UC San Diego
Ithaka S+R Research Study: Supporting Big Data Research

Stephanie Labou
Data Science Librarian
UC San Diego Library

Reid Otsuji
Data Curation Specialist
UC San Diego Library

David Minor
Program Director, Research Data Curation
UC San Diego Library

October 8, 2021

The Library
UC SAN DIEGO

Table of Contents

Executive Summary

During Winter 2020 - Summer 2021, the UC San Diego Library participated in the Ithaka S+R multi-institutional research study “Supporting Big Data Research”. The purpose of this study was to learn more about how researchers on the UC San Diego campus work with big data in their research and provide a set of recommendations to enhance and develop resources and services that will directly support and benefit this research. While definitions of “big data” may vary by discipline, we use the term here to refer to datasets large enough to be challenging to analyze within a traditional spreadsheet, or on a single computer.

In-depth interviews with twelve UC San Diego big data researchers of various academic ranks and departmental affiliations were conducted, with questions focusing on various aspects of collecting and analyzing big data, infrastructure needs, research communication and data sharing, and training and support needs. Based on these interviews, we identify three primary themes for big data researchers on the UC San Diego campus: *Curation, Open Science,* and *Training*. For each of these areas, we offer a set of recommendations for the campus to better support existing big data research at UC San Diego, as well as develop capacity to support future big data initiatives.

In the area of curation and infrastructure, we recommend:

1. UC San Diego investigate a **campus-wide storage solution that can meet a basic set of researcher needs**. This investigation should be focused particularly on long-term storage needs for data management and retention. Successful implementation of this solution would satisfy not only research community needs, but also campus and funder requirements and compliance issues.
2. This solution should include both **infrastructure and staffing to support use**. Campus should increase support for researchers needing to upload, store, preserve and curate data.

In the area of open science, we recommend:

1. UC San Diego continues to **support existing campus technical and human infrastructure to facilitate data and/or code sharing with an eye towards enabling reusability**. As big data research and data sharing become more common overall and across disciplines, campus should be ready to expand support services to meet the needs of UC San Diego researchers, such as: financial assistance with publishing in open access journals, guidance on best practices for documenting data with a goal of reuse, and tools and support to meet funder requirements (data management plans, data de-identification, metadata creation, embargo and licensing guidelines).
2. **Emphasis during promotion and tenure review on data and/or code that is made open and publicly available** with complete documentation and metadata.

In the area of training, we recommend:

1. Positioning the Library as the **centralized campus resource for supporting foundational research computing training and workshops.**
2. **Assess and develop a centralized list of big data research related training and support information that is easily discoverable and accessible** for researchers and students to use when seeking training related to big data. This will enable and assist UC San Diego students, researchers, and research personnel to find reputable, up to date, and available training resources.
3. **Establish collaborative partnerships across campus** to develop teaching and learning resources. Initiating collaborative partnerships among, but not limited to, the Library, Office for Postdoctoral Affairs, Teaching and Learning Commons, and other primary Organized Research Units is a way to facilitate conversations and collaborations about big data research training support and resources. These partnerships will facilitate connections to interdisciplinary domain expertise necessary for developing the innovative training opportunities researchers need to achieve success in their respective fields.

Introduction

Overview

During Winter 2020 - Summer 2021, the UC San Diego Library's Research Data Curation Program participated in a multi-institutional research study on supporting big data research. The study was coordinated by Ithaka S+R, a not-for-profit research and consulting organization that provides strategic advice and support services for academic and cultural communities.

The goal of the study, which was conducted in parallel with 20 other universities, was to better understand researchers' processes, particularly in the area of big data research. The UC San Diego team's focus was to learn more about how researchers on this campus work with big data and provide a set of recommendations to enhance and develop resources and services that will directly support and benefit research and researchers on our campus.

Research participants and interview process

Twelve UC San Diego affiliates representing various academic ranks and affiliations were recruited to participate in this research study. Each of the twelve participants was selected based on recommendations from various departmental contacts and direct connections with researchers, as well as their discipline expertise, academic affiliation, and rank. The selected participants' ranks were: six professors, two assistant professors, one associate professor, one postdoctoral scholar, and two project scientists. These participants came from the departments of medicine, engineering, oceanography, political science, physics, and data science. Some participants had multiple affiliations and assignments in related departments.

Following IRB approval of this research project, interviews with participants were held virtually September through December of 2020. Interviews were semi-structured, meaning each interviewee was asked the same series of predetermined open-ended questions on topics related to working with big data, including: data collection and format, infrastructure needs, research communication and data sharing, and training and support. This format of interview allowed for in-depth discussion on the topics most relevant to the interviewee, while ensuring that each interviewee was able to comment upon specific aspects of big data research as determined by interviewer questions.

Interview audio recordings were then transcribed and anonymized, and subsequently coded using the qualitative analysis software Taguette. Each team member focused their analysis on a specific theme that emerged during the qualitative analysis portion of this project, which in turn informed the three-part structure of this report.

Report and recommendations

This report targets three primary themes that were raised most often by interviewees: *Curation, Open Science,* and *Training*. The reported findings in these three key areas are followed by a *Recommendations* section that discusses in detail how UC San Diego can improve services for needs expressed as a part of these key themes. The report finishes with *Conclusions and Next Steps*, which notes related existing initiatives on campus and identifies stakeholder groups which will be needed to move forward with implementation of these recommendations.

