This can set up a potentially concerning dichotomy, wherein students trained in data science methods are highly valued for those skills and hired by researchers for the crucial reasons of time and efficiency, leaving other students to focus on domain training without gaining the experience they might need to perform their own research and compete in a tight job market. Some researchers are less concerned about this divide than others, and it may be possible to achieve a happy medium if that is valued in the department/discipline.

> *Actually the undergrads are great, they are the best ones. Data science, and if you want even better computer science, undergrads are just out of this world. We couldn't do it without them. And they know much more than our grad students. Our grad students often don't have the skills to do this, so we rely on the undergrad students… The grad students, particularly in our professional schools, are coming in to change the world, and they don't see it as a data task, so they're not as interested in big data… There's some that are interested and do try and get those skills and a lot of them are getting the data science certificate which is going across campus, but they're late to the game as opposed to our undergrads who are very well versed in skill and they're just looking for a topic. They just want a domain so we give them a domain. [Social sciences researcher]*

> *In my department we definitely have more nerdy and less nerdy students and I would say the difference is… It may be a little bit about ability, but most of it is where their interest lies. The student that is more nerdy feels that it is more in their interest to spend time on more technical projects. And the student who is less interested in that feels it's not worth their time. And I certainly don't think either one of them is wrong. If they're in the department, we're not going to be training them in data science. We're gonna be training them in [our discipline], and we're going to be showing them some things that may require some relatively sophisticated computational methods. [Social sciences researcher]*

Still, it can be challenging and even inefficient when team members trained in data science alone do not know to ask questions or direct analysis that are in line with disciplinary knowledge, which is the knowledge at the heart of the research questions being explored.

> *It is definitely a pet peeve of social scientists that a lot of people in the machine learning community don't have any "domain" knowledge that they're working with. Either just reinventing the wheel or coming up with strange interpretations of results. I think a constant source of tension in this space where we're trying to bring data scientists from CS or related fields together with people who work particular domains, it's the ports of the domain knowledge and how it's applied in the context of the analysis. [Social sciences researcher]*

> *I think that the working on public health during COVID has highlighted that issue for me because a number of modeling papers are out there now by people with no background in either biology or public health or epidemiology. And it's shocking because decisions are being made using those...But making these modeling claims and claims about the*

> *trajectory of the disease is not the right way [if] you mentally misunderstood things about how diseases are transmitted. [Life & health sciences researcher]*

In the early stages of a project, it may not be clear which data science techniques to use. Discovery of appropriate tools and methods may then involve training (or at least consultation); it can also involve creatively adjusting the resources and/or developing original techniques to meet the research need. Some researchers mentioned that it would help to be able to hire the technical expertise so as not to have to train team members from scratch just to be able to perform a particular needed function.

> *In health-related research, as I'm sure you're aware, especially now with the pandemic, data is the new oil. Everyone wants to strike oil. It's everywhere but you need some skill to get it. It's challenging. I work with a lot of the biostatistics trainees for that reason because they're highly competent in the technical aspects, but they're not studying the actual content. They're more interested in methods. [Life & health sciences researcher]*

Additionally, it can be hard to characterize who has data science skills from just looking at formal credentials, which has an impact on hiring research team members and sending team members out into the job market. Some people with formal training cannot handle real world analysis, and others with no formal training are highly skilled. The same domain vs. data science dichotomy plays out in academic hiring. Researchers noted that the domain knowledge is usually central, but while the data science skills may be needed for a productive research career, they can be considered somewhat suspect when a candidate is perceived as dividing their focus too much.

> *I don't know the answer but I think that part of the challenge is that a lot of the best work is taking place in the disciplines disconnected from people who call themselves data scientists. Yet it's in the air and people make creative applications. I'm sure the economists are doing lots of things. And the geo-engineers are doing lots of things. The climate scientists are doing lots of things. But they're not calling it data science. [Social sciences researcher]*

