Data, Software, and Advanced Computational Usage of University of Washington Research Leaders

Drew Paine
Human Centered Design & Engineering, University of Washington
pained@uw.edu

Erin Sy
Human Centered Design & Engineering, University of Washington
esy5@uw.edu

Ying-Yu Chen
Human Centered Design & Engineering, University of Washington
yingyuc@uw.edu

Charlotte P. Lee
Human Centered Design & Engineering, University of Washington
cplee@uw.edu

February 27, 2014

HUMAN CENTERED DESIGN & ENGINEERING TECHNICAL REPORT
HCDETRS_2017_01

Data, Software, and Advanced Computational Usage of University of Washington Research Leaders

A CSC Laboratory technical report by

Drew Paine, Erin Sy, Ying-Yu Chen, Charlotte P. Lee

2014

Computer Supported Collaboration (CSC) Laboratory

https://depts.washington.edu/csclab/techreports/

Department of Human Centered Design & Engineering

University of Washington

*This material is based upon work supported by the National Science Foundation under **Grant Number IIS-0954088**, an NSF CAREER award for junior faculty. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Department of Human Centered Design & Engineering, or the University of Washington.*

Abstract

Scientific research is increasingly data-intensive, relying more and more upon advanced computational resources to be able to answer the questions most pressing to our society at large. This report presents findings from a brief descriptive survey sent to a sample of 342 leading researchers at the University of Washington (UW), Seattle, Washington in 2010 and 2011 as the first stage of the larger National Science Foundation project "*Interacting with Cyberinfrastructure in the Face of Changing Science*." This survey assesses these researcher's use of advanced computational resources, data, and software in their research. We present high-level findings that describe UW researchers': demographics, interdisciplinarity, research groups, data use, software and computational use—including software development and use, data storage and transfer activities, and collaboration tools, and computing resources. These findings offer insights into the state of computational resources in use during this time period as well as offering a look at the data intensiveness of UW researchers.

Recommended Citation:
Paine, D., Sy, E., Chen, Y-Y, & Lee, C. P. (2014). Data, Software, and Advanced Computational Usage of University of Washington Research Leaders. (CSC-2014-01) Computer Supported Collaboration Laboratory Technical Reports: University of Washington.

Table of Contents

1 Introduction

Scientific research is increasingly data-intensive, relying upon advanced computational resources to be able to answer the questions most pressing to our society at large. In this report we discuss findings from our 2010-2011 survey of 342 of the University of Washington's leading researchers regarding their research work. The University of Washington, located in Seattle, WA is one of the United States leading research institutions. It is consistently ranked as a top public university in the nation, as well as the top 20 internationally. In addition, the University of Washington is ranked number one among public universities in the United States in receipt of federal research and training funding, and has been in the top five since 1975 [1]. Specifically, this survey was designed to obtain descriptive data about Principal Investigators (PIs) and their research groups, the science they undertake, their data activities, and their computational use. Here we report on the demographics of our survey's respondents before characterizing their data use and examining their use of computational resources. Before examining these findings we briefly discuss the larger context that this survey fits within followed by a detailed overview of the survey and its development.

2 Context of the Survey

Scientific research is increasingly data-intensive, relying more and more upon advanced computational resources to be able to answer the questions most pressing to our society at large. Data-intensive science is today commonly portrayed as the next frontier of scientific research practice. Data-intensiveness is typically loosely defined. One broad definition from a 2012 United Kingdom Royal Society report is: "*science that involves large or even massive datasets*" [2, p.12]. Working to better characterize what constitutes data-intensive science and determining how to support the practices and tools that enable it is a pressing matter for many scientific communities. Within the United States significant effort has been put into the development and study of cyberinfrastructure [3,4] as a means to support data-intensive science. Cyberinfrastructure are large-scale, collaborative research enterprises that apply advanced computational resources and large datasets to scientific research problems. In our field, Computer Supported Cooperative Work (CSCW), significant work has been undertaken to study the development of cyberinfrastructure and the work of data-intensive science (see [5] for a recent overview).

This survey is part of the five-year National Science Foundation (NSF) funded study "*Interacting with Cyberinfrastructure in the Face of Changing Science*."[1] This study is conducting a longitudinal examination of how scientific research practice evolves in conjunction with technological and educational practice. It is designed to empirically examine how inter- and multi-disciplinary scientists at a major research university, the University of Washington, grapple with changes in research practice in the face of increasingly data-intensive science and rapid evolution of cyberinfrastructure. As a result, this survey was designed to provide us with an initial understanding of research taking place at the university and of researcher's advanced computational use. The findings from this survey informed the development of a semi-structured interview protocol that we used to interview 20 of the 120 respondents. These 20 researchers were selected for their potential to be enrolled in the longitudinal portion of our study. Below we discuss the development of this survey before presenting findings and commentary from the responses.

[1] See https://depts.washington.edu/csclab/projects/career-ci-study/ for more details.

3 Overview of the Survey

The findings and commentary on University of Washington research leaders provided here are the result of a 51-question survey that was administered during 2010 and 2011. This survey was designed to help us begin to answer our larger study's research questions. We developed our survey using based upon guidance from multiple sources beyond our own project's research questions. These resources included a University of Washington report, conversations with members of the eScience Institute and College of Engineering Dean's Office, and institutional funding data from the UW Office of Sponsored Programs.

One resource that we drew upon was the University of Washington report "*Conversations with University of Washington's Research Leaders*" published by UW Technology and the UW eScience institute [6]. The UW Research Leaders project was an effort aimed at understanding the role of technology in current and future research projects at the University of Washington and to identify resources and services that could be of help to UW researchers in the future.

In contrast, our survey was designed to cover a larger range of topics so as to provide us with a baseline regarding researcher's work, interdisciplinarity, and how their use of advanced computational resources supports, constrains, or changes the way research is conducted and students are educated. Specifically, it was designed to provide us with information about the following aspects of each researcher who responded (a full list of the questions asked is available in the Appendix):

- *Background Information* regarding their membership in the UW community and educational history
- *Field of Research* to obtain a brief overview of their research field, the fields of their collaborators, and whether they consider themselves to be interdisciplinary
- *Student and Researcher Education* to better understand whether their interdisciplinarity affects how they educate students
- *Research Group Information* to inform us about the size and structure of their group
- *Data Use* to help us develop a sense of their data collection and sharing activities
- *Software Use* to ascertain whether or not their group develops its own software
- *Computing Resources* to help us characterize their use of advanced computational resources

These focal areas provide us with a broad overview of each researcher's group and work, along with specific details about their data and computing usage. In this report we focus on the demographics of our respondents along with their data and computing use. We do not discuss specific research areas of individual respondents or the structures of their groups to protect our informant's confidentiality.

We developed a sample of 342 researchers to whom we sent our survey in two rounds. The first round went to 300 researchers while the second went to an additional 42. When assembling our sample of 342 we used data from the UW Office of Sponsored Programs and all published University of Washington National Academy members. We wanted to be as inclusive as possible across the university when defining research leaders. As a result we used the following criteria to identify researchers to contact: 1) individuals who were in the top 20% of a College or School's research funding between 2005 and 2010, 2) membership in a National Academy, or 3) receipt of an NSF CAREER award for junior faculty between 2005 and 2011. We excluded any emeritus faculty since the aims of our larger study require active researchers and groups.

The pragmatic sampling process that we used would enable a future research team to potentially replicate such efforts.

Once our sample was compiled the survey was distributed in two rounds via the UW Catalyst WebQ tool. The first round spanned from December 2, 2012 through January 2, 2011 and was distributed to 300 researchers identified through the first two criteria listed above. A total of 103 responses were captured during this round, a 34% response rate. The second round of our survey used the same questions, but was sent to 42 UW NSF CAREER award winning faculty (43 faculty were recipients in the time frame examined however 1 had already been contacted in the first round). The second round was sent at this later time to correct the accidental exclusion of NSF CAREER awardees in the first round. The second round of the survey was available from April 19, 2011 through May 17, 2011 and 17 responses were captured during the second round, a 40% response rate. A total of 120 participants took the survey for a final response rate of 35%. Each participant was informed of their rights as a research subject in accordance with University of Washington Human Subjects Division rules. Furthermore, each researcher was offered compensation for their time with a $10 coffee gift card.