Curation

UC San Diego has a long history of support for big data initiatives. This can be traced back to several key factors: the presence of the San Diego Supercomputer Center as an official part of campus, the long-lived history of the Scripps Institution of Oceanography, and a large medical research presence are some of the most notable. In addition, over the last fifteen years UC San Diego has embarked on several campus-wide initiatives to support high-end research, including the implementation of data curation services and infrastructure. These campus-specific factors, and others, were cited by interviewees over the course of our discussions with them. In addition, a number of more general trends, expectations, and pressures were expressed, including funder mandates, domain expectations, and researcher preference.

In this section, we address the predominant trends for curation that emerged from our interviews. During the researcher interviews, we did not explicitly use the phrase "data curation," or define it specifically for interviewees. Across the interviews, however, a common shared understanding of what "data curation" means in current research processes emerged. This understanding mapped to three stages within the research workflow: data collection and creation, data analysis, and data sharing and preservation.

Curation support during data collection/creation and analysis

Two main themes emerged as common during the "creation and analysis" phase of the research process: support from campus for various storage needs, and infrastructure support for computation and networking. It is not at all surprising that when discussing needs for big data research, storage would be a common topic. Even before anyone used the term "big data," there was a tension between need and availability. This trend continues, with the most common issues being the lack of sufficient storage space, and the cost associated with storage generally. Within these broad issues, several more specific points were raised.

Storage:

> *"... challenges that I run into are where am I going to store this data? How am I going to back it up? The problem is, what is the right way to physically store the data?"*

- Several interviewees mentioned a need for storage that is directly connected to other pieces of equipment in their research workflows. This might be for reasons of speed and efficiency, the need to tune or set specific functions within the storage, or in some cases, privacy concerns. These interviewees cited challenges particularly in working with cloud storage providers and remote storage facilities. Challenges included inconsistent interfaces and APIs among storage providers, inconsistent performance when using cloud storage over time, lack of support for specific implementation details, and an overall sense of the storage being a "black box."

- Another mention was a disconnect between specific researcher or lab needs, and the types or options of storage available. For instance, there was a mention of simply needing large amounts of storage where such storage did not need to be particularly high performance, or live on any specific hardware platform. In contrast, there was a separate need expressed for infrastructure tuned for specific scientific processes.

- Beyond needed storage amounts, cost was the next significant issue. This was shared in several ways, from lab usage of storage equipment well beyond its rated lifespan, to projects that required moving data between storage locations, with decisions dictated by cost. This also has downstream deleterious effects for researchers, because it leads to inefficiencies as additional, "non-scientific" work is added (e.g., researchers having to monitor cost closely, the need to move data around unnecessarily).

- There were many positive mentions of the San Diego Supercomputer Center (SDSC) and its services related to storage infrastructure. This included direct access to hardware and support on campus, high-speed and -throughput networking attached directly between storage and labs, and the ability to share data more easily with campus collaborators.

- In contrast, there were relatively few mentions of commercial cloud providers. These mentions tended to be negative, especially around cost, but interviewees also cited complexity of integrating cloud services into local research practices. As stated by one interviewee:

  > *"In terms of resources, I'm a big fan of buying hardware, not renting it. We've done stuff on the cloud before, but it always adds like 30%, in time and cost to every project. And as soon as you spend that money, it's gone. If I spend 10 grand on a workstation that sits under my desk then it's forever mine, and I don't need to go get another grant to keep working with that hardware. It's our hardware. There are things that we could do probably if we had infinite money, we would just rent a bunch of EC2 instances and do it really inefficiently across a bunch of them and never have to worry about it."*

Computation:

"We're always hungry for more computing time."

- As with storage, there were a number of positive mentions of SDSC for computation support. In particular, several interviewees cited the Triton Shared Computing Cluster (TSCC) as pivotal to their work. Specifically, TSCC provides them an easily provisioned, customizable platform on which to run code. Also noted was the allied local support for usage of TSCC.

- The most common thread among computation discussions was the prevalence of "homegrown" code and software. That is to say, nearly every interviewee relied on code that was written within their own group and targeted to their own processes. From some, this might be a small part of their workflow. For others, it *is* their workflow. Several common tools are heavily used in the creation of code and software (e.g., GitHub, Python, tools like MATLAB). This characteristic was also tied explicitly to the use of TSCC above: it was cited as an environment well-suited to code that could be optimized for specific research needs.

- Mention was also made of needing ongoing support in the area of code and software, from code optimization to updates to porting to new systems. The biggest challenge cited was the reliance/over-reliance on student help for this kind of support. Several interviewees discussed the challenges of a constantly changing workforce within a lab that needs code and software to persist. Another challenge with student labor is the constant need for education in specific lab processes and protocols, as well as basic data and coding best practices. As stated by one researcher, “That's one of the problems I'm having now with students, is that they're so dependent on all these automatic systems that I'm constantly telling them, you’ve got to go back and look at the data. Because there may be a data quality issue.”

Curation support during data sharing and preservation

Once data has been created and analyzed, the next phases of the scientific workflow begin. Interviewees were uniformly focused on questions around *what happens to my data?* There was a spectrum of responses, but they can be grouped into two categories: data sharing and data archiving.

