UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

University of Washington Data Science Environment
Reproducibility and Open Science
Final Evaluation Findings

June 24, 2015

Prepared by

Table of Contents Page

Table of Figures

Page

Table of Tables

Reproducibility and Open Science Executive Summary

June 24, 2015
Prepared by data2insight LLC

This evaluation report communicates findings from the University of Washington Data Science Environment (UW DSE) survey administered in March, 2015 and the UW DSE key informant interviews conducted February-March, 2015. The survey was administered in five parts; each part was designed specifically for five of the six working groups. A total of 580 people responded to at least one part of the survey. Evaluators interviewed 21 key informants including department chairs, junior and senior faculty, staff, eScience Institute Data Scientists, postdoctoral researchers, and graduate students (Appendix A). In partnership with UW DSE leaders, evaluators identified survey respondents and interviewees. The UW DSE survey and interview data collection instruments were designed to answer the following evaluation questions for the Reproducibility and Open Science (ROS) working group, summarized below.

1. What does ROS mean to you and your peers? What do faculty, students, and scientists value about ROS?

All interviewees described "reproducibility" as the ability to repeat an experiment using the same inputs and achieve the same result. More than 80% of survey respondents agreed or strongly agreed with the following statements about **reproducibility**:

- Reproducibility supports greater scientific validity and integrity (97%)
- Reproducible research is a research result that can be replicated by another investigator (95%)
- Reproducibility is a basic principle of the scientific method (88%)
- Reproducibility is the ability of an independent researcher to recreate an entire experiment using the same procedures (86%)

The highest level of disagreement (15%) regarding reproducibility was with the statement: "Reproducibility can contribute to rapid discovery." However, 64% of respondents agreed or strongly agreed with this statement.

Interviewees described "open science" as sharing software, data sets, code, and publications with all people, free of charge. More than 80% of survey respondents agreed or strongly agreed with the following statements about **open science**:

- Standard open science protocols for data and code sharing should be developed (84%)
- Sharing data used in published research should be a default practice (84%)
- Open science can increase the impact of research performed (80%)
- Open science can increase the impact of software developed (80%)

There was the most disagreement among survey respondents with these statements:

- Not all data needed to be shared with all people (24% disagreed; 70% agreed or strongly agreed)
- Code developed as a part of research should be made available with every publication and/or presentation (19% disagreed; 65% agreed or strongly agreed)

Most interviewees felt that sharing data and research depended on the field, the content, and potential human risks associated with sharing data.

2. What are the current practices and barriers to effective ROS code and data sharing?

Over half (52%) of interviewees were familiar with the ROS guidelines.

Survey respondents reported using the following **software tools, environments, or data management practices** sometimes or always:

- Scripts to ensure replicable computations (70%)
- A version control system (57%)
- Open source code repositories (56%)
- Notebook environments to capture sequence of instructions (47%)
- Repositories to deposit data used in publications (42%)

The least frequently used tools and practices were workflow management (9%), literate programming (18%), and auto tools (19%).

Survey respondents indicated that they sometimes or always experienced the following **top five barriers to code and data sharing**:

- Sharing restrictions from collaborators (66%)
- Lack of existing data sets (64%)
- Concerns about having ideas stolen (64%)
- Concerns about having data stolen (60%)
- Sharing restrictions from suppliers of data (58%)

The most uncommon barriers to code and data sharing were concerns about data and/or code not being reproducible if shared (33%), financial incentives to keep data/code private (34%), and university policies (41%).

Many interviewees (48%) shared survey respondents' concerns about having data and ideas stolen. Over half of interviewees mentioned the time required to make code, data, and research open and reproducible as a barrier to code and data sharing. Over 40% of interviewees also mentioned human subjects or biology-related restrictions as well as a lack of knowledge about how to share data as barriers to code and data sharing.

This comment from a UW faculty member captured a theme that emerged in interviews: there are multiple dimensions to reproducibility: *"it's not like a single thing. I don't think there's a badge that says this work is or is not reproducible. I think that it can be more or less reproducible along a lot of dimensions."* Furthermore, two interviewees identified distinctions between 'repeatability or replicability,' and 'reproducibility;' these interviewees stated that 'repeatability or replicability' referred to the experimenters' ability to repeat the experiment, while 'reproducibility' was having someone else be able to take the same inputs and produce the same outputs. This evidence suggested the need for further exploration and classification of the dimensions of reproducibility in order to effectively measure changes in reproducibility practices and the effects of those changes.

Introduction

The University of Washington (UW) eScience Institute was established in 2008 to help ensure that UW was a leader in both advancing the techniques and technologies of data-intensive discovery and in making them accessible to researchers in the broadest imaginable range of fields.

In 2013, in partnership with the Center for Statistics in the Social Sciences, the eScience Institute secured a 5-year, $37.8M grant from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation – jointly with NYU and UC Berkeley – to conduct a distributed, collaborative experiment in creating what the foundations refer to as "Data Science Environments"—conditions in which data-intensive discovery would truly thrive. The project launched at the end of 2013.

This report summarizes the baseline evaluation findings from the evaluation of the University of Washington Data Science Environment (UW DSE). These findings emerged from analysis of survey data collected in March, 2015 and key informant interviews conducted February through March, 2015. The findings and conclusions in this final report are intended for use by UW DSE leaders to inform strategic and operational decision making and action taking to improve UW DSE quality, implementation effectiveness, and desired outcomes.

The UW DSE baseline evaluation plan, developed in partnership with the UW DSE Ethnography and Evaluation Working Group, (Appendix B) focused on UW DSE leaders' evaluation questions about the following processes and outcomes:

1. **Reproducibility and Open Science**
    a) **Increase awareness of need for tools, code, and data sharing/publishing in research labs/undergraduate education**
2. Software Tools, Environments, and Support
    a) Increase technique and technology adoption by scientific community
    b) Increase adoption of data-intensive discovery
3. Education and Training
    a) Increase number of: user-friendly tools, data science courses and curricula, pi-shaped educators
    b) Increase cross-department events and training
4. Career Paths and Alternative Metrics
    a) Increase awareness of need for additional career impact measures
    b) Increase recruiting of data scientists
5. Working Space and Culture
    a) Increase number of data scientists and domain scientists working together on projects
    b) Increase inter-domain collaborations

The evaluation implementation included data collection using the following methods. See Appendix C for detailed data collection and analysis methodologies.

Methodological overview

1. Interviews

In early January 2015, data2insight worked with UW DSE working group leads to identify a purposeful sample of 26 key informants. UW DSE working group leads identified key informants to participate in interviews based on their knowledge and proximity to the UW DSE or their potential to offer a valuable outsider's perspective. This purposeful sample included senior and junior faculty, department chairs, eScience Institute data scientists, post docs, and graduate students. Data2insight evaluators interviewed 21 individuals (81% response rate).

2. Surveys

In February 2015, eScience Institute program management administered the UW DSE survey using Catalyst (UW web survey administration platform). The eScience Institute program management administered the UW DSE survey from March 2 – April 27, 2015 to one purposeful sample (N=58) and three random samples (N=1853).

The survey was administered via Catalyst in five parts:

1. Software Tools, Environments, and Support
2. Reproducibility and Open Science
3. Education and Training
4. Working Space and Culture
5. Career Paths and Alternative Metrics

Purposeful sample

Data2insight worked with the UW DSE Ethnography and Evaluation working group and created the purposeful survey sample (N=58) from members of the UW DSE data science community. These individuals received all five survey sections.

Random samples

Data2insight worked with the UW Library Data Services Coordinator and the UW DSE Ethnography and Evaluation working group to identify the random survey sample (N=1853).