> *This is one of the clearly recognized issues. You have the domain side and you have the computer scientists and the statisticians. How do you bring these two together? It's interesting. I've really seen that become a problem in our field. We have people, whether it be graduate students or postdocs, who have that technical computer science or machine learning skill. I'm not talking about the hardware side of it, I'm talking about the software, algorithmic side of it, machine learning expertise. It's a real challenge for them when it comes to looking for a job because there's a lot of hesitation. Our own department has the same problem. We hesitate to hire somebody who has spent their time working in machine learning because the sense is that they've become very focused on the tools as opposed to the science. [Physical sciences & engineering researcher]*

RECOMMENDATIONS

There are arguably two main challenges that Berkeley faces as big data research becomes increasingly common. One challenge is that the range of discrete data operations happening all over campus, not always broadly promoted, means that it is easy to have overlapping services and resources -- and siloes. The other, related, challenge is that Berkeley has a distinctive data landscape and a long history of smaller units on campus being at the cutting edge of data activities. An example is SDA (Survey Documentation and Analysis), started at Berkeley, which still underlies the online analysis tool of the social science data repository ICPSR (Inter-university Consortium for Political and Social Research). How can these be better integrated while maintaining their individuality and freedom of movement?

An additional challenge, of course, is the relative lack of funding that public institutions receive as compared to private ones. Berkeley researchers do much with what they have, but with the rising prevalence of big data activity nationally and globally, they need to be poised to operate smoothly in a new environment. The Bay Area as a base of operations can be a challenge in itself, with cost of living outstripping salaries, and the vast army of tech workers often being able to command more than Berkeley can pay. Within its funding challenges, how else can Berkeley address this need?

Based on these realities and informed by the interviews conducted, the following recommendations are suggested for the University Library and campus partners to successfully support big data research at UC Berkeley moving forward:

1. Create a research-welcoming “third place” to encourage and support data cultures and communities.

Two central challenges that researchers expressed in navigating the big data landscape at Berkeley are communication (i.e., knowing what is going on with research on campus) and collaboration/interdisciplinarity (i.e., being able to have strong connections internally as well as externally). A related challenge is bridging the gap between data science and domain knowledge. To address these challenges, campus units should work to create a “data culture” on campus, which can infuse everything from communications to curricula, and which is timely in light of the new Data Science facility being planned as a “gateway” to campus, both literally and figuratively.

One way to operationalize this idea is to utilize the concept of the “third place,” first outlined by Ray Oldenburg. In this model, people have homes and workplaces, but also need informal public gathering spaces, such as coffee houses, markets, and often libraries. With Berkeley as the common workplace, the proposed data communities can be less structured, and welcoming of all kinds of research, including big data. This can happen in, but not be limited to, the Library, and can be in both physical and virtual spaces. Goals of these “third places” would include encouraging open exploration and conversation across silos, disciplines, and hierarchies while centering Justice, Equity, Diversity, and Inclusion (JEDI) as a core principle.

- *The University Library, in partnership with Research IT, conducts continuous inquiry and assessment of researchers and data professionals, to be sure our efforts address the in-the-moment needs of researchers and research teams.* This can be ongoing focused interviews, a structured survey, or a fuller research exploration such as this Ithaka project. Another possibility is to find ways to work within research teams, possibly with students or other "embedded" personnel, so as to create solutions from inside out, rather than vice versa.

- *The University Library, in line with being a "third place" for conversation and knowledge sharing, and in partnership with a range of campus entities, sponsors programs to encourage cross-disciplinary engagement*. This programming should be inclusive of a range of perspectives and disciplines, beyond the traditional STEM (Science, Technology, Engineering, and Math) focus. Potential examples: a series in which a range of researchers talk about their research, aimed at undergrads, a scholarly publishing mixer aimed at grad students, or ideally something dreamed up by students and researchers themselves. This initiative can also include working to make connections and encourage synergies between domain experts and data science experts. Possibilities in this area could include using the "Collaborating with Strangers" model[11], meetups, exhibits, panel discussions, programming during Love Data Week[12], a research guide, etc.

- *Research IT and other campus units institute a process to explore resource sharing possibilities across teams of researchers in order to address duplication and improve efficiency.* Resource sharing can include staff support as well as physical resources and equipment, but needs to take into account the frequency and urgency of research team needs. This calls for improving campus-wide communication about research activities and services, leading to a common understanding of and increased transparency related to university-wide research landscapes. It may first be necessary to do background research on what approach would be most effective.