The findings presented here offer commentary on the 120 responses that we received. We examine the demographics of our respondents in our first findings section below. Questions that were asked using free response text entry boxes were qualitatively coded to group similar responses in to categories [7]. When applicable quantitative information regarding responses to questions is provided. We note that since this survey was not designed for statistical analyses to be performed.

4 Findings

The 120 responses received for our survey provide an interesting picture of the state of research at the University of Washington in 2010 and 2011. In our findings sections we examine the demographics of our respondents to emphasize the perspective our data captures before discussing findings regarding data and computational usage.

4.1 Demographics of our Respondents

Our survey sample was intentionally designed to span the University of Washington research community. The 120 responses received however represent a narrower swath that primarily covers researchers in medicine, the natural sciences, engineering, and environment. The reader should keep this in mind when examining all findings presented in this report. We had hoped to hear from scholars in the social sciences and humanities regarding their computational uses, since Digital Humanities and Social Sciences are an emerging area of data-intensive work. Unfortunately they were underrepresented in our final set of responses.

To understand our respondents' membership in the UW community we asked questions about their research position and their educational background. We asked our respondents to

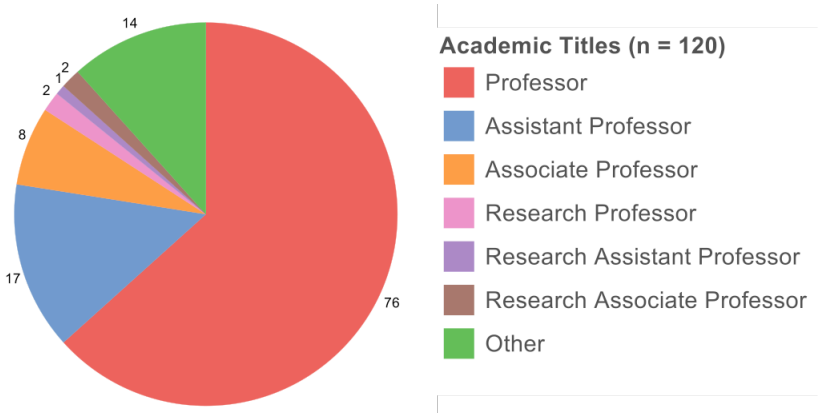

Figure 1. Academic titles held by the respondents to our survey. Taken from responses to Q3.

provide their title in a short response entry as a way to gauge the seniority of those who answered our survey. As Figure 1 shows, 76 out of 108 (70%) respondents are either tenure-track or research professors. Examining the entries we grouped our respondents according to a common hierarchy of academic ranks: Professor, Assistant Professor, Associate Professor, Research Professor, Research Assistant Professor, and Research Associate Professor. Respondents who provided an answer that did not fall under these academic titles were classified as "Other." Examples classified as other include titles such as Associate Dean, Principal Investigator, Research Scientist, and one Affiliate Professor from Australia.

In addition, we had our respondents indicate which college or school they are a member of by selecting one choice from a list of UW Colleges and Schools. The choices for this menu were derived from the University of Washington's "Academics and Research" webpage and may be viewed in the Appendix under Q4. A point to note here is that four of our respondents did not answer this question. Three of these respondents are members of the UW Applied Physics Laboratory, a distinct unit within the university, while the remaining respondent indicated that they are a member of Undergraduate Academic Affairs. As Figure 2 illustrates, the majority of our respondents are from the School of Medicine (34 of 120, 28%), the College of Arts and Sciences (28 of 120, 23%), the College of Engineering (21 of 120, 18%), and the College of the Environment (18 of 120, 15%). While respondents to this survey are heavily concentrated in the natural sciences, we were attempting to obtain responses from scholars outside of these areas with our overall sample. For example, a few of our respondents are from the Information School (3 of 120, 3%) and the College of Education (2 of 120, 2%), among the other less represented colleges or schools.

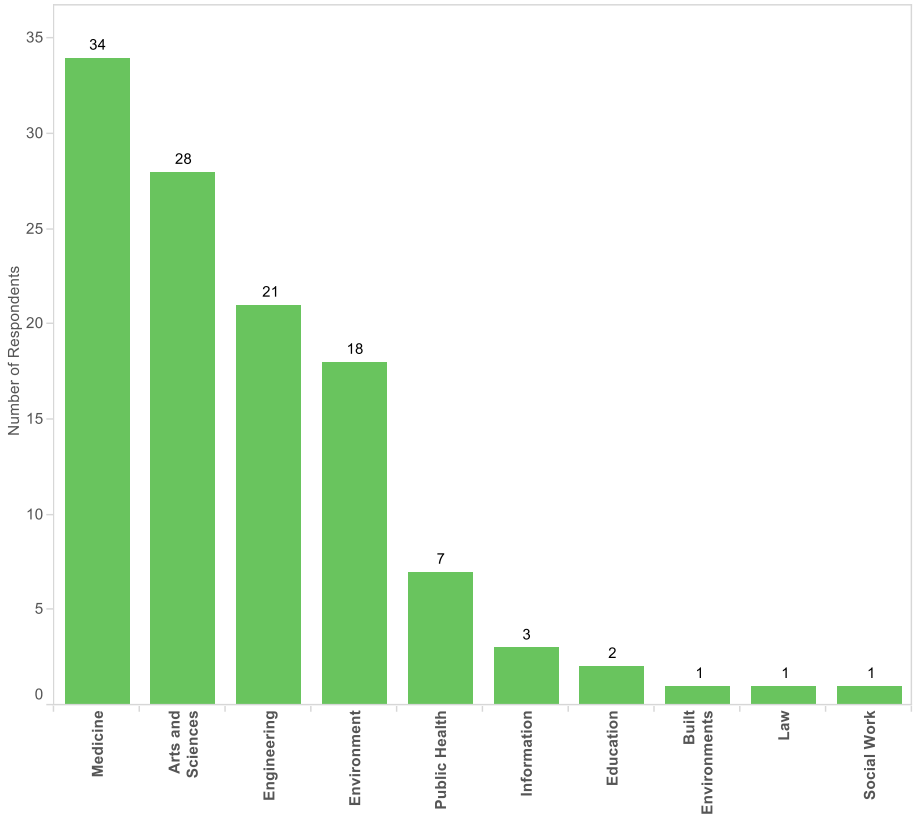

Figure 2. Breakdown of the schools and colleges respondents are members of. Taken from responses to Q4.

We also asked our respondents a set of short response questions, inquiring which department they are in and whether they had other departmental or institutional affiliations such as affiliate or adjunct faculty positions. In addition, we asked about their educational background, specifically what field their undergraduate, masters', and doctoral or professional degrees is in, and when their received their highest degree. The responses ranged widely in this question, from 1959 to 2008. The majority of our respondents received their last degree between 1970 and 1990; however 40 of our respondents received their last degree after 1990. All of these questions provide us with an overview of the demographics of our survey responses.

To understand our participants' work as a researcher at UW we included a set of questions examining their field of research along with the fields their peers and students come from. We first asked respondents to briefly describe their primary field of research. Our respondents' primary fields of research span a broad range of topics such as bacterial metabolism, ocean circulation, cardiovascular disease epidemiology, and earthquakes (among myriad other examples).

We also asked respondents to describe what field their peers and their research employees or staff members come from outside of their primary field in a short response question. In addition, we asked them to list the top five conferences and journals they most frequently submit to, once again in a short response. Responses to both of these questions were highly variable. We asked these questions to help us gain a qualitative sense of each individual's interdisciplinarity, related to our next subsection, and to help us when assessing whether we might wish to further study a group's work in the longitudinal portion of our study.

4.1.1 The Interdisciplinarity of our Respondents

In addition to the above inquiries about a researcher's background, we asked with a yes or no question as to whether they consider their research to be interdisciplinary. This was of interest due to the overall project's aim of studying interdisciplinary scholars. A total of 108 out 120 (77%) of our respondents consider their research to be interdisciplinary. While we did not interact further with the 12 who answered no it was striking that an overwhelming majority considered themselves to be an interdisciplinary scholar and so few did not. How these individuals define interdisciplinary is more than likely variable, for example they might consider themselves to draw upon methods or practices from fields related to theirs but these might all fall within a larger overarching discipline leading an outsider to not view their work as all that interdisciplinary. We did not offer a definition for the term so it was up for interpretation by the respondent.