The random sample originally identified by the UW Librarian included all non-emeritus faculty (N=1,077), graduate students (N=1,497), and research scientists/engineers (N=387) from the following departments:

- Oceanography
- CSE
- Astronomy
- Physics
- Biology
- Sociology
- Applied Math
- Biostatistics
- Genome Sciences
- Statistics
- iSchool
- Human Centered Design and Engineering

The random survey samples consisted of 1,853 of these individuals. The first random survey sample received survey prompts from three of the five working group survey sections: Education and Training, Software Tools, Environments and Support, and Reproducibility and Open Science. The second random survey sample received survey prompts from two of the five working group survey sections: Career Paths and Alternative Metrics and Working Space and Culture. The third random survey sample received survey prompts from one of the five working group survey sections: Career Paths and Alternative Metrics. (This administration was necessary to achieve the desired sample size of N=150 for this survey section. In the second random sample miscoded skip logic resulted in too few survey respondents.

1. What does ROS mean to you and your peers? What do faculty, students, and scientists value about ROS?

1.1. Reproducibility

Survey respondents overwhelmingly agreed that reproducibility supports greater scientific validity and integrity (97%) and that it is a basic principle of the scientific method (95%). While most (64%) did agree or strongly agree that it can contribute to rapid discovery, 15% disagreed, and 21% didn't know.

More than 80% of survey respondents agreed or strongly agreed with the following statements about reproducibility:

- Reproducibility supports greater scientific validity and integrity (97%)
- Reproducible research is a research result that can be replicated by another investigator (95%)
- Reproducibility is a basic principle of the scientific method (88%)
- Reproducibility is the ability of an independent researcher to recreate an entire experiment using the same procedures (86%)

The highest level of disagreement (15%) regarding reproducibility was with the statement: "Reproducibility can contribute to rapid discovery." However, 64% of respondents agreed or strongly agreed with this statement.

Table 1 lists respondents' level of agreement with each reproducibility statement provided in the survey.

Table 1: **Survey respondents' level of agreement with the given statements about reproducibility**

Statement	N	Don't know	Strongly disagree	Disagree	Agree	Strongly agree
Reproducibility supports greater scientific validity and integrity	208	2%	0	0	30%	67%
Reproducibility is the ability of an independent researcher to recreate an entire experiment using the same procedures	209	2%	0	3%	36%	59%
Reproducibility is a basic principle of the scientific method	209	5%	0	7%	34%	54%
Reproducibility can increase the productivity of current and future researchers	209	14%	0	6%	38%	41%
Reproducibility is the ability of an independent researcher to recreate an entire experiment using the same procedures	208	4%	0	10%	45%	40%
Reproducibility can contribute to rapid discovery	206	21%	0	15%	46%	18%

Interviewees (N=21) all agreed that reproducibility meant the ability to repeat an experiment using the same inputs and achieving the same result.

While some interviewees identified reproducibility as applying to themselves, that reproducible research was when they or their own lab could reproduce their work, most interviewees identified reproducible research as research where someone else could apply your methodology to your data and reproduce the same results. Two interviewees identified this distinction as the difference between "repeatability" or "replicability" and "reproducibility;" these interviewees stated that the former referred to the experimenters' ability to repeat the experiment, while the latter was having someone else be able to take the same inputs and produce the same outputs.

Interviewees felt that their peers and others in their departments, despite some potential confusion on the distinction of the terms "repeatability/replicability" and "reproducibility," held the same general understanding of reproducibility. The following comments are representative of key informant interviews:

"The standard definition of [reproducibility] is that you can perform—given the input data that someone produced in an experiment, you can recreate the analysis that they performed and get the same results from that analysis."

– Staff

"[Reproducibility means] enough information in the methods that someone could reproduce exactly what you did given the same resources."

– Post Doc/Grad Student

"[Reproducibility means] that all the data that are generated should be able to be replicated by someone who is an expert in the field; somebody who is an expert in the field should be able to do what you did."

– Department Chair

All interviewees clearly stated that the value of reproducibility was that it validated scientific research; reproducible research ensured that the research was executed and reported correctly. Even interviewees who felt that few individuals in their field practiced reproducible research, still highly valued it. One data scientist stated, *"If it's not reproducible, it's not science."* A department chair echoed this sentiment, *"It isn't science if it isn't reproducible."*

Interviewees strongly felt that reproducibility *"provides a better foundation from which new science can happen."* Interviewees valued that reproducible research practices provided an audit trail that allows others to review your work and ensure your work was done accurately. Interviewees also valued that reproducible research "*removes a little bit of the subjectivity"* from analyses. For example, two different department chairs made the following statements:

"If we're going to make predictions, which is all science is really good for, then those predictions have to be robust, and I think you can't know that unless they're reproducible."

"Reproducibility and independent ways of coming to similar conclusions have been extremely important in forming some of the most surprising and interesting discoveries in astrophysics."

A handful of social science interviewees from Anthropology, Sociology, and HCDE stated that reproducibility *"is not emphasized very much"* in their fields. These interviewees felt that though

reproducibility of research and experiments remained very important, *"empirical reproducibility should be classified as something different"* given the nature of field data collection. These interviewees felt that with field research, reproducibility is best defined more narrowly, as *"the ability to take the code and the data and then combine them to obtain the figures and tables and other results that appear in publication… it's computational, statistical, reproducibility…in field based disciplines, you've got field data collection… and you could never hope to go back and collect that exact data set again."*

Other interviewees (N=4) articulated that reproducibility was a spectrum both in definition and in practice. They described the spectrum of current reproducible practices as ranging from preserving every detail of your research process—including computational platform information—to taking no measures to ensure your research could be reproducible by someone else. A data scientist explains how different aspects of reproducibility can be confusing:

"I think sometimes there's some confusion… I think there's less thought around what is the starting point, you know, when you say you are going to reproduce something… sometimes people subdivide it in terms of reproducibility in total versus reproducibility of just the computations that were done to produce-- to yield the result. So at that point, you're starting with known data, what you don't know is necessarily… what was the computational environment, which operating system, which libraries…"

The **spectrum of current reproducibility practices** is discussed in section 2.1.

1.2. Defining open science

Over 80% of survey respondents (N=209) agreed that standard open science protocols and practices of data and code sharing should be developed. Eight out of ten also agreed that open science could increase the impact of research and software development. While open science was perceived to be valuable, most survey respondents (70%) agreed that not all data should be shared with all people. In contrast, 24% of respondents disagreed that not all data needed to be shared with all people.

More than 80% of survey respondents agreed or strongly agreed with the following statements about open science:

- Standard open science protocols for data and code sharing should be developed (84%)
- Sharing data used in published research should be a default practice (84%)
- Open science can increase the impact of research performed (80%)
- Open science can increase the impact of software developed (80%)

Table 2 lists respondents' level of agreement with each open science statement provided in the survey.

Table 2: **Survey respondents' level of agreement with the given statements about open science**

Statement	N	Don't know	Strongly disagree	Disagree	Agree	Strongly agree
Open science can increase the productivity of current and future researchers	208	20%	0	4%	37%	39%
Sharing data used in published research should be a default practice	207	8%	0	8%	46%	38%
Standard open science protocols for data and code sharing should be developed	208	12%	0	4%	48%	37%
Open science can increase the impact of software developed	207	18%	0	1%	44%	35%
Open science can increase the impact of research performed	209	17%	0	2%	46%	34%
Open science is the movement to make scientific research and data accessible to all	209	17%	0	4%	46%	33%
Open science can increase the impact of papers published	207	23%	0	4%	42%	30%
Code developed as part of research should be made available with every publication and/or presentation	209	16%	0	19%	40%	25%
Not all data need to be shared with all people	209	7%	4%	20%	52%	18%

All interviewees (N=21) agreed on these main concepts of open science: open science refers to sharing data, code, and software used in research with all people, everywhere in the world, free of charge with permission for reuse. For example:

"[Open science means] that the data are available to anyone, for free, any time."