- *The University Library partners with the Division of Computing, Data Science, and Society (CDSS) to explore possibilities for data-dedicated physical and virtual spaces to support interdisciplinary data science collaboration and consultation*. Proposals for the Center for Connected Learning in Moffitt Library and The Gateway building may offer opportunities for interactive spaces where collaborative programs could be hosted. This will be a long-term process, and will be complex given the constraints involved, but having presence in each others' spaces and programs will allow for the day-to-day interactions that lead to closer collaboration and integration of services.

- *A consortium of campus entities develop a data policy/mission statement, which has as its central value an explicit justice, equity, diversity, and inclusion (JEDI) focus/requirement.* Additional values can be directed towards avoiding silos, incentivizing interdisciplinary work, and potentially incorporating (big) data literacy throughout curricula.

[11] For more information about "Collaborating with Strangers" model, please see: https://www.alastore.ala.org/content/collaborating-strangers-facilitating-workshops-libraries-classes-and-nonprofits

[12] For more information about UC Love Data Week, please see: https://uc-love-data-week.github.io/

2. Enhance the campus computing and data storage infrastructure to support the work of big data researchers across all disciplines and funding levels.

Researchers expressed gratitude for campus computing resources but also noted challenges with bandwidth, computing power, access, and cost. Others seemed unaware of the full extent of resources available to them. As a result, some research groups and collaborations have developed their own storage solutions and computing clusters rather than rely on or contribute to campus resources. While this makes sense for large international collaborations, how does campus ensure that its computing and storage options meet the needs of smaller collaborations and encourage them to leverage those resources?

- *Research, Teaching & Learning and the University Library partner with Information Services & Technology (IST) to conduct further research and benchmarking in order to develop baseline levels of free data storage and computing access for all campus researchers.* To be clear, campus currently offers a faculty computing allowance and free unlimited cloud storage through Google Drive and Box. However, as new limits are imposed on those cloud solutions at the same time that storage and computing needs continue to grow, it is important to establish new baselines. Further research will be essential to establishing those baselines. The University Library should work with Research IT to further investigate disciplinary needs for data storage and computing in order to understand disciplinary differences. In addition, Research, Teaching & Learning should continue its regular research and academic engagement benchmarking projects and transparently share the results. Some researchers opined that UC Berkeley should be setting the standard for big data research infrastructure, but they perceived that other institutions had more robust computing resources. Sharing the results of regular environmental scans would confirm or refute those perceptions. In addition, the results could be used to define aspirational goals for cutting-edge infrastructure to support the campus vision for data science. Once new baselines are established, there should also be competitive and transparent pricing for researchers who need more than the baseline.

- *Research IT and the University Library work with campus to develop further incentives for funded researchers to participate in the Condo Cluster Program for Savio and/or the Secure Research Data & Computing (SRDC) platform.* Many researchers mentioned using Savio, but others built their own servers and clusters without joining the program. Increasing participation would expand high performance computing capacity and access for all researchers while preserving priority access for contributors. What incentives might help grow that program?

- *The University Library and Research IT partner to develop and promote streamlined, clear, and cost-effective workflows for storing, sharing, and moving big data*. The UC Berkeley Storage Changes Transition Team -- a cross-unit collaboration -- is already developing alternatives and working on plans to transition the largest users in preparation for unlimited storage ending. At the same time, campus service providers (e.g., librarians, Research IT consultants) need to develop an understanding of the options to provide effective consulting to new and existing researchers. This includes advice and training on data curation for researchers so that they can avoid storing unnecessary data as campus sees further restrictions on storage limits. In addition, some researchers alluded to the challenges inherent in moving big data between computers, and many of them seemed unaware of tools such as Globus. Devoting resources to making Globus an option for more researchers, and then developing and promoting effective workflows using these tools would benefit researchers.

3. Strengthen communication of research data and computing services to the campus community.

In the interviews, researchers directly or indirectly expressed a lack of knowledge about campus services, particularly as they related to research data and computing. In light of that, it is important for campus service providers to continuously assess how researchers are made aware of the services available to them.