If an individual answered that they consider their research to be interdisciplinary then we asked four follow-up questions. These four questions ask respondents if their interdisciplinarity affects their approaches to training future researchers in their field. We inquired with one yes, no and not applicable question and three yes or no questions as to whether the interdisciplinary nature of their research affects how they approach classroom education of their students, how they educate researchers working in their lab, how they mentor advises, and whether it affects their decisions when recruiting lab members (see Q20-Q23 in the appendix for the exact questions).

The majority of our respondents indicated that the interdisciplinary nature of their research affects their approaches to classroom education (84 of 108, 78%), as seen in Figure 3 column 1. Fifteen respondents indicated Not Applicable, which leads the reader to conclude that they are not engaged in classroom education. However, since follow-up questions were not asked we cannot be sure. Second, 92% of our respondents (97 out of 106 responses), Figure 3 column 2, indicated that their research affects how they educate researchers working in their

lab. Third, 94% (101 of 107 responses), Figure 3 column 3, indicated that the interdisciplinary nature of their research affects how they mentor their advisees. Fourth and finally, 87% of the responses (94 out of 106), Figure 3 column 4, indicated that being interdisciplinary affects how they recruit members to their lab. While we did not ask detailed follow-up questions to add context to these answers the high percentage of yes answers to each does suggest that researchers at UW find the interdisciplinary nature of their research to impact their education and advising work.

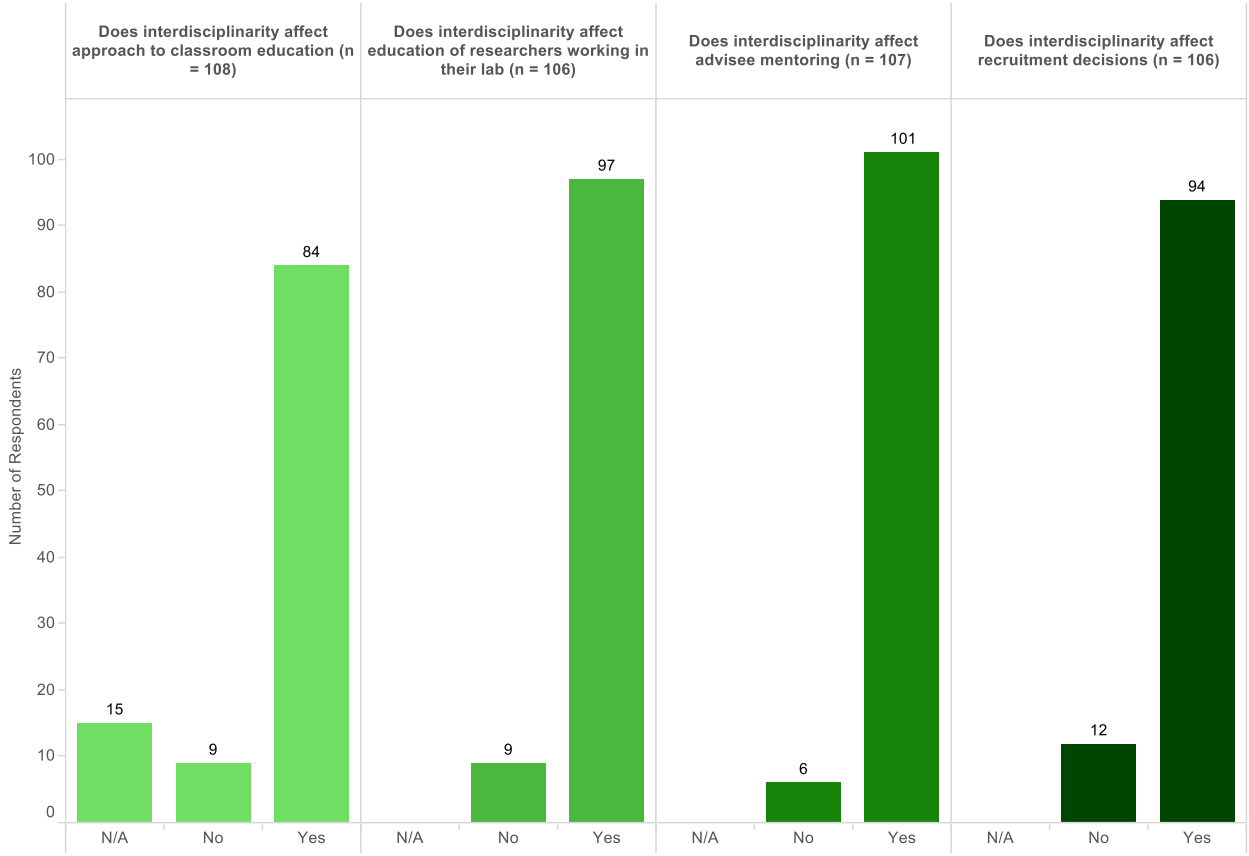

Figure 3. The effect of interdisciplinary nature of research on researchers' approach to classroom education of their students (column 1), education of researchers working in their lab (column 2), advisee mentoring (column 3), and decisions when recruiting lab members (column 4). Taken from responses to Q20-Q23.

4.1.2 Research Group Structure

The final section of background questions that we asked our respondents examined the research agenda of the group, the number of funded grants they had at the time, and the number of members of different ranks in the group (see Q24-Q31 in the Appendix). These questions were designed to help us develop a broad understanding of the work being undertaken and the size of the group doing the work. In this report we do not examine the agendas of the groups to protect the privacy of our survey participants.

We inquired about the number of each respondent's funded grants as one element in our understanding of their group's size. The number of funded grants that respondents had at the time of the survey ranged from a minimum of zero up to a maximum of 51, with an average of six grants. The majority of responses fell between 1 and 15 grants per PI, with only five responses above 15. With this question we were looking to obtain information that would help us to ensure that we can follow groups with a variety of projects taking place. The number of

members in a group of each category varies widely by respondent, with respondents consistently having doctoral and post-doctoral students along with research scientists in their group. Many members also reported engaging undergraduate researchers in their group's work. However, it is interesting to note that most respondents did not indicate having many Masters students as researchers in their group.

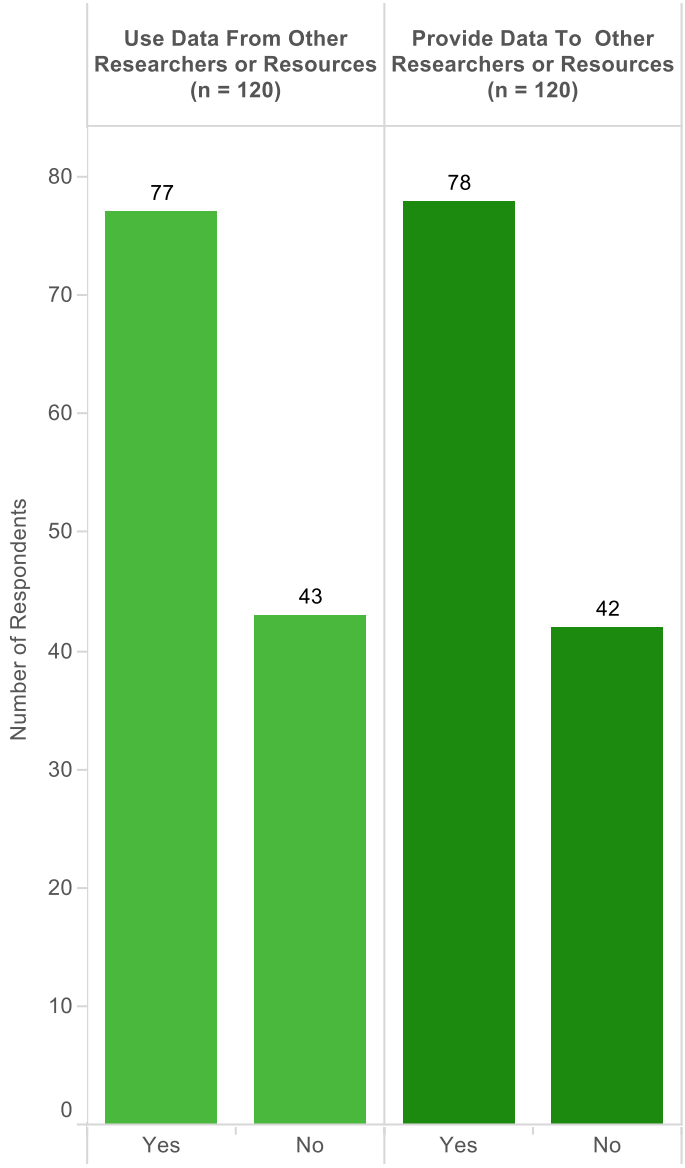

Figure 4. Comparison of how often researchers use data from (left) and distribute data (right) to other researchers or resources to answer research questions. Taken from responses to Q33-Q34.