– Department Chair

"When I think of "open science," I think of science that's not just scientists accessing that data but it's anyone; people from developing countries, people that aren't scientists… open science is… It would just be the ability to access."

– Faculty

Additionally, most interviewees stated that open science meant free access to all journal publications for all people. Interviewees felt particularly strongly about practicing open science and freely sharing publications and data in the case of publicly funded research. One department chair's comment illustrated this opinion, *"Where public resources are substantial investments… I think the data, the products, and software interfaces, that expertise, the archive, things like that, some of the hardware access that you need to support all of that, should be publicly accessible and useful."*

Approximately one-third of interviewees were unfamiliar with the term "open science" and stated that it was not something they talked about with their peers or in their work. The interviewees unfamiliar with the term came from a variety of departments including social, physical, and life sciences.

Despite initially stating that they were unfamiliar with the term, these interviewees described open science similarly to those who were familiar with the concept and the term.

These individuals described their sense of open science as *"making data available to other people," "publishing early versions of your paper on repositories, making complete data sets available,"* giving datasets to others to use, *"free licensing, the ability to get data code,"* and full open access to study results with the permission to use the data.

However, few interviewees felt that all of these concepts of open science needed to be practiced in all research.

1.3. Interviewees believed open science made scientific research more efficient and credible.

Most interviewees valued open science because (a) it allowed scientists and researchers to build off one another, reducing duplicated efforts and increasing efficiency and speed of discovery, and (b) it allowed scientists, researchers, and the general public to identify errors and increase the speed and accuracy of scientific practice. Interviewees also valued open science because much of their own research relied on open or publicly shared data; interviewees valued making research and data funded with public dollars publicly available. The following quotes are representative of interviewee comments.

"You want as much new science to happen as efficiently as possible, and the greater ability people have to access what you've done, the more likely they're not going to have to repeat what you did and instead just build from that. So [open science] provides a better foundation from which new science can be developed."

– Post Doc/Graduate Student

"I think that the faster that you can get access to all of the particulars of a measurement the faster science can correct itself."

– Data Scientist

"[Open science] is a core tenet of science… I mean, that's the scientific method."

– Staff

"In some ways, [open science] is more work, but on the other hand, it's much more cost-effective. Why generate the same data over and over and over again?"

– Department Chair

"[Open science] is the ability to show that I'm right or to have others show that I might not be right, I guess to test the fact that I'm right. And then the ability to have the result reused and someone else could analyze the data in a new way or use the code to analyze their own data in a single way. So it's reuse, so it's research integrity and research reuse."

– Faculty

1.4. There were differing opinions about the practice of open science among survey respondents and interviewees.

Survey respondents disagreed most often with these statements:

- *Not all data needed to be shared with all people.* While 24% disagreed with the statement, 70% agreed or strongly agreed with the statement.
- *Code developed as part of research should be made available with every publication and/or presentation.* Nineteen percent of respondents disagreed with this statement. In contrast, 65% agreed or strongly agreed with the statement.

There was also disagreement across interviewees regarding the extent to which the principles of open science were feasible or should be followed. Over half of interviewees (N=12) articulated that the practice of open science fell on a spectrum. These individuals felt that not all open science practices were feasible or possible in many types of research. Most interviewees who felt that open science should be practiced on a spectrum, felt that sharing data and research should depend on the research field, the content of the data, and any potential human risks data sharing could pose.

Only a handful of interviewees (N=2) felt strongly that open science should be a strictly followed practice (i.e. “all data should be shared with all people”). This difference in opinion mirrored the disagreement seen in the UW DSE survey, where 24% of survey respondents disagreed with the statement “not all data should be shared with all people.”

2. What are the current practices and barriers to effective ROS code and data sharing?

2.1. Current ROS practices

Most survey respondents reported using scripts, version control systems, and open source code repositories for code and data sharing. However, less than 10% reported using workflow management systems.

Survey respondents **sometimes** or **always** used the following software tools, environments, and data management practices:

1. Scripts to ensure replicable computations (70%)
2. A version control system (57%)
3. Open source code repositories (56%)
4. Notebook environments to capture sequence of instructions (47%)
5. Repositories to deposit data used in publications (42%)

The least frequently used tools and practices were workflow management systems (9%), literate programming (18%), and auto tools (19%). Table 3 provides a complete list of survey prompts and response rates.

Table 3: **Survey respondents' frequency of use for listed tools, environments, and data management practices**

ROS Practice	N	Never	Sometimes	Always
Scripts to ensure replicable computations	198	30%	32%	38%
A version control system	202	43%	32%	25%
Notebook environments to capture sequence of instructions	204	53%	33%	14%
Repositories to deposit data used in publications	204	58%	31%	11%
Open source code repository	206	44%	45%	11%
Internal code reviews to ensure reproducible results	205	60%	33%	7%
Licensing	203	63%	31%	6%
Literate programming	203	82%	15%	4%
Automated tools to document the computational environment	205	81%	16%	3%
Virtualization to archive the environment in which a code runs	206	74%	24%	2%
Workflow management systems to track series of experiments	205	91%	9%	0

2.2. Over two-thirds of the interviewees were familiar with ROS guidelines.

The UW DSE ROS working group had developed and shared a set of guidelines for reproducible and open science practices. These guidelines included suggestions for practicing version control, replicable computations, data and code provenance, sharing and archiving, and replicable environments. Eleven interviewees were familiar with and had seen the ROS guidelines; five interviewees had not[1].

Similar to the ROS practices among UW DSE survey respondents, version control, using scripts for replicable computations, and replicable environments were the most common ROS practices among interviewees. GitHub was the most common version control tool, with approximately one-third of interviewees from the Information School, Astronomy, Physiology and Biophysics, Physics, and Biology, stating that they used GitHub. Some interviewees also described using GitHub as a tool with their students to share and preserve code, asking students to comment on all of the code shared. The following quotes were representative of interviewee comments regarding current ROS practices.

"Everything [on the reproducibility and open science guidelines], for example, is by necessity a part of many of the things I'm directly involved with, like the Sloan Digital Sky Survey NASA project, and certainly a part of future planning the department has for the Large Synoptic Survey Telescope and related projects. I think those are places where-- I can't claim that this is uniformly true across all of astronomy, but I think in those places, which are places where I've actively worked recently, I think each and every one of those things is something that is routinely engaged in and, indeed, required."

– Department Chair

"Yes, these [reproducibility and open science guidelines] are all things that I do every day."

– Faculty

"We have all of our code on GitHub; it's freely available, anybody can download it. I think it's very important that the data become public and that the codes are public so that you can check why you have a different answer than somebody else… I think it should be public and visible to people, but I think doing it separately is important."

– Data Scientist

"Everything I do is under version control. Replicable computations, yeah, pretty much everything I do is script based. Data and code sharing, again, most anything that's of reasonable size goes on to GitHub... Replicable environment… I release my AWS images but I don't do all of my sort of local computations in a virtual machine just because of overhead issues."

– Post Doc/Grad Student

"I tend to try and adopt practices that lean in that direction, but yeah. I don't know. For example, we don't share data sets very often. Most of the data sets that I work with can't be shared. But I put all of my software on GitHub so that it's open source… I mention that in the papers that I write about it… I try and make those available to the public. I also contribute to open source projects that I use when I'm doing my research…"

– Post Doc/Grad Student

[1] Only interviewees interviewed during data collection (N=16) were specifically asked this question; we did not ask this question of individuals interviewed in the preliminary interviews (N=5). A copy of the guidelines can be found at the UW DSE ROS woking group webpage (http://uwescience.github.io/reproducible/guidelines.html).