Data sharing:

> *“Just being a data archive is not very exciting, right? The question is, how do you now open up the data to the community?”*

- Not surprisingly, there were several kinds of data that are not appropriate for sharing. These included all the expected data classes: patient data, human subjects broadly, data tied to specific commercial funding, etc. But these represented a small minority of outputs. Most researchers were eager to have their data made available and usable. Most of the variability in data sharing was due less to specific intent (e.g., “I don’t want to share”), and more to workflows and commitment (e.g., “I just don't have much time to spend on it”). As one researcher stated,

  > *“Data management and data handling, data science is something that isn't in our domain. I have been trained to think about [my domain] and I want to get to those [domain-specific] answers. And I see it [sharing data] more as a responsibility rather than something that I get excited about.”*

- A number of interviewees actually used the word “curation” to describe this stage of work. Many discussed how much time was spent (often graduate student time) cleaning data post-analysis to make it available to other sources and users. While most commonly this was an in-house process, some researchers and labs looked to external help, outside of the university. However, this didn’t tend to be a worthwhile activity. In fact, one interviewee noted their lab going in the opposite direction:

> *“... [for curation] consulting, I haven't yet found too many people to help with that, but I've found that our infrastructure here at UC San Diego has been much stronger than anything I can get elsewhere. And what we're trying to do is to then use that for our colleagues. So to actually reverse the script, to have it be such that people come to us for these services.”*

- Even when they didn’t explicitly use the word curation, interviewees frequently mentioned needs including, “data sharing,” “data accessibility” and “data discoverability.” Some mentioned these as explicit funder requirements, while others said they are just part of good scientific practice. Regardless of the origin, these topics we mentioned repeatedly and are discussed in more detail in the “Open Science” section.

Data archiving:

> “*And then that data, where does it go? It goes into a journal publication as a PDF file but that's really just an image of the data, it's not really the data. So that data sort of disappears, we put it on our website but I guess when I retire it might disappear someday.”*

Most interviewees expressed unease, at best, with the long-term preservation of their data. No interviewee viewed long-term preservation of their data as a solved problem, which is in contrast to every other topic discussed. Reasons for wanting or needing long-term preservation varied. Several interviewees expressed it as a function of ensuring professional legacy; others talked about funder requirements; still others emphasized a frustration that data that was expensive and time-consuming to create would just disappear. Regardless of the reason, several threads emerged:

- For some interviewees, cost was the main factor. Since the vast majority of data created was funded via grants, there was no clear way to pay for any kind of maintenance beyond the specified research years. In some cases, interviewees had to content themselves with letting data sit on local storage systems until these systems stopped functioning. A common task was to deposit a small amount of data, as required by a journal, in a repository. Some interviewees did have a commitment to maintaining data for the long term, but this was usually done as part of a larger access plan, rather than for preservation *per se*.

- Another factor was scale. Several interviewees had data of multiple petabytes, and there is currently no good long-term solution for data of this magnitude. Typically these researchers make a “best effort” to keep what they can, while explicitly acknowledging that loss is inevitable.

- Some interviewees actually used the interview to find out about options for preserving their data. For them, they had either stopped trying to solve the problem of long-term preservation, or just were not aware that alternative options do exist.

- For most of the interviewees, barring the multi-petabyte users, there was a desire to find the best places to archive their data. Many expressed an interest and willingness to continue using the university Library for the task, if they were already using it. Others expressed frustration that campus as a whole didn't provide a consistent solution for this issue. As one interviewee said:

  > "*So one of the things we could use help with is how do we archive this stuff? I'm not going to be here much longer … so I want to figure out how do we keep this dataset going on forever because we have ... graduate students going back to data that was 10, 15, 20 years old and they're still discovering new things that are in the data because it's so rich.*"

Open science

In recent years, the cultural move towards open science — an approach to scientific research where all research outputs, including data and analysis code, are publicly available — has led to an ever-increasing amount of data available for replication and reuse. This trend towards sharing data is especially important for “big data”, which is often labor-intensive and computationally expensive to create and store. Concurrent with the open science movement, data professionals have been urging researchers to make their data FAIR (findable, accessible, interoperable, reusable) in order to fully realize the potential of broadly shared data. Adoption of data sharing practices and FAIR guidelines is increasing across disciplines, but as with any new process, real-world implementation often falls short of the ideal due to resource limitations.

In this section, we first provide an overview of “open” platforms and processes used by researchers and general attitudes about open science, then discuss the nuances and caveats associated with the general tendency towards openness and sharing displayed by interviewees.

Commonly used platforms

Open software and code

Mentions of open or free software platforms by interviewees tended to outnumber mentions of proprietary software, although this was variable by domain. Nearly all interviewees mentioned using GitHub, a free (or paid, for certain features) web platform for code repositories and version control, as a platform to share code not only with lab members or collaborators, but also with the broader academic community. Multiple interviewees noted their reliance on Python as the programming language of choice, as well as Jupyter Notebooks, particularly as a method for code dissemination. Both Python and Jupyter Notebooks are open-source, free and available for anyone to install and use. Other open and/or free (depending on venue) software and platforms referenced included: R/RStudio, LaTex, SQL, Google Drive, Google Earth Engine, and Unix/Linux. Mentions of proprietary platforms were mostly of MATLAB, a programming language, with a few references to more niche discipline-specific programming languages and other platforms.

Open data

Sharing data was often noted as a community-driven effort, unsurprising for “big data,” which is often monetarily and computationally expensive to produce. Multiple interviewees reported using data from existing public domain-specific data repositories (often funded by the same organizations that funded their research). Other interviewees mentioned hosting their own domain-specific data repository. In these cases, the repository ranged from a small platform, sharing only their own data, to a more robust system that allowed other researchers in the domain to also deposit data. When research depends on access to other people's data, the general consensus seems to be a positive attitude towards open data/data sharing. As one researcher put it,

> *"...open data promotes better tooling, which then promotes better science. You can't create really good, broadly generalizable methods that work for everything, unless you have access to lots of different kinds of data."*

Open access publishing

In addition to avenues for sharing data and code, interviewees mentioned publishing peer-reviewed articles as a primary method of disseminating their results. Many expressed support for preprints, as well as having analysis reproducible in a public Jupyter Notebook. However, in regards to traditional publishing avenues, attitudes towards open access were more mixed. A few interviewees noted that open access publishing tended to be more expensive and if their grant funding didn't include monetary support for publishing, then a cheaper, non-open access journal would be preferred. Conversely, other interviewees noted that library assistance with funding for open access publishing meant they were more likely to publish their results in an open access journal. Only one researcher mentioned eScholarship, the University of California institutional repository, in the context of a platform to make technical research documents citable, without the more lengthy peer review process.