4.2 Characterizing Participant's Data Use

One of the main goals of our longitudinal study is to follow data-intensive research work. Key to our survey then was obtaining an understanding of each researcher's use of data. Therefore, in this section of the survey we asked researchers about their data use (see Q32-Q39 in the Appendix).

We first inquired with a yes or no question as to whether respondents were collecting their own data stores. This question enabled us to make sure that we study empiricists who are involved in the production of new datasets. A total of 108 of the 120 respondents (90%) answered yes that they collect their own data stores or sets. This showed us that most of the survey respondents are engaged in some amount of empirical work themselves.

We also inquired with a free response question as to where researchers were housing their data. The majority of responses indicated some form of server locally in their group's workspace or here at UW. However, some respondents did indicate that they use systems such as cloud computing, a national repository or data archive, or a collaborating institution's server or resource. Beyond these two questions we inquired about respondents' data sharing activities as well as their data handling.

We next asked four questions about these researchers' data sharing. First we inquired as to how often participants were using data from, or providing data to, other researchers in order to answer their research questions. As Figure 4 shows, almost two-thirds of the respondents often use data (77 of 120 responses, 64%) and provide data (78 of 120 responses, 65%) to other researchers or resources to answer their research questions. Second, we wished to develop a broad understanding of how many of a researcher's funding agencies and publication venues require them to share their data. Using two Likert-scale questions we asked

whether All, Some, or None of these entities require respondents to share their data. The responses to this question were mixed when compared to our other two questions about data sharing. As Figure 5 shows, most participants indicated that their funding sources do require them to share data (39 responded All, 53 responded Some, out of 119). In addition, many respondents answered that publication venues were requiring them to share their data to have a paper published at the time of the survey (22 responded All, 44 responded Some, of 119). It is however interesting to note that 53 of the 119 responses indicated that none of their publication venues were requiring them to share their data to be published. These four questions as a set suggest that researchers were accustomed to sharing their data at some level in 2010 and 2011.

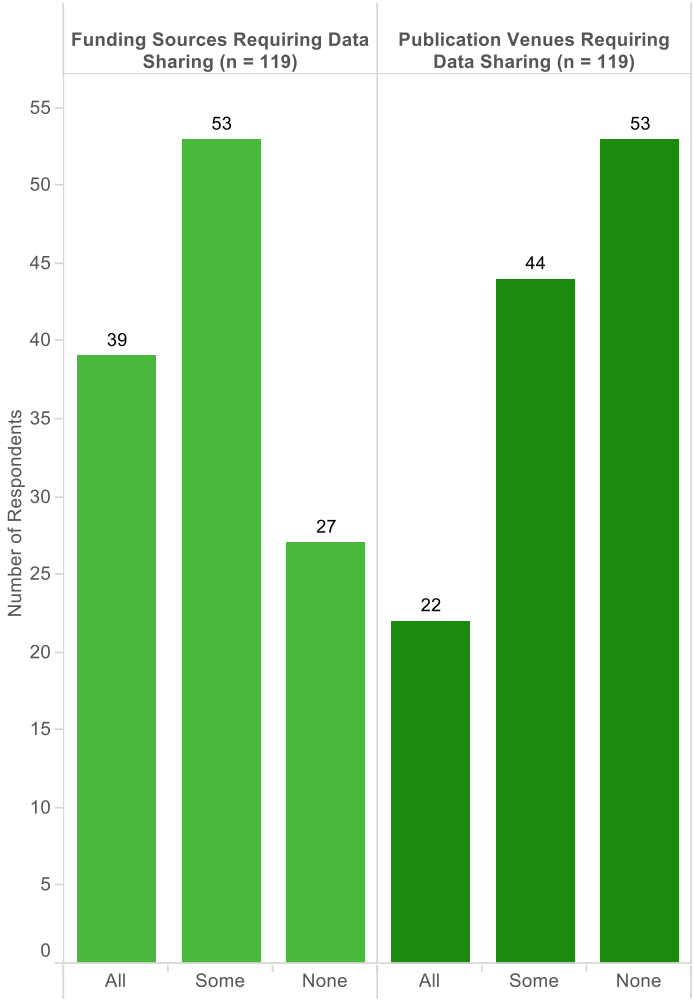

Figure 5. Number of researchers whose funding sources (left) and publication venues (right) require them to share data. Taken from response to Q35 and Q36.

Finally, we asked two questions about the amount of time respondents spend engaged in "data handling" activities[2]. We purposefully did not define what activities count as handling data, other than to note that they should not include the amount of time spent collecting or analyzing data. Our intent was to try and begin to tease out how much time is spent doing "other" work with data, such as processing datasets after they are collected to prepare them for analysis. First, we

[2] We thank Bill Howe of the University of Washington eScience Institute for suggesting that the term "data handling" be used for this question.

asked the survey participant to enter the percentage of hours per week that were spent by their group handling data. These responses varied wildly, from 0% of a group's time per week to 90%. We followed up this question by asking the respondents to indicate using a Likert-scale how this percentage of hours had changed in the last five years. The responses to this question were mixed as is visible in Figure 6. Over half (66 of 116 responses, 57%) of the participants indicated that their data handling percentage had increased by some amount, while 41% (48 of 116 responses) of the participants indicated that there had been no change. It is hard to say exactly how participants were spending more time having to "handle" or process data than they previously were, since we cannot know what activities each person counts as handling. The responses to this question do suggest that there is some change in the amount of time participants must spend working with their data. This change may perhaps be taken as an indicator of increasing data-intensiveness for these researchers. However, it may also simply indicate increased complexity in their work with the data that they already have.

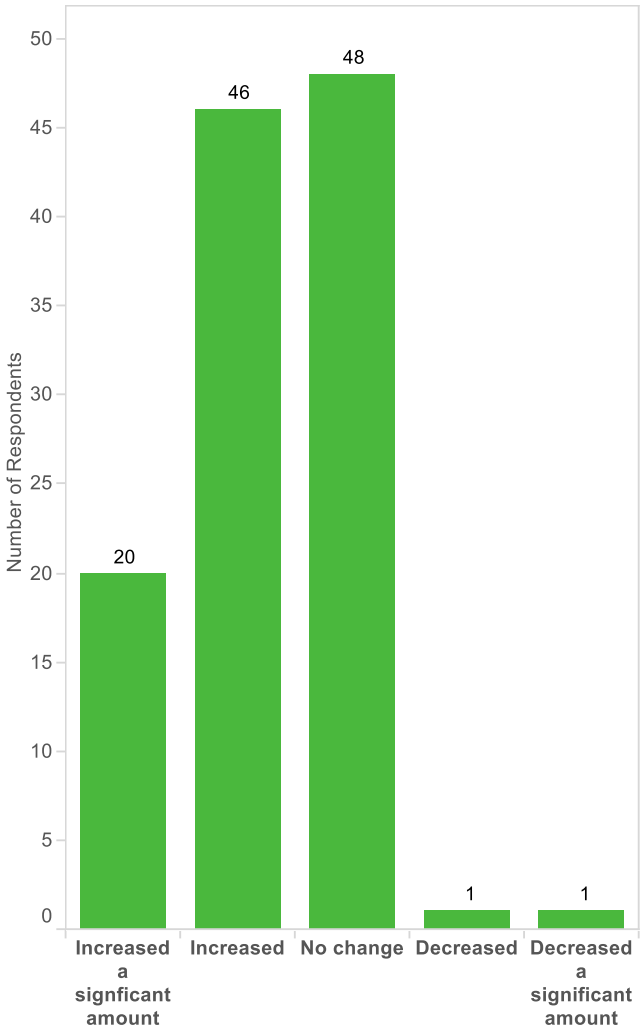

Figure 6. Researchers' view of how time spent on data handling in their research group has changed in the last five years. Taken from responses to Q39.

4.3 Examining Researcher's Software and Computational Use

The final two sections of our survey inquired about researcher's software use and computing resources. We briefly asked whether or not their group develops software as a part of their research work before delving into their data storage and transfer requirements, collaboration tools, and sources of computing resources.