"In my lab we've improved our versioning habits. I've always had, you know, like being able to use Git, or GitHub, or Bitbucket, or versioning; versioning code, versioning papers better, that's improving. Also versioning data to some degree, but not really yet, we're still working really hard on that and I don't think we're going to solve that anytime soon. But in our lab, my lab specifically, versioning. We've gotten much better at versioning, so that's probably the big improvement we've made."

– Faculty

"Many journals charge extra to publish in open-access format. When I have the opportunity to pay extra to make it open-access, I never do it. My feeling is that-- I mean, as soon as the thing is published, I put it on my website, which journals don't really like that, but they also can't really stop it… I think, in terms of raw data, it's immeasurably better today than it was 10 years ago, because there're repositories for raw data and for supplementary data and things like that, that were never available before, except by making a request to the author, which was a huge nuisance to make the request, and an even bigger nuisance to fulfill the request. And now that's just not an issue. And so the UW has its own repository for those kinds of data, because the journals require that those be permanent repositories. I think that's working really well, and I'd love to be able to point-- people say, "Hey, could you send me the raw data of this?" And I said, "Hey, guess what-- it's already online."

– Department Chair

Sharing and archiving all code used in published research was also a common practice among interviewees. Many interviewees discussed sharing and archiving their data as well, though sharing data was a significant challenge for many individuals who worked with either human subjects or proprietary data.

While only one interviewee stated specifically using a notebook (i.e. iPython notebooks) to capture and document their processes, several interviewees discussed having their own informal processes in place to track their work flows, in addition to requiring that their students keep notes or slides with intermediate results. One interviewee also described using literate programming (e.g., Sweave, Knitter) to create dynamic reports and automatically update data in papers.

2.3. ROS barriers

The most common barriers to code and data sharing identified by survey respondents were sharing restrictions from collaborators, lack of existing data sets, and concerns about having ideas or data stolen. The least common barrier was concerns about lack of reproducibility if the data were shared. Another barrier to code and data sharing offered by respondents was lack of training.

Survey respondents' top five barriers to code and data sharing (sometimes or always experienced):

1. Sharing restrictions from collaborators (66%)
2. Lack of access to existing data sets (64%)
3. Concerns about having ideas stolen (64%)
4. Concerns about having data stolen (60%)
5. Sharing restrictions from suppliers of data (58%)

The most uncommon barriers to code and data sharing among survey respondents were concerns about data and/or code not being reproducible if shared (33%), financial incentives to keep data/code private (34%), and university policies (41%).

Interviewees' top five barriers to code and data sharing:

1. The time it took to make code, data, and research open and reproducible (N=11, 52%)
2. Fear of being scooped or having data stolen (N=10, 48%)
3. Insufficient sharing platforms (N=9, 43%)
4. The monetary cost of sharing (N=9, 43%)
5. The inability to share data and/or code due to human subject or biological physical data, and Institutional Review Board requirements (N=9, 43%)

Figure 1: **Survey respondents reported frequency of experience with sharing restrictions from collaborators as a barrier to code and data sharing, grouped by sample type and respondents' self-selected professional identity**

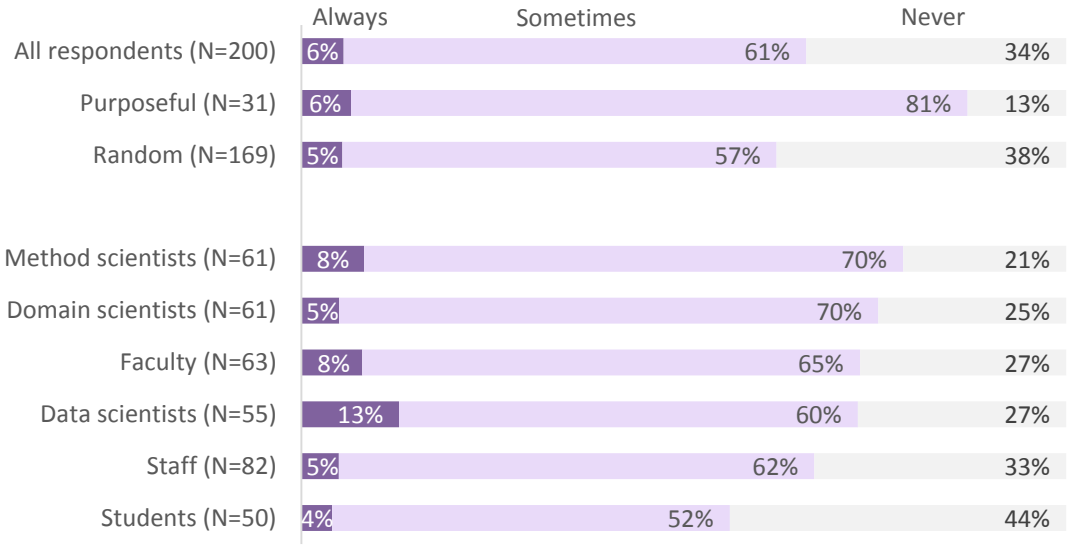


Figure 2: **Survey respondents reported frequency of experience with a lack of access to existing data sets as a barrier to code and data sharing, grouped by respondents' self-selected professional identity**

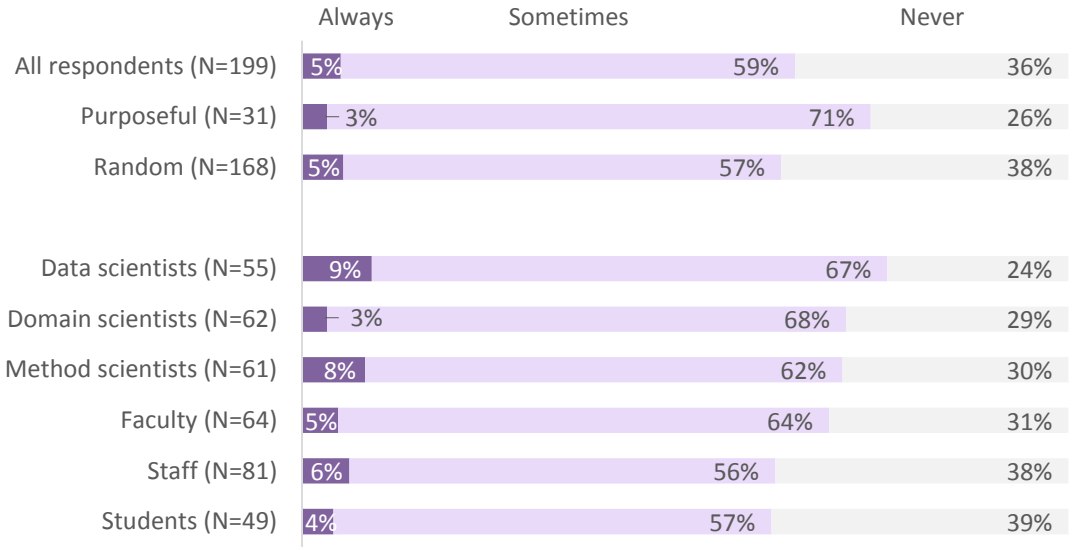


Figure 3: **Survey respondents reported frequency of experience having concerns about ideas being stolen as a barrier to code and data sharing, grouped by sample type and respondents' self-selected professional identity**