"Open" is a spectrum

Based on the responses from interviewees, it is clear that "openness" in the context of research data and/or code is not an open/closed binary, but rather a spectrum: on one end is the practice of making data and code (and associated documentation) freely, publicly available, and accessible, and on the other end is a completely closed system of research.

Where interviewees reported falling on this spectrum was a product of disciplinary norms, funder and journal requirements, privacy considerations, time and interest available to spend on documentation and curation, and personal preference. As might be expected, given the prevailing positive cultural attitude towards open science, most interviewees fell somewhere in the middle of the openness scale, where "open" meant either sharing their own data within a consortium or with collaborators, or using other people's (variably documented) publicly available data.

Based on interviewee comments, we were able to categorize researcher scenarios within a "spectrum of openness", which we define to include the following, from most to least open:

- Fully open, in practice: data and code are freely, publicly available. Metadata and other documentation are provided in enough detail to enable reuse.

- Fully open, in theory: data and code are freely, publicly available. Documentation is sparse, or lacking, making reuse challenging by anyone not already familiar with the exact methodology.

- Data/code shared with collaborators: data and code are shared within existing groups of collaborators and upon request from other researchers, in which case an offer of collaboration may be expected.

- Data shared within a consortium: data is generated by and shared within a large consortium of institutions and researchers. Data is not shared outside the defined consortium. Code is developed by local groups and may or may not be shared.

- Not shared, but able to be recreated: data is not shared, either due to terms of service, or data size constraints, but methodology is detailed enough for users to recreate dataset from the original source. Code to do so may or may not be shared.

- Not shared, not able to be recreated: data is not shared in any form, due to funder mandate, privacy issues not alleviated by de-identification methods, or other restrictions.

Incentives and barriers

So why do researchers share, or not share, data and/or code? While some interviewees reported sharing because of a philosophical agreement with the concept of open science or a strong sense of community within the discipline (i.e., if you use other people's data, you should in turn share data you collect yourself), most reported more tangible incentives.

Incentives to sharing:

> *"People will not [share] because it's the right thing [to do],*
> *they will do it because there's a benefit [they] get out of it."*

- Multiple interviewees mentioned that journal requirements for sharing were the most effective and motivating. This is due to the influence journals wield. As noted by one researcher,

  > *"There's definitely incentive at the publication stage because you sometimes can't publish or submit if you don't have it [data] available somewhere open access."*

- Along with journals, grant agencies were mentioned as an incentivizing source, with grant requirements ranging from community-accessible data or code being a key requirement of funding (e.g., when funding supports a community platform or effort for sharing data and/or code) to general funder guidance (e.g., new NIH data sharing policy).

- Other concrete reasons researchers mentioned for sharing data and/or code were related to tenure and promotion. For instance, making data open and reusable can lead to an increase in citations, a traditional metric of impact, and published data and/or code can be included as part of the academic record under review.

- On a smaller scale, some interviewees mentioned encouraging open sharing of data and code within their lab so that when students graduate, no work is lost. Using, for example, a code sharing repository platform like GitHub is beneficial when student researcher turnover is high and efficiency will be drastically reduced if each new student must develop code from scratch.

- Whatever the incentive to share may be — journal or funder requirements, credit during academic reviews, or for internal logistical reasons — the crucial point made by interviewees was that there must be either a requirement or a tangible benefit of some kind.

Ultimately, whether data and/or code were shared was the result of a resource and value judgement between incentives to sharing on one hand and barriers to sharing on the other. Along with fairly strong incentives to share data and/or code, interviewees reported a variety of barriers to sharing, from individual-level rationales to funder mandates and domain-specific data considerations.

Barriers to sharing:

"Data sharing is resource intensive."

- A clear and inviolable barrier to sharing is the presence of potentially identifiable data. Privacy considerations are valid reasons to restrict data access, which multiple interviewees noted. While medical data is an oft-cited example of data that can and should be kept private and restricted, interviewees also noted a host of other examples, including: data on sensitive topics, copyright issues with image data, instances were data sharing would breach the trust (if not the confidentiality) of the study group, and research data that would require written permission from foreign governments to release.

- Even when de-identification methods were available, researchers tended to be cautious in terms of sharing. This attitude was described by one researcher who said:

  > *"We usually will not release the data because even if we try our best to de-identify it, if it is released and there are some issues with it, it can have some consequences that we don't really want to deal with."*

- Another data type that was often not shared was social media-related data. For many social media platforms, even though posts are "public" and therefore theoretically accessible for research, re-sharing bulk data violates the terms of service. In these cases, interviewees mentioned that they would instead share enough methodology so that anyone interested in the exact data could go through the appropriate channels (usually via API, either credentialed or not) in order to recreate the dataset described.

- Another barrier to sharing identified by interviewees was at the funder level. Although not common, certain entities such as government agencies include directives within grants prohibiting public sharing of data (e.g., data from Department of Defense funded projects may be considered confidential or restricted access).