4.3.1 Software Development and Use

To understand our respondent's software use, we inquired with a yes or no question as to whether their research group develops software. The responses to the question were divided almost perfectly in half. Sixty of the 119 respondents (50%) said yes they develop software while 59 of the 119 respondents (50% again) said that they did not. Those who indicated that they develop software were asked to fill out a free response question to list the software that their group develops. Responses were varied, listing a variety of software. However, two responses for this question were common. Thirteen of our respondents (13 of 102 responses, 15%) indicated that they develop their own software or write their own codes and scripts to collect or analyze data internal to their group. In addition, nine of our respondents (9 of 102 responses, 9%) indicated that they use specialized software for data analysis such as

Interactive Data Language (IDL), MATLAB, and COMSOL Multiphysics. Additional examples include software for quantitative analysis of three-dimensional microscopic images, the use of software to control instruments deployed in the ocean, or epidemiologic visualization software for public health practitioners.

4.3.2 Data Storage and Transfer Activities

To build upon our earlier questions about data use and to connect such activities to their computational requirements we inquired about each individual's data storage and transfer activities through four Likert-scale questions (see Q42-Q45 in the Appendix). The first question about data storage asked how much storage they were using to store their data at that time while the second asked about the volume of data they were producing per month.

The amount of data being stored by a research group varied among the 109 responses that we received, see Figure 7. Two answers were prominent for this question. Thirty of the respondents indicated that they were storing between 1 and 50 gigabytes of data while in contrast 40 responded that they were storing more than a terabyte of data. It is highly possible that three years later the majority of responses would be skewed towards higher amounts if scientific research is indeed becoming more data-intensive, if measured purely by dataset size.

Likewise, the answers regarding the amount of data an individual produces per month varied widely. The majority of responses fell between less than 10 megabytes and up to 50 gigabytes, see Figure 7, however 24 researchers answered that they were producing 50 or more gigabytes of data per month. Aside from noting the variability in quantities of data produced per month it is not possible to directly infer much from these answers. It is of interest to examine connections between the quantities being produced and different scientific fields in future studies.

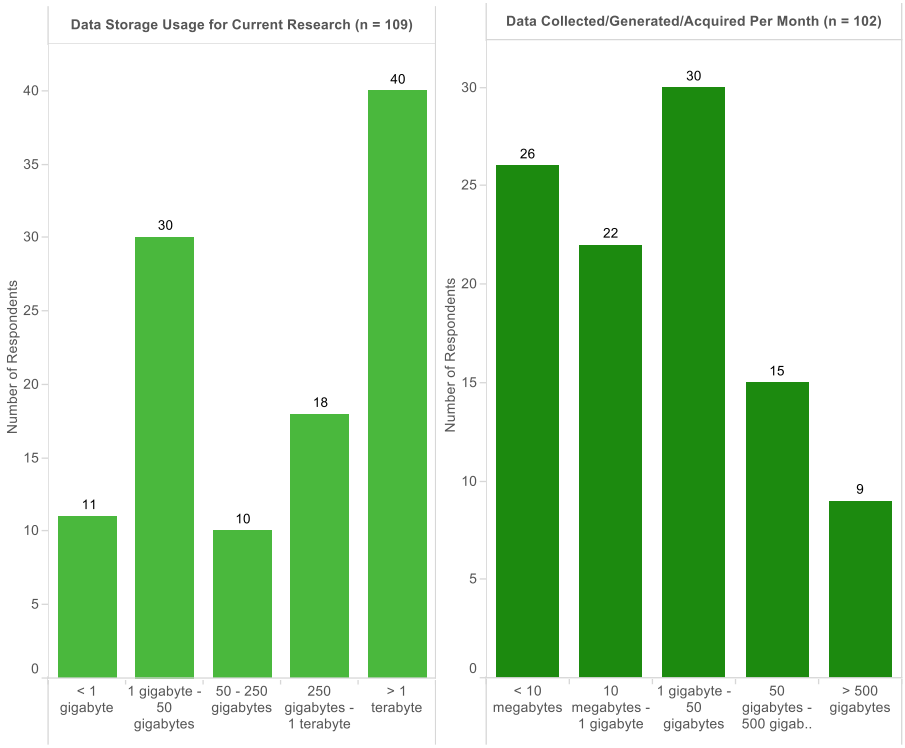

Figure 7. Comparison between researchers' use of data storage (left) and volume of data collection per month (right). Taken from responses to Q42 and Q43.

Finally, we asked respondents to answer a Likert-scale question about the frequency with which they transfer large datasets. This question was designed to give us a rough understanding of the need to move data around the UW campus and off-campus. As Figure 8 shows, most of our respondents indicated that they seldom or never transfer large datasets. We did not ask any follow-up questions to provide context for these responses so it is hard to say whether bandwidth constraints play a role in these answers or if there was simply no need. It could be the case that researchers with large datasets simply store them on remote computing resources and work with them on such systems, rather than locally housing them.

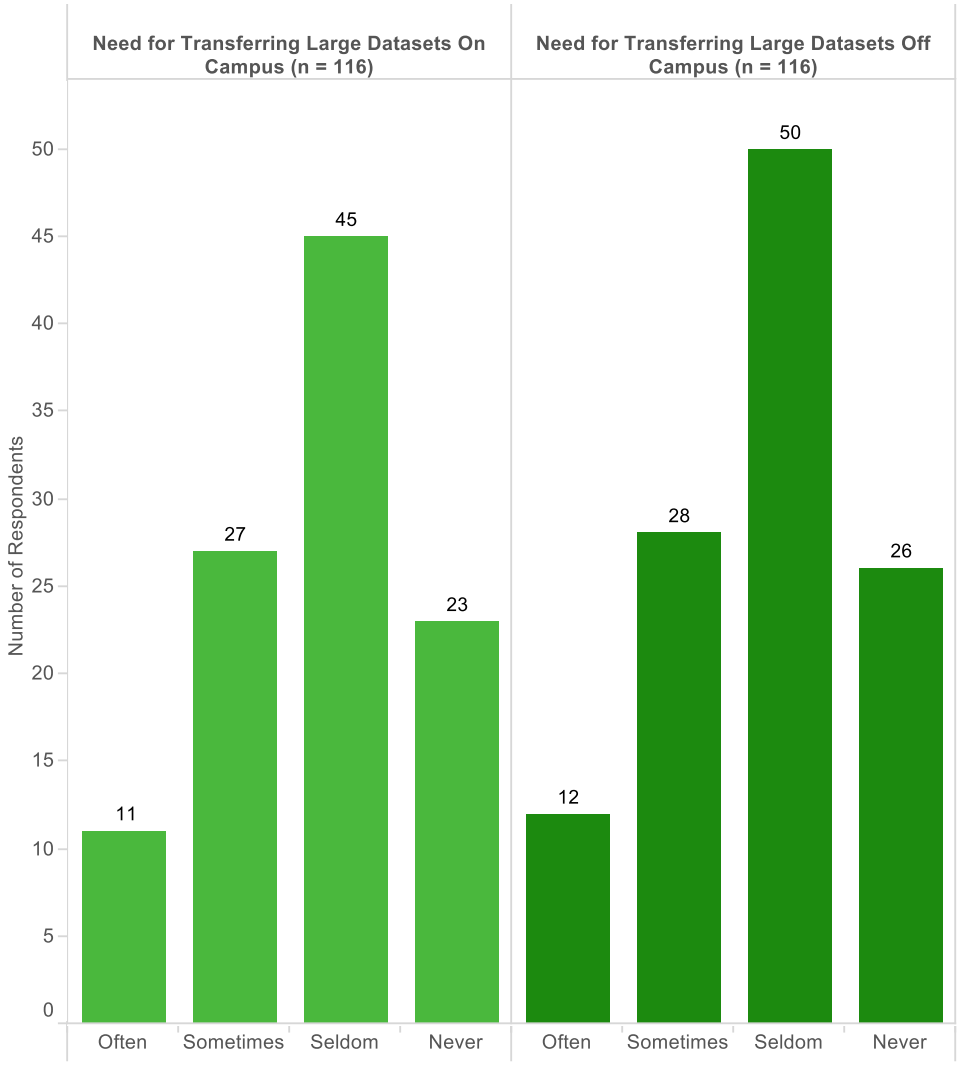

Figure 8. Frequency of researchers' need to to transfer large datas sets on and off campus. Taken from responses to Q44 and Q45.