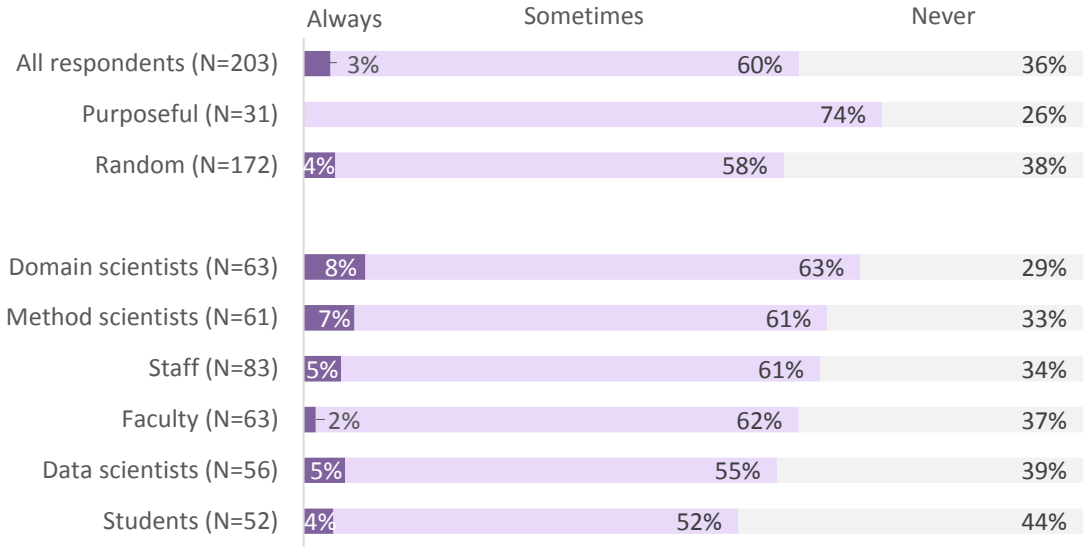


Figure 4: **Survey respondents reported frequency of experience having concerns about data being stolen as a barrier to code and data sharing, grouped by sample type and respondents' self-selected professional identity**

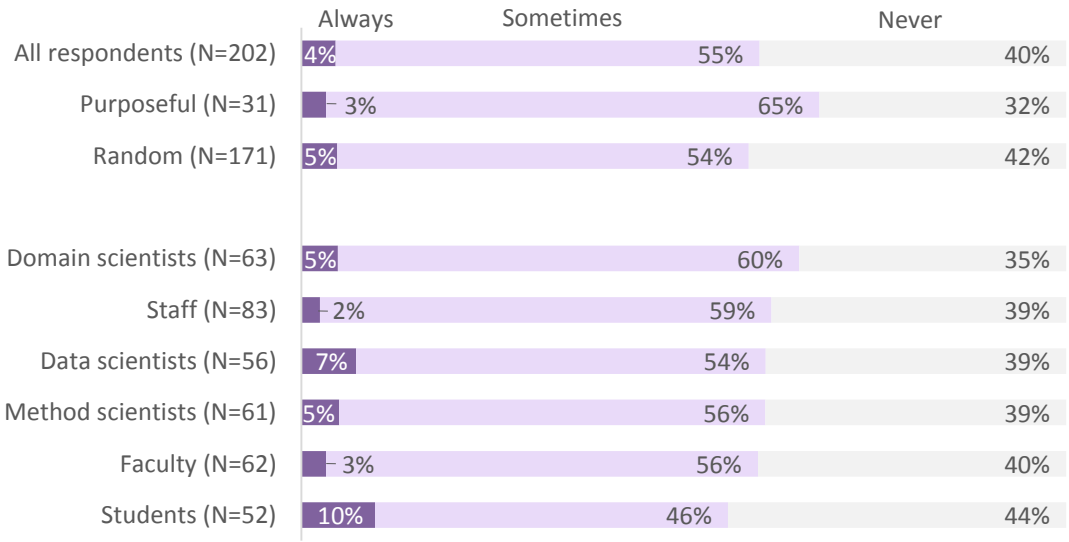


Figure 5: **Survey respondents reported frequency of experiencing sharing restrictions from suppliers of data as a barrier to code and data sharing, grouped by sample type and respondents' self-selected professional identity**

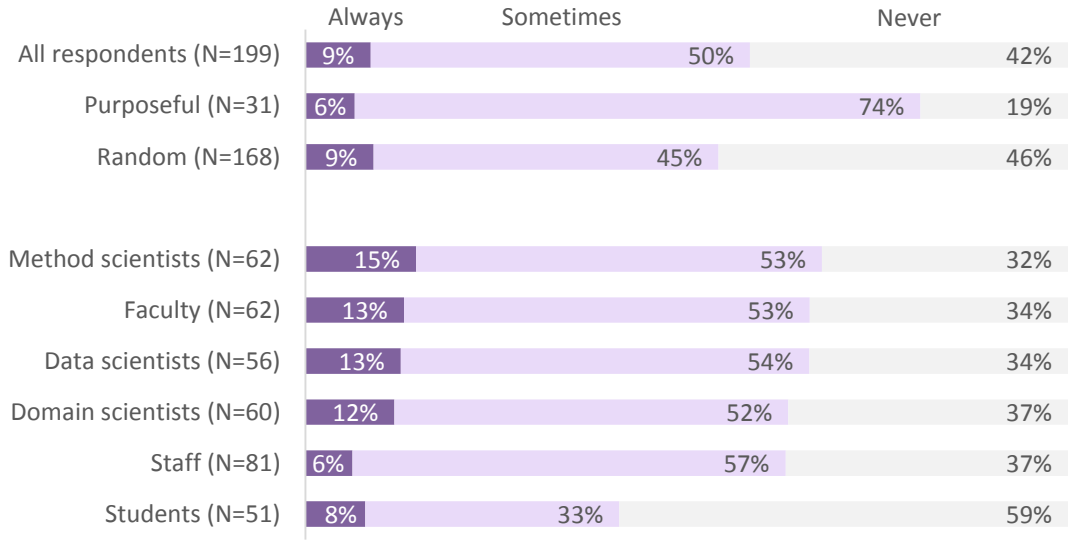


Table 4: **Survey respondents wrote-in the following code and data sharing barriers that they frequently face.**

Other barriers to code and data sharing	Number of respondent mentions
Lack of training	10
Insufficient time	7
Insufficient funding	6
Gap in understanding of qualitative vs. quantitative research and a lack of respect for other disciplines	5
Privacy issues (Human Subjects Board and IRB)	5
Lack of incentives	4
Storage space	2
Work required	2
Insufficient caution about drawing causal conclusions from big data	1
Lack of interest	1
Lack of personnel needed to share code and data	1

Interviewees reported experiencing an average of four barriers to code and data sharing with the frequency of barriers experienced ranging from one to seven . Interviewees in the library sciences, information school, HCDE, sociology, anthropology, communications, and statistics departments reported having experienced an average of five ROS barriers compared to individuals from biology, computer science and engineering, physics, oceanography, physiology and biophysics, electrical engineering, astronomy, and applied math—who experienced an average of three barriers per person. Junior faculty interviewees were the most likely to experience barriers to code and data sharing with an

average of 5.25 barriers per person reported. The following quotes were illustrative of interviewee comments.

"A lot of those [ROS practices] take a long time, and I feel like for my goals as an academic I don't think right now that those efforts are rewarded… it's not one of the metrics to which I am held"

– Faculty

"The principal [barrier] is the desire among investigators to be the first to use their data and so the unwillingness to share. And then the second barrier is once the data [is] released the lack of support for that data… Somehow there should be credit given for good sharing…. sometimes programmers are given bonuses for their code that's used by others. And if somehow there was [something] like citation credits for papers. If there was credit given for data…. somehow that has to figure in there because right now mandating [open science] doesn't work. There has to be an incentive system."

– Department Chair

"People feel sometimes a lot more reluctant to practice [open science] if they don't feel that they're vested in some way in recognizing the benefit of all the time that they put into those best practices. So ways that can make people feel that they are getting something out of that… I think that that sometimes can be a barrier, and a tendency to try to hold onto those things as proprietary."