- In some cases, the main barrier to sharing was due to prevailing norms within the discipline. One researcher noted:

  > *"If it was left up to me, I would have probably published [data] and said, 'Look, this is the data, here's how you use it, knock yourself out.' However, [in my discipline, this practice is] actively discouraged."*

- Along the same lines of disciplinary standards, there were a couple mentions of the fear of getting scooped in a highly competitive field, if data were openly available. However, most interviewees were open to sharing data after a set period of time during which they reserved the right to publish first (i.e., a data embargo).

- Aside from these more entrenched barriers — privacy concerns, funder requirements, disciplinary culture — interviewees also noted another major barrier to sharing: the lack of resources needed to share data properly. Data sharing is resource intensive in terms of time and infrastructure: metadata needs to be created, additional documentation may be required to provide full context, and there needs to be a proper interface for data access, which may or may not be pre-existing. As one researcher noted, "It takes a cognitive effort to say, 'Okay, where are my data assets? Where did they come from? How do I package them? Who will I give access to? How will they access it?' That takes a person suddenly to do that full-time." Another researcher summed up this quandary, noting: "It's not an incentive to share if [the process to share] is creating an additional burden".

- Sharing code is equally labor intensive. Code documentation is crucial for reusability and as one interviewee noted, "I want to release all of my code, but I don't really have time to clean up [every script] and then...release them…". Additionally, there is often an expectation by users to keep code up to date, which creates a burden on the originating researcher: "I don't want people coming to me from outside or even inside...to keep asking…[is this] a bug, or this doesn't work, or I want this feature...".

These types of barriers are, of course, not restricted to "big data" research and can be found in a wide variety of research types. The one "big data" barrier interviewees mentioned was data size itself as a complication for data sharing. In order to make data open/publicly accessible, there needed to be existing infrastructure not only to host the data, but also to make it easily downloadable by interested parties in a machine-readable format. Such infrastructure may or may not exist for the domain and data format/type in question, and researchers may need access to a supercomputer or other high performance computing cluster in order to handle the downloaded data.

Importance of shared usability

Although some researchers who share their data may be doing the bare minimum to fulfill any sharing requirements, many researchers who share genuinely want others to reuse the data in future research. In order to effectively and efficiently reuse pre-existing data, detailed metadata (e.g., number of samples, if there is missing data, any relevant pre-processing and data provenance, etc.) is necessary. As such, a minimum level of curation is essential for shared usability; without it, open sharing of data and code is merely lip service to the concept.

Curation, therefore, serves as both incentive (data will be reused) and barrier (curation is a non-trivial task).

> *"There is a big difference between something being available*
> *in principle and in practice, because software without instructions,*
> *it's just a bunch of bytes and gibberish, right?"*

- Interviewees were keenly aware of the need for metadata and other contextual information, but most identified the lack of time as the main hindrance to sharing their own data/code in a well-documented format.

- Conversely, expectations were high when trying to use other people's data/code. The scarcity of well-documented machine-readable data was noted by many interviewees, multiple of whom had anecdotes about needing to manually extract data, or pertinent information about data, from published papers. One interviewee noted:

  > *"...how crazy it is that researchers put really useful data into papers in non-machine-readable format. And then anybody who wants to use it has to go in and curate it out...[which requires] individuals pouring over many, many papers to pull out data".*

- Multiple interviewees mentioned that the ultimate example of reusable data and code would be a public Jupyter Notebook that starts from the raw data (which is also accessible) and goes over every processing and analysis step, all the way to final figures and results. While some interviewees said they did this with their own research, the majority did not.

Training

Training and support at UC San Diego play a key role in the success of big data research. Training from various resources directly impacts how researchers conduct their research activities and achieve successful results in their areas of expertise. During the course of the interview process for this project, each of the interviewees was asked questions related to receiving, or not receiving, data-related training, from the beginning of their early research work throughout their careers. Additionally, the interviewees were asked about anticipated future needs for training and support that would be most beneficial in the area of big data.

In this section we address the primary training themes that emerged from the interviews. These themes were: training issues, training resources beneficial for scholars, and the Library's role in training support.

Training Issues

When discussing the topic of training for big data research topics, two issues were mentioned across interviewees: personnel or staff training issues, and the specific need for training to learn foundational research computing skills or required software tools.

Personnel or staff training issues:

"Graduation killed the code."

- Several interviewees discussed personnel or staff training issues in their labs or research work. One of the primary concerns for several of the interviewees was the unavoidable issue of graduation or transferring to another institution (i.e., when graduate students or post-doctoral staff leave the lab or university). The institutional knowledge, experience, and expertise leaves with them, which in turn requires increased lab resources and significant time dedicated to retraining new staff or moving on to another project.

- As stated by one interviewee, a major concern with graduate students working on projects is:

  > *"in academia, historically, graduation killed the project; graduation killed the data, or, in computer science in particular, we like to say 'Graduation killed the code.' That sets off that whiplash moment that we see in many domains, where people change topics essentially on that expiration date of the graduate student. There is a constant concern to make sure not everything dies with the graduation of a student."*

Another common personnel training issue that was mentioned frequently is finding training resources for research personnel or students to learn new methods for solving problems related to working with data. When asked a follow-up question about advising colleagues or students who need to learn new methods or solve a statistical problem, interviewees often mentioned the problem of not knowing specific training resources to recommend for themselves, research personnel, or students.

Several issues emerged related to discovering training:

- Multiple interviewees acknowledged their training was received through trial and error during self teaching. In the case of research personnel, on-the-job training from collaborators or peers was the primary method for learning new tools, programs, or processes.

- Practical experience was highlighted as another important factor for developing skill sets and training for incoming and long-term personnel. One interviewee stated that post-docs, or a visiting researcher that works in a lab, who do not have specific training in the basics, become more of a training problem, noting, "I do not want to sit down with them for an hour and tell them how to do [programmatic file navigation] and all that kind of basic stuff." Lack of foundational computational skills leads to additional time and effort needed to explain the basic concepts repeatedly to incoming students or research staff.