4.3.3 Collaboration Tools

In our next cluster of questions we used two free response questions to inquire about the widely available and domain specific tools researchers were using to collaborate. Respondents were asked to list the tools that they use to work with collaborators. We listed a few examples such as Wikipedia, Skype, email, teleconferencing, and screen sharing tools. Since the individuals answering these questions were able to enter any combination of tools or systems we clustered similar tools into overarching categories.

Figure 9 illustrates that the majority of our respondents use email, Internet conferencing tools such as Skype, Adobe Connect, GoToMeeting Webinar, web conferencing, and screen sharing tools, as well as teleconferencing, and web based collaboration tools such as Wikipedia, Google Docs, and CMS. Our respondents also noted that they use file exchange tools such as Dropbox and FTP, and chat tools such as IM and iChat to collaborate with other researchers. A couple of our respondents did comment that they meet face to face with their collaborators. The use of UW resources such as Catalyst Workspace, Center for Studies in Demography and Ecology (CSDE), and WebQ also came up in these responses.

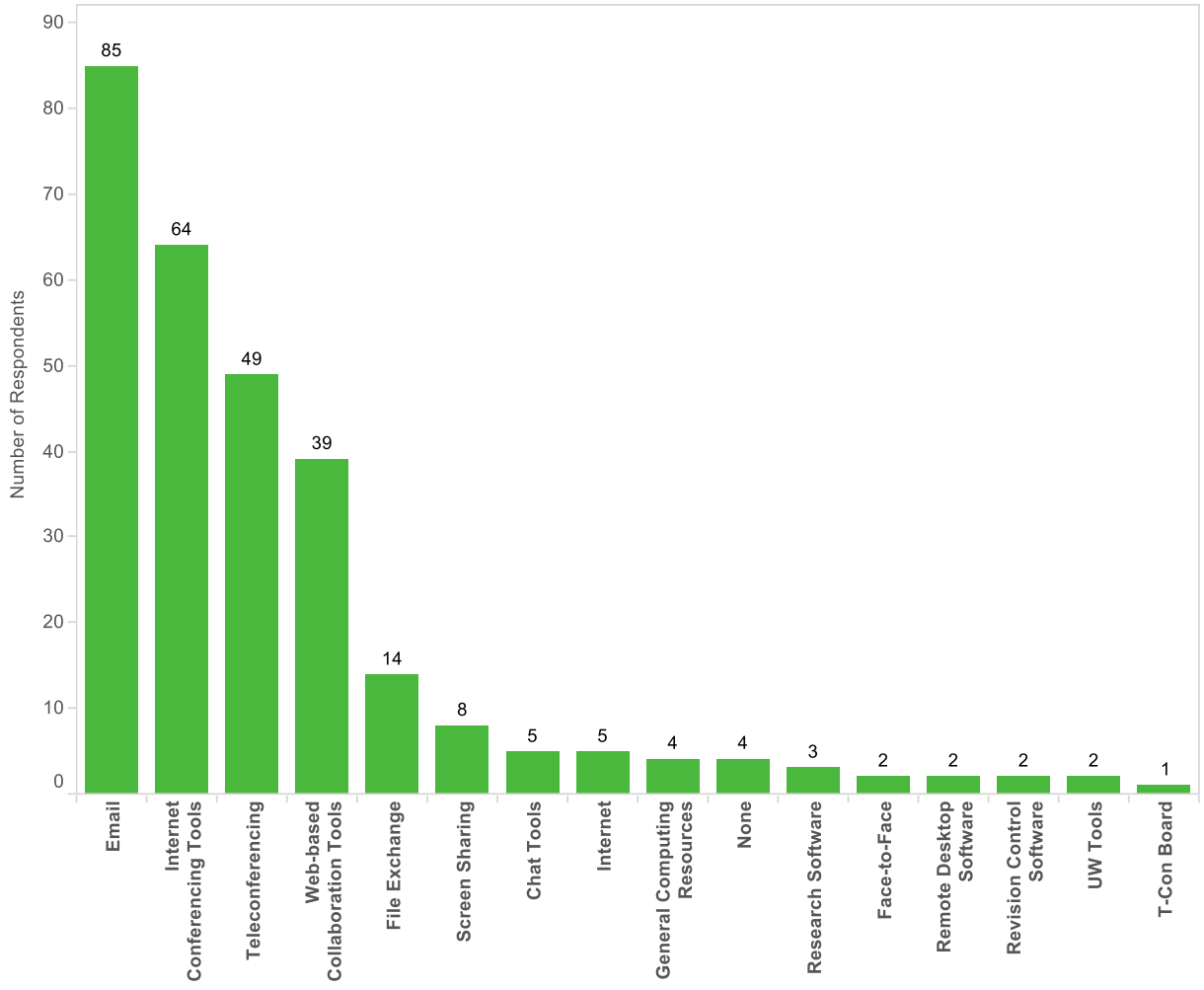

Figure 9. Categories of widely available tools that researchers use to work with collaborators. Taken from responses to Q46.

4.3.4 Computing Resources

The final cluster of computing questions that we asked our respondents inquired about the sources of their computing resources (see Q48-Q50 in the Appendix). The first question asked where the researcher obtains computing resources. We listed Hyak, Teragrid, college IT department, Amazon EC2, and Microsoft Azure as examples of what might be relevant answers since we used a free response question. We next used a Likert-scale question to gauge respondent's familiarity with Cloud Computing since it was beginning to be more widely used at the time of the survey. Finally, we inquired as to whom researchers were turning to for expertise

on computing resources using another free response question. For both of the free response questions we categorized answers that are discussed below.

Looking at where researchers were obtaining computing resources the majority of our respondents refer to their colleagues, specifically their local, in house, and group members as resources, see Figure 10. Most of our respondents obtain their computational resources through IT such as their college IT, department IT, UW IT, local IT, lab IT, and own IT group. Furthermore, a few of our respondents indicated that they obtain their resources commercially such as Amazon, SAS, and Oracle, as well as federal agencies such as TeraGrid, OpenScienceGrid, National Institutes of Health (NIH), National Energy Research Scientific Computing Center (NERSC), National Science Foundation (NSF), and Centers for Disease Control and Prevention (CDC). One individual obtains their computational resources independently, one through books at the bookstore, and one through their research grant. The variability in the answers to this question makes it clear that respondents interpreted computing resources to mean a wide variety of things. We do however see that the places people obtain computational resources from align with their responses to our question about who they turn to for expertise.

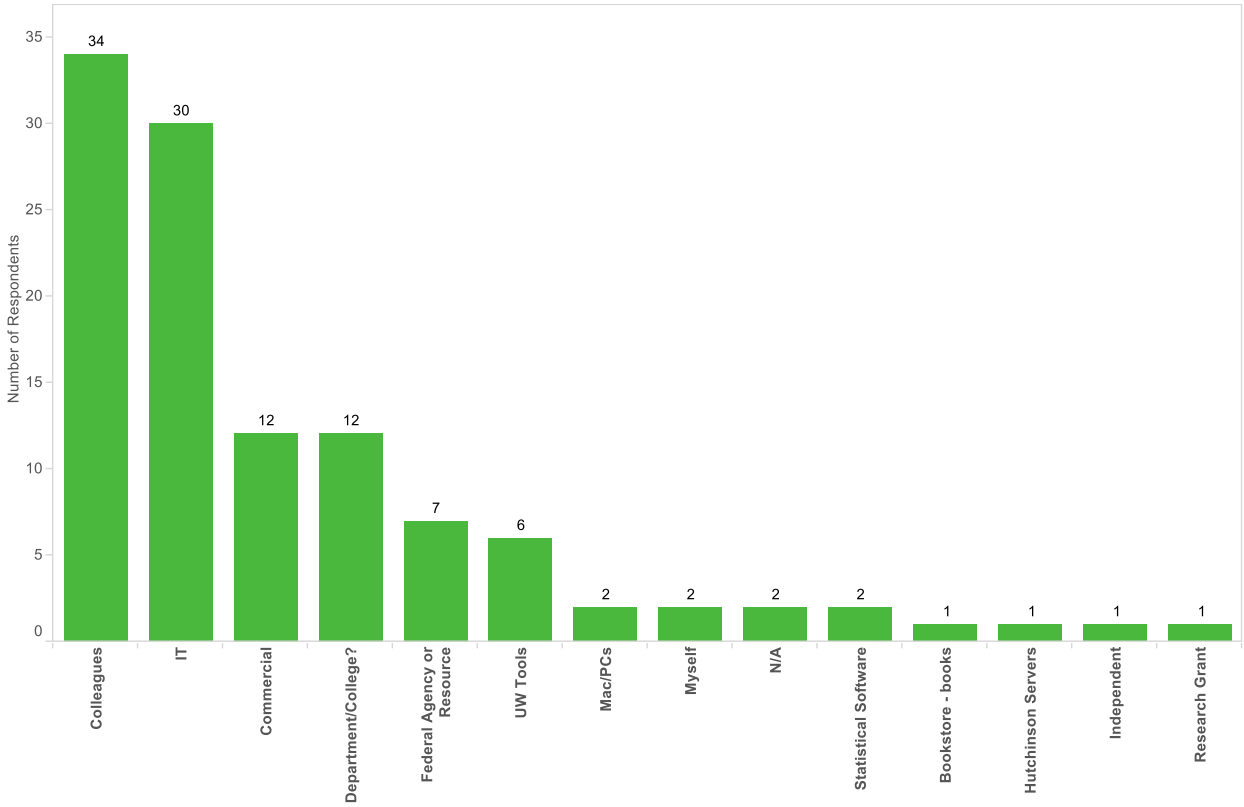

Figure 10. Location of computational resources obtained by researchers. Taken from responses to Q48.