– Department Chair

"I think it's perfectly okay, and it's a standard in this field to have a period of time in which your data is not publicly available. There's a couple reasons for that. One is simply to verify that the data is good quality and that you understand it. Putting out data that you don't understand or that might have weird problems in it that you haven't figured out and you can't tell people about tends to make you look bad. It doesn't help anything and nobody can get anything informative out of it, so there's no value… I think having a little bit of time to let the people who have put the effort-- who spent years of their life building the instrument get some science out of it before everybody else gets a crack at it is fair."

– Data Scientist

"This subject of privacy requirements that 'sharing should be the default'-- I think this is extremely discipline-specific, and I think there are many disciplines where that will just sound naïve and a turn-off… there's no way that somebody analyzing tax records, which we have in the census data center, can be expected to be making the data available… there are situations with that kind of data whereby you might be allowed to publish some summary, but then if you publish some summary that could then prevent somebody else from publishing another summary, because if those two summaries were put together somebody could then identify individuals, and so the census office restricts which things are being put forward. So somebody thinking that they were living up to an ideal of making all the data available on one particular analysis they did… could then block some other researcher from publishing another summary of the data… So there are lots of thorny issues there, and they are being thought about in a lot of detail."

– Department Chair

"I think saying 'Sharing should be the default' is…. if it said sort of 'wherever possible' I think it would be reasonable… while I certainly am in favor of data being put forward in this way, people should be aware of the ethical issues that are arising here, and the track record is not always very good."

– Department Chair

"In my field it's all human produced data. That's the problem is that I'm a social scientist. I study people, and there are either human subject issues or licensing issues."

– Faculty

Additional barriers named by interviewees included:

- Lack of knowledge about how to share data or how to make data and code reproducible (N=7, 33%)

  "People don't understand just how much they have to describe their data along the way in order to make it usable and sharable into the future. And it's hard-- it's not easy-- it's complicated, it takes time-- and people just want to analyze the data. They don't want to sit and describe it for somebody else."

  – Staff

  "You know, I think that there's a lot of older faculty and less computationally savvy graduate students and undergrads that have never even heard of version control. And then there's people in the data science group that adhere to all of these in every project they do and that's fantastic. So yeah, huge variability."

  – Post Doc/Grad Student

- Lack of incentive for sharing code, data, and research (N=6, 29%)
- Insufficient access to data (N=6, 29%)
- Inability to share data due to its proprietary nature (N=6, 29%)
- Perfectionism and a fear of sharing data or code that is not perfect or of high enough quality (N=5, 24%)
- Restrictions imposed by collaborators (N=1, 5%)

  "It turns out the thing that slows me down are data agreements between UW and the industry… that slows me down a lot. Because the industry are scared to share their data, and they have all the interesting data…"

  – Faculty

- Lack of standards for how to share code and data or how to produce reproducible research (N=1, 5%)

Insufficient sharing platforms and the **monetary cost** of sharing data were the two barriers that most often occurred as a pair; the **time required** to make code and data shared and/or reproducible and **IRB and Human Subjects data** were the second most frequent pairing of barriers.

Social scientist interviewees were the most vocal about insufficient incentives to encourage reproducibility in their fields. For example:

"I would argue, that some of social science is… a lot of published social sciences, perhaps depends on one or two cases in data. Generalized notion of would someone find this same thing if approaching it by a different method, say a statistical method or… You know, I think there's a lot of, at least in my own world, a little skepticism that what you see in a given paper comes out of a lot of playing around before you get there… And how sensitive the results are to that playing is never really disclosed or fully appreciated… there's a lot of users out there that are trained to be able to do something in a very cookbook-ish fashion. And so there's not always a real understanding of what they did."

– Department Chair

2.4. Grassroots inclusive outreach and education that meets researchers where they are on the ROS spectrum is key to improving ROS on the UW campus.

Interviewees agreed by and large that ROS practice on the UW campus needed to be viewed as a spectrum. Interviewees expressed that not all individuals were going to be able to implement 100% of ROS guidelines. Thus, there was a desire for ROS leaders to recognize this spectrum, meet individuals and teams wherever they are on that spectrum, and support their incorporation of ROS into their work in feasible ways.

"For me, [reproducibility] is not like a single thing. I don't think there's a badge that says this work is or is not reproducible. I think that it can be more or less reproducible along a lot of dimensions."

– Department Chair

The primary suggestion interviewees made for how the UW or the ROS working group could support increased reproducibility and open science in research was to increase engagement, outreach, and education activities across the UW campus. Interviewees suggested incoming graduate students and junior faculty as ideal targets for these efforts. Specifically, interviewees identified providing tutorials on how to use GitHub, explaining what it means to do versioning and how it can be helpful in one's research, and giving lessons on its use. For example, one faculty member stated that *"I think continuing to encourage/try and disseminate the idea of source code repositories, version control repositories for both analytic tools chains and papers is great. I think that's probably the lowest hanging fruit. So i.e. put everything on GitHub."*

Interviewees also discussed the importance of being accepting, inclusive, and non-judgmental while doing outreach and education. They suggested crafting guidelines that take into account the differences across departments and fields and welcoming individuals to begin by incorporating one ROS practice into their research at a time. The following quotes illustrate that how ROS leaders engage faculty, administration, staff, and students is important to effective engagement and ultimately changing the UW culture.

"I'm on a mailing list, and I get all of these emails; I feel like a lot of it is very preachy and judgmental, and I think that's counterproductive, because it makes other people not want to listen to the reproducibility group. It's a lot of like "thou shalt do this," right, and if thou shalt not do that it's like, okay, now you're a terrible person, right?… the recommendations are not flexible… philosophically I'm for it, but I think the recommendations that they have right now are not easy to follow, and it's easier to ignore all of it rather than to take some of it."

– Department Chair

"I think that there's a distinction to be made between limits versus reasonable expectations. And I think that sometimes there is the expectation that everybody engage in open science. And that's good in theory and that comes across as like oh, why would you object to that? What possible objection could you have? And I do see cases where people have legitimate objections. I know of cases where there [are] labs that are relatively small labs of two or three people. And they have a software model that they've built up over years. And if they release that model they sort of lose the advantage that they've accrued through the development of this technique. And that might be something that's absorbed by a lab that's three times as large and is funded better. The counter argument to that is well how can you trust it if you can't look at it? And that's a fair counter argument. And so there's certainly, I think, very much two sides to this and I think sometimes the side of sort of the small lab trying to protect its time investment isn't really viewed as valid, and I think there's an element to that that should be considered."

– Faculty

Conclusion

- *Reproducibility and Open Science*: **While interviewees and survey respondents largely agreed on the definitions of and benefits to ROS, more classification is needed. The fear of being scooped, concerns of having ideas stolen, and data sharing restrictions were the biggest barriers to code and data sharing.**

The evaluation findings suggest that the UW could benefit from further definition and classification of "reproducibility" across multiple dimensions and to clarify the difference between "repeatability" and "replicability." Additionally, the evidence revealed a need to facilitate consensus within the UW community regarding the practice of open science and how to determine when unrestricted sharing is appropriate and when it is not. Finally, interview data suggests that it is not only important to classify definitions and define best practices, but also to facilitate ROS uptake in ways that meet researchers where they were on the ROS spectrum. Interviewees indicated that graduate students and junior faculty were especially important stakeholder groups to reach out to, educate, and train to be ROS experts and evangelists going forward.

Preliminary data showed that those who have used the Data Science Studio and participated in eScience Institute educational activities found the eScience Institute culture to be welcoming and egalitarian. A welcoming and egalitarian approach will serve the ROS working group well as the group engages with researchers at the UW going forward.

ROS Definitions

All interviewees described "**reproducibility**" as the ability to repeat an experiment using the same inputs and achieve the same result. Nearly 100% of survey respondents agreed with the statement, "Reproducibility supports greater scientific validity and integrity."