- *The University Library partners with Research IT to establish a process to reach new faculty across disciplines about campus data and compute resources.* There have been existing efforts between the Research Data Management (RDM) Program (a Library and Research IT partnership) and Research IT at large to connect with “new” (within 2 years) faculty members in order to make them aware of campus research data and computing services. Moving forward, this outreach should be formalized so these providers can consistently connect with new faculty as they join UC Berkeley. As it relates to big data researchers in particular, a challenge for campus services is becoming aware of the technical needs of researchers early enough to provide support as well as advocate for needed support. By meeting with new faculty early in their time at Berkeley, campus service providers can better determine the data and computational needs of researchers in order to guide them most effectively in establishing their workflows.

- *The University Library partners with Research IT and CDSS (including D-Lab and BIDS) to develop a promotional campaign and outreach model to increase awareness of the campus computing infrastructure and consulting services*. Researchers often did not know about the full range of resources and services available to them, and there is a need to better promote systems and tools such as Savio, Globus, and the Secure Research Data Compute (SRDC) platform. A coordinated campaign among several campus partners might reach more faculty and researchers successfully, and ongoing work to create a campus data services mapping and referral resource (led by the Library in partnership with Data Peers Consulting) will be another way to communicate campus data service providers and areas of expertise moving forward.

- *The University Library develops a unified and targeted communication method for providing campus researchers with information about campus data resources -- big data and otherwise*. Taking inspiration from a proven model such as that used at the University of Illinois (i.e., their Data Nudge program), the Library Data Services Program (LDSP) should consider developing a similar method to reach researchers about available campus data resources. There are various partners on campus with whom the Library can partner to develop and implement this strategy including Research IT with D-Lab and BIDS as part of CDSS.

4. Coordinate and develop training programs to support researchers in "keeping up with keeping up."

One of the most-cited challenges interviewees stated in terms of training is that of keeping up with the dizzying pace of advances in the field of big data which often necessitates learning new methods and tools. Even with postdoc and graduate student contributions, it can seem impossible to stay up to date with needed skills and techniques.

Accordingly, the focus in this area should be to help researchers to keep up with staying current in their fields. The Library in particular should support researchers by creating/helping them create alerts and feeds of information that are brief and to the point, and arrive regularly in a preferred format. This could range from customized searches, and brief alerts from the Library such as those mentioned in the communication section above, to webinar series and online guides (some of which are already in place).

- *The University Library addresses librarians/library staff needs for professional development to increase comfort with the concepts of and program implementation around the research life cycle and big data*. This can include library-offered workshops as well as online offerings.

- *The University Library's newly formed Library Data Services Program (LDSP) is well-positioned to offer campus-wide training sessions within the Program's defined scope, and to serve as a hub for coordination of a holistic and scaffolded campus-wide training program*. Training opportunities can be offered by the Library in response to requests and current trends; general areas include (but are not limited to):
    - machine learning and natural language processing
    - text mining
    - data processing
    - availability of data collection tools such as Qualtrics and REDCap
    - qualitative and mixed methods (although not generally used with big data)
    - ethical use of big data tools and methods (such as artificial intelligence)
    - image management and digitization

  The LDSP may also want to consider offering researcher-suggested training that is broader in scope, but not offered elsewhere on campus, such as project management and organization.

  With its cross-disciplinary status, the University Library is in a unique position to collate and curate a list of offerings for targeted big data training from relevant entities across campus. Such a holistic approach, with scaffolding of competencies being addressed, can include partnering with and/or promoting offerings by entities such as Research IT, Data Peer Consulting, the D-Lab, and CDSS at large.

- *The University Library's LDSP, departmental liaisons, and other campus entities offering data-related training should specifically target graduate students and postdocs for research support*, recognizing and addressing their central roles on research teams. Targeted offerings could include training around networking/social media/general research dissemination, and scholarly publishing, to name just two.

- *CDSS and other campus entities investigate the possibility of a certificate training program -- targeted at faculty, postdocs, graduate students -- leading to knowledge of the foundations of data science and machine learning, and competencies in working with those methodologies*. The certificate will support training in how to balance domain knowledge and research needs with intensive data science methods and analysis experience, for those whose research involves these methods but who are not specifically CDSS-affiliated.