In our last survey question we asked our respondents whom they go to for expertise on computing resources. As examples we listed graduate students, UW IT, and the eScience Institute. Most of our respondents responded that they go to some form of IT group for computing resources, see Figure 11. Out of the free response entries, IT arose broadly as a category that includes college IT, department IT, UW IT, and group-specific entities such as “local IT,” “group IT,” and “own hired IT.” Most of our respondents also go to their graduate and post-doctoral students, and members of their own group. A few of our respondents use

community organizations such as NERSC, Joint Institute for the Study of the Atmosphere (JISAO), the Pacific Northwest National Laboratory (PNNL), and the Oak Ridge National Laboratory (ORNL) as sources of expertise. In addition to UW tools and the eScience Institute, a small handful of our respondents refer to outside peers for expertise on computing resources such as their friends, spouse, and peers outside of UW.

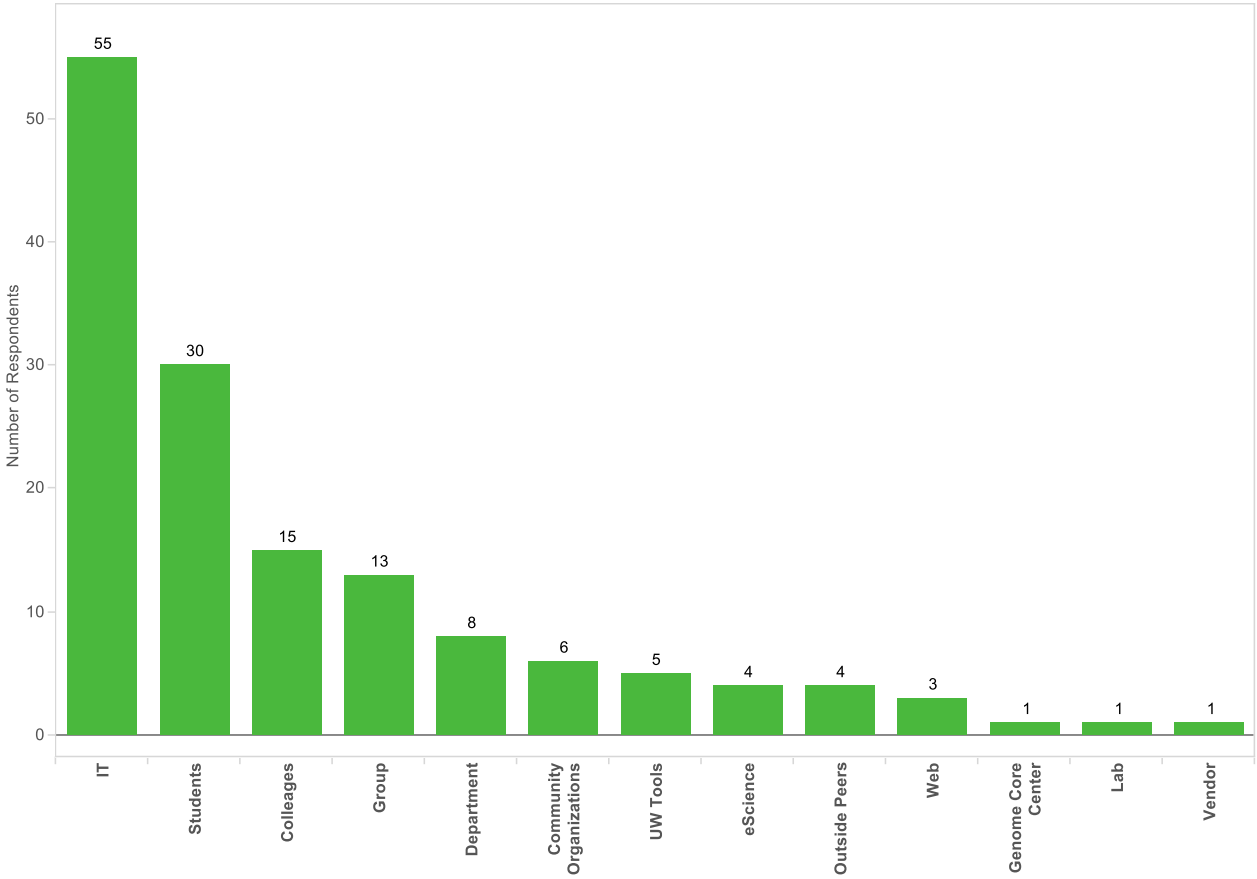

Figure 11. List of individuals rsearchers go to for expertise on computing resources. Taken from responses to Q50.

Finally, before asking whom researchers turn to for computing expertise we inquired about their experience with cloud computing in their research group through a Likert-scale question. We were interested in seeing how common it would be for researchers to have experience with such services at this point in time. Unsurprisingly, a total of 80 out of 108 respondents (74%) indicated that they had either no experience with or were unfamiliar with cloud computing, see Figure 12. Only 38 of the 108 respondents (35%) had some or extensive experience. Three years later it may be the case that more researchers would have more extensive experience using cloud computing resources in their research, however the high reliance on local IT services might keep this from being the case.

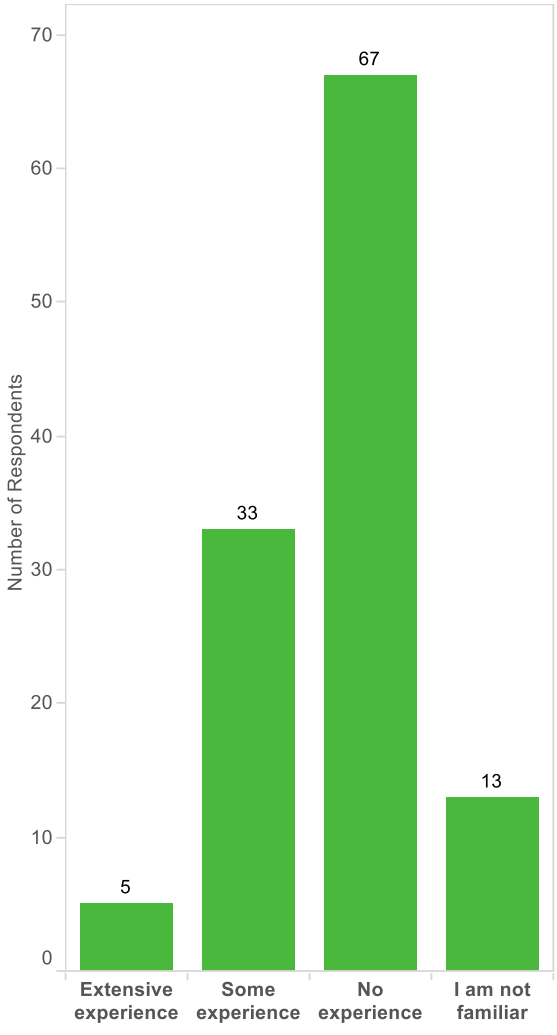

Figure 12. Researchers' experience with cloud computing in their research group. Taken from response to Q49.

5 Summary

The responses to our survey offer a look at the data intensiveness of a swath of the University of Washington research community. While by no means representative of every scholar at the university they do offer an interesting look at the use of data and computational resources by many researchers at one point in time.