Interviewees generally described "**open science**" as sharing software, data sets, code, and publications with all people, free of charge. More than 80% of survey respondents agreed or strongly agreed with the following statements about open science:

- Standard open science protocols for data and code sharing should be developed (84%)
- Sharing data used in published research should be a default practice (84%)
- Open science could increase the impact of research performed (80%)
- Open science could increase the impact of software developed (80%)

ROS Practice

Not surprisingly, the practice of ROS was where opinions diverged. In both the survey and interviews we saw divergence in the group regarding open science practice:

Survey prompts:

- Not all data need to be shared with all people (24% disagreed; 70% agreed or strongly agreed)
- Code developed as part of research should be made available with every publication and/or presentation (19% disagreed; 65% agreed or strongly agreed)

There was also disagreement across interviewees regarding the extent to which the principles of open science were feasible or should be followed. Over half of interviewees articulated that the practice of open science was a spectrum. These individuals felt that not all open science practices were feasible or possible in many types of research. Most interviewees who felt that open science should be practiced on a spectrum thought sharing data and research should depend on the research field, the content of the data, and any potential human risks data sharing could pose.

At baseline (March 2015), over half of the UW faculty, staff, and graduate students were familiar with the ROS guidelines and/or using scripts to ensure replicable computations, version control systems, and open source code repositories. The most common barriers to code and data sharing were:

- Sharing restrictions from collaborators
- Lack of access to existing data sets
- Concerns about having ideas stolen
- Concerns about having data stolen
- Sharing restrictions from suppliers of data
- Time required to make code, data, and research open and reproducible
- Human subjects or biology-related restrictions
- Lack of knowledge about how to share data

DSE Working Group:

Reproducibility and Open Science

Baseline evaluation highlights

June 24, 2015

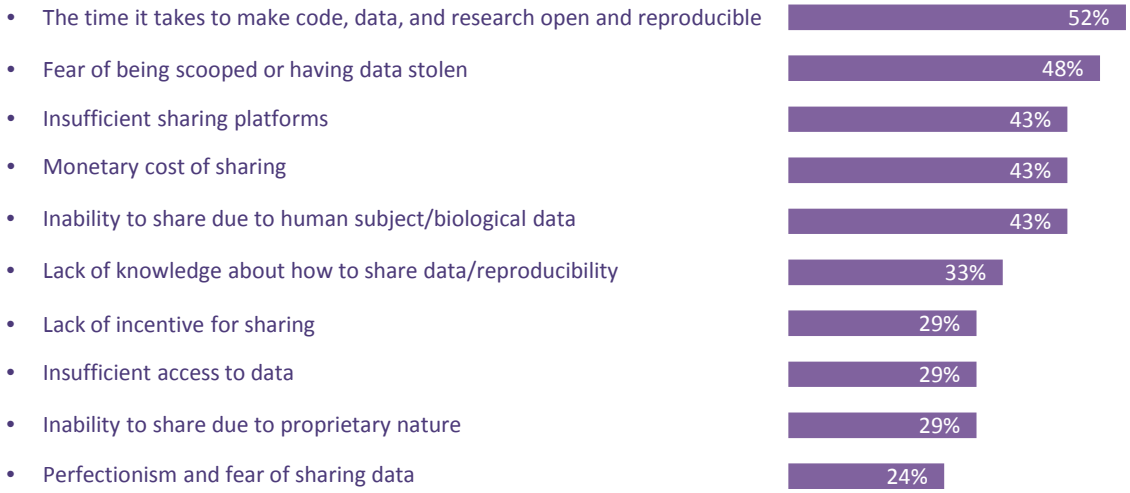
UNIVERSITY of WASHINGTON eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Reproducibility and Open Science (ROS) on the UW campus

- All interviewees described reproducibility as the ability to repeat an experiment using the same inputs and achieve the same result.
- Interviewees described open science as the open sharing of software, data sets, code, and publications with all people, free of charge.
- Most interviewees felt that sharing data and research depended on the field, the content, and potential human risks to sharing the data.
- 11 (69%) interviewees were familiar with the ROS guidelines.

Barriers to code and data sharing (N=21)



Opportunities for development

- Meet researchers where they are on the ROS spectrum

"For me, [reproducibility] is not like a single thing. I don't think there's a badge that says this work is or is not reproducible. I think that it can be more or less reproducible along a lot of dimensions."

- Department Chair

- Offer more outreach, education, and training on campus, especially to graduate students and junior faculty (e.g., Github training)

"I'm on a mailing list, and I get all of these emails; I feel like a lot of it is very preachy and judgmental, and I think that's counterproductive because … philosophically I'm for it, but I think the recommendations that they have right now are not easy to follow, and it's easier to ignore all of it rather than to take some of it."

- Department Chair

data2insight
Find wisdom in information

DSE Working Group:

Reproducibility and Open Science

Baseline evaluation highlights

June 24, 2015

UNIVERSITY of WASHINGTON eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Overview

- Respondents agreed that reproducibility supported greater scientific validity and integrity.
- Most agreed (64%) that reproducibility can contribute to rapid discovery; 15% disagreed.
- Respondents felt that standard open science protocols for data and code sharing was a good thing.
- 24% believed that all data should be shared with all people.
- Sharing restrictions from collaborators was the #1 barrier to code and data sharing.
- Scripts were the most common tool used for data management.

Respondent beliefs about **reproducibility** (N=209)

- Reproducibility supports greater scientific validity and integrity, 97%
- Reproducible research is a research result that can be replicated by another investigator, 95%
- Reproducibility is a basic principle of the scientific method, 88%
- Reproducibility can contribute to rapid discovery, 64%

Respondent beliefs about **open science** (N=209)

- Standard open science protocols for data and code sharing should be developed, 84%
- Sharing data used in published research should be a default practice, 84%
- Open science can increase the impact of research performed, 80%
- Open science can increase the impact of software developed, 80%

Software tools, environments and data management practices used (N=206)

- Scripts to ensure replicable computations, 70%
- A version control system, 57%
- Open source code repositories, 56%
- Notebook environments, 47%
- Repositories to deposit data used in publications, 42%

Barriers to code and data sharing (N=203)

- Sharing restrictions from collaborators, 66%
- Lack of access to existing data sets, 64%
- Concerns about having ideas stolen, 64%
- Concerns about having data stolen, 60%
- Sharing restrictions from suppliers of data, 58%

data2insight
Find wisdom in information

Survey Demographics:

Reproducibility and Open Science

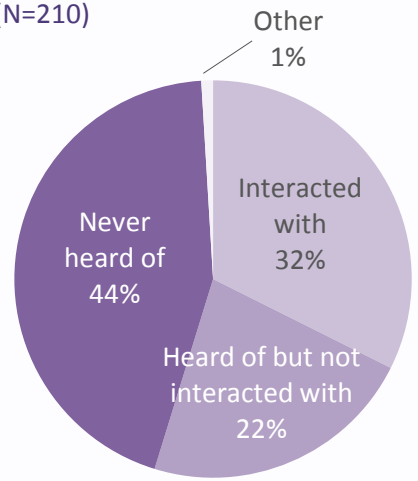
Baseline evaluation highlights

June 24, 2015

UNIVERSITY of WASHINGTON eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Familiarity with UW DSE (N=210)



Position at UW (N=210)

- Research scientist/engineer, 30%
- Graduate student, 25%
- Tenured faculty, 18%
- Postdoc, 10%
- Non-tenured faculty, 9%

Professional identity (N=210)

- Domain scientist, 30%
- Method scientist, 30%
- Data scientist, 27%
- Other, 25%
- Engineer, 15%

Primary department (N=210)

- Other, 17%
- Oceanography, 12%
- Genome Sciences, 11%
- iSchool, 11%
- Physics, 10%