- Another significant issue identified by interviewees was the need to have practical training available for students to learn how to be data literate. Students learning how to conduct big data research are often focusing on learning the latest computing technology or methods and spending less time learning about the quality of the data they work with. Referring to the availability of training related to data literacy, one interviewee stated,

  > *"This reliance on automated systems diverts students from focusing on the quality of the data. Which will lead them to disregard the need to receive practical training to learn how to thoroughly examine data for quality issues."*

Specific need for training:

> *"Since I'm centered in the department of medicine, nobody has the training, so all of this falls on my lap. It would be really awesome if the students and postdocs could have some training."*

Related to the importance of practical experience is another common issue mentioned by interviewees: the need to receive training *before* research staff begin working on data-intensive research projects. Whether training is received through trial and error, self-teaching, or practical on-the-job experience, some form of basic or advanced training is necessary for research groups consisting of new and experienced research personnel and students working on various projects.

- Several interviewees stated that students and postdocs, research personnel, faculty, and collaborators working in labs, conducting big data research, were all candidates for training. The interviewees acknowledged communication barriers related to technical tasks can occur between research staff with differences in discipline expertise and ensuring a common level of knowledge via training would improve project work by improving communication between research group members.

- Interviewees identified two specific skill levels when asked about training resources for colleagues or students: required training at the basic or foundational skill level, and supplemental intermediate or advanced training for experienced research staff. Among the responses, there was consensus that training is necessary at the foundational level when researchers are new to labs or projects.

- A number of interviewees mentioned research personnel would benefit from learning several foundational research computing and open-source data science related tools. These included software and programming tools such as Python, R, MySQL, GIS, LaTex, Unix Shell, Github, Jupyter Notebooks, Hadoop, machine learning, and visualization tools.

- The majority of interviewees identified the most useful basic training topics as: computing infrastructure on campus, basic elements of accessibility and transferability, critical thinking and data exploration, how to discuss research fundamentals, how to talk about the data itself, and learning how to talk about big data.

- Many interviewees lead or work in labs supporting big data with research personnel that have varying degrees of research computing experience and skill sets. The intermediate and advanced researchers who focus on machine learning or deep learning utilize data literacy skills to improve communication about their work. Furthermore, these researchers are the developers for coding and software development tools who would benefit from additional training resources in these specialized areas. In turn, these technically skilled researchers become *de facto* trainers, providing peer training within their research lab environment. However, an adverse side effect is the time lost to conducting research when helping train novices on foundational skills.

Training resources beneficial for scholars

Various training resources related to big data research are widely available and beneficial for researchers. Peer learning, open educational resources (OERs), and formal training programs or campus departments were identified by the interviewees as significant resources most useful for scholars. These training resources were utilized by interviewees to meet their need for personal skill development or recommended for their research staff to better support research activities.

Peer learning

The most common and readily available training resource mentioned was peer-to-peer learning within the research lab environment. Peer learning was noted as an organic occurrence between all levels of staff and students in research lab groups.

- Several of the interviewees mentioned their labs had interdisciplinary staff pools which naturally developed a “cross-pollination” for an internal exchange of expertise.

- Another peer learning benefit identified was the pairing of full-time research staff with graduate students. This pairing somewhat addressed the persistent cycle of losing expertise when “everything dies with the graduation of a student”.

Open educational resources

OERs have had a significant impact on how researchers find, recommend, and receive training for themselves, their research lab personnel, and students. Such resources have made learning new skills and topics a low-barrier, cost-effective approach for supporting big data research training.

- A few of the interviewees mentioned open educational training and support resources as an easily accessible and readily available resource for their labs. For example, one interviewees mentioned The Carpentries, which is an international community of volunteer instructors who develop and teach open and accessible lessons focusing on foundational research computing topics.

- Additionally, a variety of OERs widely available online, covering various topics that focus on research data analysis, were mentioned as primary or supplemental resources for researchers wanting to learn or apply new skills or tools in their work. OERs mentioned specifically by interviewees included funder-sponsored workshops such as a big data workshop sponsored by the National Institutes of Health (NIH) and other National Science Foundation (NSF) or NIH funded software development data analysis types of workshops.

Educational training from campus departments

In addition to OERs, interviewees mentioned participating in or recommending local computing and data analysis workshops related to big data from UC San Diego Organized Research Units such SDSC or academic departments such as the Computer Science and Engineering (CSE), and Bioengineering academic departments. These local training options have been beneficial to their research activities and easily accessible when they are aware of the local training that is available.

Fee-based educational programs and workshops

There are also a significant number of fee-based educational programs, workshops, and training opportunities available online from commercial vendors and at research institutions like UC San Diego, agencies such as NSF or NIH, and at conferences.

- When asked where they would advise a colleague or student to seek out training, multiple interviewees responded that they use, and would recommend, online educational services and conferences. One interviewee went even further, saying:

> *"I send my students wherever they want to go. So I've sent them all over the world...A lot of universities now will have NSF or NIH funded software development data analysis kinds of workshops. So I will send them to summer schools, I will send them the boot-camps, whichever ones I want and encourage them to go to those kinds of things just because it's, in my opinion, a very valuable and useful set of skills and good networking too."*

- Another interviewee mentioned the UC San Diego Extension classes as a resource for fee-based educational programs and classes.

Formal educational degree programs

Formal education classes are the primary means through which the interviewees were trained to learn computing and analysis tools for their current research. The primary training acquired by the interviewees during and/or after completing their degrees was continuing education and on the job experience. These *ad hoc* acquired skills represented the only "training" on topics such as data management for many interviewees throughout their careers.