The number of responses indicating the need to share data as a result of funding agency or publication venue requirements points to policy changes beginning the process of encouraging research to be more open. The fact that 50 percent of our respondents develop software as a part of their research foregrounds many scholar's continued need for customized tools to support their scientific inquiry. However the use of widely available and often simple software to support collaboration suggests that researchers are making due with much software off the shelf, rather than expending the effort to produce something custom for their work.

The quantities of data being generated per month being a gigabyte or more by 54 of the 102 researchers who responded to our survey illustrates the growing size of datasets in scholar's work, in addition to the relatively large datasets they indicated that they already had at the time. The lack of a need to transfer datasets around campus or off of campus aligns with respondent's answers that many of their computing resources are local to their group or otherwise maintained by a larger organization. Likewise the reliance upon students and local information technology groups for computing expertise points to the opportunity for University of Washington entities to offer continued training support, to increase awareness of already available resources, as part of their missions. While our survey was not designed to directly examine areas for the university to improve its computational offerings and support, as the work behind the Conversations with Research Leaders report was, we do find that at a broad level the scholars who answered our survey were relying upon local resources and knowledge to support their work. The mention of the eScience Institute as a resource for researchers only two years after its founding at the university points to its potential to impact a large amount of the work here at UW, in addition to the support already offered by various information technology organizations around the campus.

6 Acknowledgments

The authors would like to thank all of the researchers who took our survey. In addition, Bill Howe of the University of Washington eScience Institute, Cara Lane of the UW Office of Learning & Scholarly Technologies, the University of Washington Office of Sponsored Programs, and David Castner Associate Dean of Infrastructure in the College of Engineering all offered valuable feedback and data that informed our survey. Toni Ferro (HCDE PhD student) contributed to the development and initial analysis of the survey in 2010-2011.

This material is based upon work supported by the National Science Foundation under Grant Number IIS-0954088, an NSF CAREER award for junior faculty. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7 References

[1] University of Washington. (2012). Academics and Research. http://www.washington.edu/discover/academics. Last accessed Feb. 26, 2014.

[2] The Royal Society Science Policy Centre. (2012). Science as an Open Enterprise (DES24782): UK Royal Society. http://royalsociety.org/policy/projects/science-public-enterprise/report/

[3] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., et al. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington D.C.: National Science Foundation.

[4] Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). Understanding Infrastructure: Dynamics, Tensions, and Design (Workshop Report).

[5] Jirotka, M., Lee, C. P., & Olson, G. M. Supporting Scientific Collaboration: Methods, Tools and Concepts. Computer Supported Cooperative Work (CSCW), 22, 4-6 (2013), 667-715. DOI= http://dx.doi.org/10.1007/s10606-012-9184-0.

[6] Fournier, J., Koester, G., Giacomini, C., Lewis, T., & Washington, W. (2009). *Conversations with the University of Washington's Research Leaders: Final Report.* University of Washington Technology & eScience Institute: University of Washington. http://www.washington.edu/itconnect/learn/research/

[7] Miles, M. B., & Huberman, A. M. Qualitative data analysis: An expanded sourcebook. Sage Publications, Incorporated, 1994.

8 Appendix: Survey Questions

The questions from the survey are presented below along with the question type and any constrained answer lists. The pages of the survey were divided into sections. The headings are presented in bold between questions.

Note: If not otherwise noted the question was a free response text entry.

Q1. IRB Acknowledgment

Background Information

Q2. What is your name?
- Answer required

Q3. What is your title?

Q4. Which college or school are you in?
- *Pick list:*
  - Arts and Sciences
  - Built Environments
  - Business
  - Dentistry
  - Education
  - Engineering
  - Environment
  - Graduate
  - Information
  - Law
  - Medicine
  - Nursing
  - Pharmacy
  - Public Affairs
  - Public Health
  - Social Work
- This list was pulled from http://www.washington.edu/discover/academics/ in Autumn 2010. Through the administration of the survey we determined this is not comprehensive, for example the Applied Physics Laboratory is considered to be a distinct unit.

Q5. Which department are you in?

Q6. What other departmental or institutional affiliations do you have, if any?

Q7. What field is your undergraduate degree in?

Q8. What field is your masters degree in?

Q9. What field is your doctoral or professional degree in?

Q10. In what field(s) have you held post-doctoral positions?

Q11. What is the year in which you received your highest degree?
- Numeric entry field

Field of Research

Q12. Please describe your primary field of research in two to four sentences.

Q13. Do you have a secondary field of research?
- Yes/No radio buttons

Q14. If you answered yes, please describe your secondary field of research.

Q15. On your currently funded projects, what fields other than your primary field do your peers come from?

Q16. On your currently funded projects, what fields other than your primary field do your students come from?

Q17. On your currently funded projects, what fields other than your primary field do your research employees or staff members come from?

Q18. What are the top 5 conferences and journals you most frequently submit to?

Q19. Do you consider your research to be interdisciplinary?
- Yes/No radio buttons

- If an individual answered No they were not asked Q20-23.

Student and Researcher Education

Q20. Does the interdisciplinary nature of your research affect how you approach classroom education of your students?
- Yes/No/Not Applicable radio buttons

Q21. Does the interdisciplinary nature of your research affect how you educate researchers working in your lab?
- Yes/No radio buttons

Q22. Does the interdisciplinary nature of your research affect how you mentor advisees?
- Yes/No radio buttons

Q23. Does the interdisciplinary nature of your research affect your decisions when recruiting lab members?
- Yes/No radio buttons

Research Group Information

Q24. What is the name of your research group?

Q25. Please describe the research agenda(s) that drives the projects your research group undertakes.

Q26. How many funded grants does your research group currently have?
- Numeric entry field

Q27. How many post-doctoral researchers are members of your research group?
- Numeric entry field

Q28. How many research scientists are members of your research group, other than listed as post-doctoral researchers above?
- Numeric entry field

Q29. How many doctoral student researchers participate in your research group?
- Numeric entry field

Q30. How many masters student researchers participate in your research group?
- Numeric entry field

Q31. How many undergraduate student researchers participate in your research group?
- Numeric entry field

Data Use

Q32. Do you collect your own data stores or sets?
- Yes/No radio buttons

Q33. Do you often use data from other researchers or resources to be able to answer your research questions?
- Yes/No radio buttons

Q34. Do you often provide data to other researchers or resources in order for them to answer their research questions?
- Yes/No radio buttons

Q35. How many of your funding sources require you to share data?
- All/Some/None Likert-scale

Q36. How many of your publication venues require you to share data?
- All/Some/None Likert-scale

Q37. Where do you house your data stores or sets?

Q38. What percentage of hours per week are spent by your research group handling data? This does **NOT** include collecting or analyzing data.
- Numeric entry field

Q39. How has this data handling percentage changed in the last five years?
- Likert-scale with the following options:
  - Increased a significant amount

- Increased
- No change
- Decreased
- Decreased a significant amount

Software Use

Q40. Does your research group develop software?
- Yes/No radio buttons

Q41. If you answered yes, please list the software you develop. Please include the name and its purpose.

Computing Resources

Q42. How much data storage are you using for your current research?
- Likert-scale with the following options:
  - < 1 gigabyte
  - 1 – 50 gigabytes
  - 50 – 250 gigabytes
  - 250 gigabytes – 1 terabyte
  - > 1 terabyte

Q43. Approximately how much data do you collect/generate/acquire per month?
- Likert-scale with the following options:
  - < 10 megabytes
  - 10 megabytes – 1 gigabytes
  - 1 gigabyte – 50 gigabytes
  - 50 gigabytes – 500 gigabytes
  - > 500 gigabytes

Q44. How frequently do you need to transfer large datasets around campus?
- Likert-scale with the following options:
  - Often
  - Sometimes
  - Seldom
  - Never

Q45. How frequently do you need to transfer large datasets off campus?
- Likert-scale with the following options:
  - Often
  - Sometimes
  - Seldom
  - Never

Q46. What widely available tools do you use to work with collaborators?

Q47. What domain-specific tools do you use to work with collaborators?

Q48. Where do you obtain your computational resources?

Q49. How much experience does your research group have with cloud computing?
- Likert-scale with the following options:
  - Extensive experience
  - Some experience
  - No experience
  - I am not familiar with cloud computing

Q50. Who do you go to for expertise on computing resources?

Summary

Q51. To receive your $10 gift card please provide your University of Washington box number.