CLOSING THOUGHT

This report provides insight into the current state of big data research and use of data sciences methodologies at UC Berkeley. As the campus moves forward on initiatives to develop data science infrastructure and communities, the research team hopes that this report's findings and recommendations inform and inspire the vision of campus leadership, particularly as it relates to the future of big data at UC Berkeley.

> *"The tsunami is coming. I sound like a crazy person heaping warning, but that's the future. I'm sure we'll adapt as this technology becomes more refined, cheaper… Big data is the way of the future. The question is, where in that spectrum do we as folks at Berkeley want to be? Do we want to be where the consumers are or do we want to be where the researchers should be? Which is basically several steps ahead of where what is more or less the gold standard. That's a good question to contemplate in all of these discussions.*
>
> *Do we want to be able to meet the bare minimum complying with big data capabilities? Or do we want to make sure that big data is not an issue? Because the thing is that it's thrown around in the context that big data is a problem, a buzzword. But how do we at Berkeley make that a non-buzzword?*
>
> *Big data should be just a way of life. How do we get to that point?"*
> *[Physical sciences & engineering researcher]*

ACKNOWLEDGEMENTS

The research team would like to express their gratitude to several people who contributed to this project. Salwa Ismail served as the project sponsor, signing UC Berkeley up for the project, assembling the team, and sharing advice when needed. Danielle Cooper, Dylan Ruediger, and Rebecca Springer at Ithaka S+R developed the interview protocol and provided training on the protocol, recruitment strategies, and transcript coding principles. Anthony Suen participated in these training sessions, shared relevant insights from the Division of Computing, Data Science & Society, and contributed to early planning stages and discussions. Celia Emmelhainz offered generous and valuable insights on qualitative methods and software, and Tara Van Hoorn earned our eternal gratitude for skillfully transcribing the interviews. Danielle Cooper from Ithaka S+R, our Ithaka study cohort partners at UC San Diego (i.e., David Minor, Stephanie Labou, Reid Otsuji), and our UC Berkeley colleagues -- Salwa Ismail, Anthony Suen, Rick Jaffe, and Anna Sackmann -- all provided valuable feedback on the report.

REFERENCES

Association of Research Libraries. 2019. “The Data Science Revolution.” *Research Library Issues*, no. 298.

Baker, Sarah Elsie, and Edwards, Rosalind. 2012. “How Many Qualitative Interviews Is Enough.” Discussion Paper. National Centre for Research Methods. http://eprints.ncrm.ac.uk/2273/.

Burton, Matt, and Liz Lyon. 2017. “Data Science in Libraries.” *Bulletin of the Association for Information Science and Technology* 43 (4): 33–35. https://doi.org/10.1002/bul2.2017.1720430409.

Burton, Matt, Liz Lyon, Chris Erdmann, and Bonnie Tijerina. 2018. “Shifting to Data Savvy: The Future of Data Science in Libraries.” http://d-scholarship.pitt.edu/33891/.

Corbin, Juliet M., and Anselm L. Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Fourth edition. Los Angeles: SAGE.

Creswell, John W. 2002. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Upper Saddle River, N.J: Merrill.

Creswell, John W. 2007. *Qualitative Inquiry & Research Design: Choosing among Five Approaches*. 2nd ed. Thousand Oaks: Sage Publications.

Guest, Greg, Arwen Bunce, and Laura Johnson. 2006. “How Many Interviews Are Enough? An Experiment with Data Saturation and Variability.” *Field Methods* 18 (1): 59–82. https://doi.org/10.1177/1525822X05279903.

Li, Chan, Susan Edwards, Mohamed Hamed, Tor Haugan, and Becky Miller. 2019. “UC Berkeley Library Faculty Survey 2018 Report.” https://escholarship.org/uc/item/9p90t88d.

Maxwell, Dan, Hannah Norton, and Joe Wu. 2018. “The Data Science Opportunity: Crafting a Holistic Strategy.” *Journal of Library Administration* 58 (2): 111–27. https://doi.org/10.1080/01930826.2017.1412704.

Oliver, Jeffrey C., Christine Kollen, Benjamin Hickson, and Fernando Rios. 2019. “Data Science Support at the Academic Library.” *Journal of Library Administration* 59 (3): 241–57. https://doi.org/10.1080/01930826.2019.1583015.