Training support from the Library

The UC San Diego Library currently serves as a primary campus resource for research data management resources and training support. As discussed by multiple interviewees, academic libraries provide the resources, facilitate the training, and house the information researchers need to accurately and efficiently manage their big data research. One interviewee stated that when working on their PhD, the library at their institution "had a lot of wonderful resources embedded in the library, all kinds of workshops and a core group of support staff." The interviewee went on to say that,

> *"By going to the workshops [at the library], you would meet more of the staff and then you would learn that they have a knack for a certain thing, and then you could seek out more help from them for one-on-one type of issues that you're having."*

Academic libraries are well positioned as an educational hub for training workshops and can serve as a primary research and training resource: libraries have accessible physical spaces and the expertise and availability of discipline specialists who provide consultation, training, and support before, during and after researchers invest time in attending a workshop or training event. At UC San Diego, the Library's Research Data Curation Program and its ecosystem of data literacy expertise is challenged to meet the big data research needs of the increasing numbers of students and researchers who need the foundational educational opportunities outside of the classroom or research lab environment. With the rapid evolution and complexity of new hardware and software, and the quantities of data being produced, it is necessary to plan for scaling up support for big data students and researchers through training programs offered by the Library. By entering into collaborative agreements with various allied specialty programs, access to resources and training can be better coordinated for greater efficiency. These collaborative training partnerships can deliver introductory workshops in data literacy, data management, and foundational computing, targeting specific research groups that need this additional support to skill up research personnel and enhance research productivity.

Recommendations

Curation

In contrast to previous surveys of the campus community, interviewees demonstrated a strong grasp of the need for storage and compute infrastructure that supports long-term curation of data. However, having the expertise and/or staff to accomplish this work varied widely. Based on the interviews, several key needs, and potential solutions, were identified:

Campus should define a highly available storage infrastructure that supports the most common research needs. This recommendation is put forward with full knowledge that no solution will meet all needs. However, there are enough commonalities that could be addressed with a single, robust option. Some specific characteristics of this infrastructure:

- Provides basic active storage that can be accessed by campus labs and equipment via high-speed networks. Almost every interviewee expressed a need for this service. It was also commonly mentioned that cloud solutions in this arena are usually expensive and not optimal for local research purposes. An on-premise solution is strongly recommended.

- As noted, this infrastructure should take advantage of the high-speed network campus has already spent significant money installing and configuring. This would provide researchers the opportunity to tie their specific labs and equipment to the storage service.

- Beyond active storage, this infrastructure should either provide long-term storage that can exist beyond grants, or explicitly tie to such a service. This should be an automatic feature, so that researchers don't need to consider multiple storage systems over the life of a grant.

- Because it is managed by campus administrators, this infrastructure can be designed and maintained so that it satisfies all current and future security and compliance requirements. Researchers wouldn't have to worry about maintaining systems on their own. More importantly, campus would have stronger management to help avoid major security issues.

- Further, this infrastructure would provide campus administrators with an opportunity to assist researchers in being compliant with data management plans and proposals. For example, if a funder demands to be shown that data are available and maintained properly for long-term access, this can more readily be shown in a system that is under active campus management.

"*Infrastructure" in this context is not just physical: it also includes a variety of support staff charged with a range of tasks*. It's never been adequate to simply install equipment and expect people to use it. More importantly, if this infrastructure is to go beyond basic storage, and become a true data management tool, staffing assistance is required. For example:

- Interviewees consistently commented on funder requirements for data sharing and long-term retention. Most also commented this is not something typically within their area of expertise, nor part of their budgets. Campus should provide staff who can provide at least basic help in creating metadata, uploading data in ways it can be easily accessed, and assisting with general usage questions.

- This campus infrastructure should be cited in proposals as meeting or exceeding funder requirements. Staff can assist in providing appropriate language and descriptions for these proposals. In doing so, they can also provide first steps to solving a crucial campus problem — a way to make sure researchers are complying with their data management plans.

Open science

Collecting, processing, and analyzing big data is time intensive and computationally expensive, and openly sharing data and code is a sensible approach to making big data research more efficient and effective. Given the overall positive attitude from interviewees towards sharing data and/or code (when not prohibited by valid barriers), the manifold benefits of sharing data and code, and the lack of time and resources being the number one self-reported barrier to sharing, the campus priority should therefore be to assist researchers with incorporating principles of open science into their existing workflows.

Campus resources devoted to helping researchers make their work public are necessary for supporting a campus-wide move towards open science. This includes financial assistance with publishing in open access journals, guidance on best practices for documenting data with an eye towards reuse, and tools and support to meet funder requirements (data management plans, data de-identification, metadata creation, embargo and licensing guidelines).

- Supporting researchers in this area does not entail starting from scratch, but rather making better use of current campus resources already available to researchers and expanding such resources when needs are not being fully met. For example, there already exists on campus a framework for supporting various aspects of big data research. Within this framework, the Library's Research Data Curation program provides support for data documentation and data deposition into the Library's research data collections. Similarly, Research IT Services can assist with large data transfers and other data-related infrastructure needs, as can SDSC and Nautilus.

- Strategically investing not only in technical, but also in human, infrastructure in programs like these will be necessary to scale such services to support departments across campus, especially as data and code sharing become more common across disciplines and more often required at the funder level.

A renewed emphasis on appropriate credit during promotion and tenure review for data and/or code that is made open and publicly available with complete documentation and metadata. As evidenced by interviewee responses, the time and resources needed to share data and/or code is non-trivial but beneficial to the researchers' discipline at large.

- As such, encouraging researchers to reflect on their own sharing practices in review packets — whether by linking to a public version of their research output or commenting on why data are not shared — allows researchers engaged in the crucial, but too often overlooked, work of sharing data and/or to showcase their contributions to their discipline at large.

- Of course, it is equally important to recognize that not all types of research lend themselves to being publicly available, whether from privacy considerations, funder directives, disciplinary standards, or other issues.

Achieving the appropriate balance of making sure researchers who share their data and/or code receive credit for this additional work during promotion and tenure evaluations, while not penalizing researchers who, by nature of their specific situation, are unable to share publicly, will take considerable effort and planning. However, we believe that the benefits in this case will far outweigh the upfront work needed to implement properly.

Training

Interviewees generally reported that they had received research data training either on-the-job or were self-taught. The desire for the availability of dedicated resources and training focusing on the needs of big data projects was universal among the participants in this survey. There was consistent acknowledgement that training to support big data research at UC San Diego is available, but generally unrecognized, and necessary to enhance the skills of students and research lab personnel. The recommendations for training and support are as follows:

- Identify and organize the library specialists and key campus stakeholders to collaboratively work together on training. The goal is to identify dedicated big data training topics and establish resources for promoting and teaching data literacy to respective teams across all disciplines.

- Expand collaborations with postdoctoral researchers. Collaborations between the library and research support units, such as the Office of Postdoctoral Scholar Affairs, can establish unique discipline-specific training opportunities in data literacy, data management and foundational computing. Postdoctoral researchers seeking instructional opportunities would be an invaluable resource for developing discipline-specific foundational instructional activities for research groups.

- Identify a physical space to facilitate the collaborative process. The implementation of the plan developed by the proposed multi-disciplinary data literacy training collaborative can include regularly scheduled classes and workshops, in-person training and drop-in clinics. Creating this dedicated physical space, such as a Research Center within the library, would offer students and researchers “an environment to meet more staff which can establish the potential for collaborations or solving issues” (as described by one interviewee).

- Identify and develop a case study-based training program for supporting and enhancing big data research. As one interviewee noted, “Learning about well-managed programs and efforts that have lots of data and trying to unpack why that works, and how what worked there, could be applicable to me”. Indeed, case studies have been shown to be an effective strategy to teach and learn new skills and tools. A comprehensive review and assessments of currently available training related to big data at UC San Diego is needed. A compilation of currently accessible and available training opportunities provided on campus would immediately assist the research community and also evaluate the need for a case study-based training program.

- Provide free and fee-based training services outside of the formal academic setting. There were requests from interviewees for training options outside of the formal academic setting, for example, group specific classes, individual tutoring, or certification programs. One interviewee recommended “consulting happy hours” or drop-ins for free support and assistance. Additionally, interviewees referenced and supported the benefits of paid consulting services when there is a need for assistance beyond what could be provided by free support.

- Facilitate campus data training activities that enable peer-learning and academic recognition. More informal or less guided activities with peers can be extremely valuable, with one interviewee advocating learning from participation in programming/coding competitions and data hackathons. Engaging in competitions of this nature could build enthusiasm for similar activities offered by the Library or research units, as these types of competitions also serve as team building exercises to break the monotony of the classroom or laboratory setting. These activities enable peer learning, increased proficiency, and a valuable opportunity for networking. Providing information and organizing competitions would benefit students and researchers interested in gaining hands-on, practical experience in a less formal shared learning environment.

Conclusion and next steps

The in-depth interviews with UC San Diego researchers working with “big data” indicate a number of potential avenues for the campus to better support researchers in the areas of curation, open science, and training. While the interview group was relatively small, compared to the number of researchers currently at UC San Diego, it included researchers from a variety of domains and ranks, and the needs mentioned were consistent across interviewees.

More broadly, this report comes at a fortuitous moment in time. Several groups at UC San Diego are looking at the next generations of infrastructure, technology, and services for campus researchers. While there is overlap among these efforts, they each have a specific focus and audience. Taken together, campus leadership has an opportunity to make significant improvements in how research, particularly big data research, is supported on campus.

Examples of allied processes include:

- Creation of “Blueprint 2030” by the Research Compute and Data Services Committee. This document is presenting a vision for research support for the campus over the next decade. It is based on a series of interviews with campus leadership, a large survey of campus researchers, and targeted focus group discussions. There are notable overlaps in the recommendations in the “Blueprint” and this Ithaka report.

- “Accelerating Public Access to Research Data Initiative,” led by the Library and other organizations on campus. Based on recommendations from a multi-year program involving stakeholders from around the country, this initiative is looking at how to improve transparency and reproducibility of scientific results, increase scientific rigor and public trust in science, and - most importantly - accelerate the pace of discovery and innovation through the open sharing of research results. The group is examining the establishment of policies around research, and how UC San Diego must invest in the infrastructure and support necessary to achieve the desired aspirations and aims of the policies.

In addition to these processes, it is important that this report is shared with a wide spectrum of campus stakeholders, all of whom are impacted by the recommendations in this report and whose support will be required to fully implement the recommendations. This includes service providers, decision and policy makers, and user groups. Examples include:

- Offices of Research Affairs and Academic Affairs
- Leadership of the San Diego Supercomputer Center
- UC San Diego Faculty Senate
- Information Technology groups in both campus and health IT systems

- Campus Research Institutes

In conjunction with the other campus-wide efforts mentioned above, and in collaboration with relevant stakeholders, the recommendations in this report aim to position UC San Diego as a continuing leader in data-intensive research for many years to come.