Race (N=194)

- White / Caucasian, 85%
- Asian, 12%
- Hispanic / Latino, 5%
- Other, 2%
- American Indian / Alaska Native, 1%
- Native Hawaiian / Pacific Islander, 1%
- Black / African American, 0%

Age (N=177)

Mean age: 42 years
Range: 23 - 90

Gender identity (N=197)

- Male, 52%
- Female, 47%
- Self-identify, 1%

Education (N=203)

- Doctoral, 57%
- Bachelor's, 23%
- Master's, 19%
- Professional degree, 1%
- Alternative credential, 0%
- Associate's degree, 0%

Years in academia (N=203)

- 0-2 years, 19%
- 3-4 years, 11%
- 5-6 years, 10%
- 7-8 years, 7%
- 9-10 years, 5%
- Over 10 years, 48%

data2insight
Find wisdom in information

Survey Demographics:

Reproducibility and Open Science

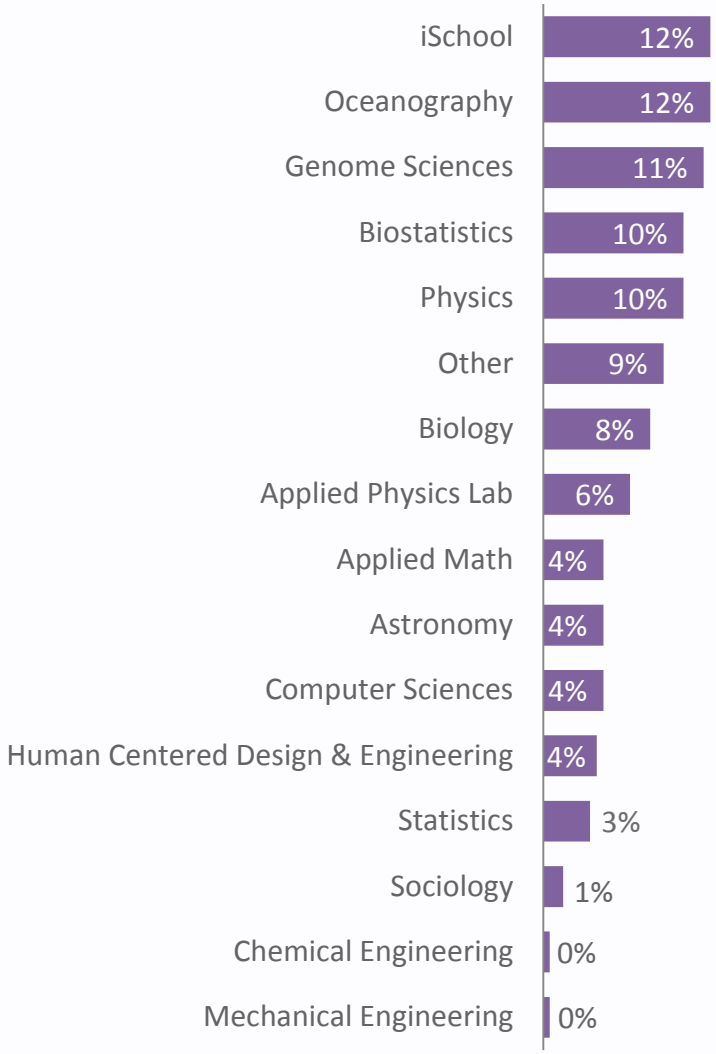
Baseline evaluation highlights

June 24, 2015

UNIVERSITY of WASHINGTON eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

All departments, including Applied Physics Lab (N=210)



data2insight
Find wisdom in information

Appendices

A: Demographics

B: Evaluation Plan

C: Methodology

D: IRB Human Subjects Worksheet

E: Theories of Action and Theories of Change

F: Preliminary Interview Protocol

Appendix A: Demographics

Demographics from all survey respondents (N=347)

Table 1: **All survey respondents: Level of familiarity with the UW DSE**

Familiarity with UW DSE	N (%)
I had not heard of the UW DSE prior to taking this survey	178 (51%)
I have heard of the UW DSE, but I have never interacted or worked with individuals involved in the UW DSE	80 (23%)
I sometimes interact and work with individuals involved in the UW DSE	50 (14%)
I often interact and work with individuals involved in the UW DSE	21 (6%)
I always interact and work with individuals involved in the UW DSE	14 (4%)
Other	4 (1%)

Table 2: **All survey respondents: Position at UW**

Position at UW	Percent of respondents (N=347)
Research scientist/engineer	35%
Graduate student	27%
Tenured faculty	15%
Postdoctoral student	10%
Non-tenured faculty	7%
Research faculty	5%
Other	5%
Data scientist	1%
Undergraduate student	0

Table 3: **All survey respondents: Professional identity**

Professional identity	Percent of respondents (N=347)
Domain scientist	36%
Method scientist	28%
Data scientist	30%
Engineer	20%
Other	20%

Table 4: **All survey respondents: Primary department**

Primary department	Percent of respondents (N=347)
Other	18%
Oceanography	12%
Genome Sciences	12%
iSchool	11%
Physics	9%
Biology	9%
Biostatistics	9%
Computer Sciences	6%
Human Centered Design & Engineering	4%
Astronomy	4%
Applied Math	4%
Statistics	4%
Sociology	3%
Mechanical Engineering	0
Chemical Engineering	0

Table 5: **All survey respondents: Gender identity**

Gender identity	Percent of respondents (N=329)
Male	55%
Female	44%
Self-identify	1%

Table 6: **All survey respondents: Race/ethnicity**

Race/ethnicity	Percent of respondents (N=322)
White/Caucasian	84%
Asian	13%
Hispanic/Latino	3%
Other	2%
American Indian/Alaska Native	1%
Native Hawaiian/Pacific Islander	1%
Black/African American	0

Table 7: **All survey respondents: Age in years**

Age in years	N	Minimum	Maximum	Mean
Age in years	296	22 years	90 years	41 years

Table 8: **All survey respondents: Academic credentials**

Academic credentials	Percent of respondents (N=334)
Doctoral	51%
Bachelor's	25%
Master's	23%
Professional degree	1%
Associate's	0
Alternative credential	0

Table 9: **All survey respondents: Years in academia**

Years in academia	Respondents N (%)
0-2 years	67 (20%)
3-4 years	43 (13%)
5-6 years	35 (11%)
7-8 years	24 (7%)
9-10 years	18 (5%)
Over 10 years	147 (44%)

Software Tools, Environments and Support, Education and Training, and Reproducibility and Open Science survey demographics (N=211)

Table 10: **Level of familiarity with the UW DSE among survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Familiarity with UW DSE	Respondents N (%)
I had not heard of the UW DSE prior to taking this survey	93 (44%)
I have heard of the UW DSE, but I have never interacted or worked with individuals involved in the UW DSE	47 (22%)
I sometimes interact and work with individuals involved in the UW DSE	37 (18%)
I often interact and work with individuals involved in the UW DSE	19 (9%)
I always interact and work with individuals involved in the UW DSE	12 (6%)
Other	2 (1%)

Table 11: **Position at UW among survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Position at UW	Percent of respondents (N=210)
Research scientist/engineer	30%
Graduate student	25%
Tenured faculty	18%
Postdoctoral student	10%
Non-tenured faculty	9%
Research faculty	5%
Other	4%
Data scientist	2%
Undergraduate student	0

Table 12: **Professional identify of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Professional identity	Percent of respondents (N=210)
Domain scientist	30%
Method scientist	30%
Data scientist	27%
Other	25%
Engineer	15%

Table 13: **Primary department of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Primary department	Percent of respondents (N=210)
Other	17%
Oceanography	12%
Genome Sciences	11%
iSchool	11%
Physics	10%
Biostatistics	10%
Biology	8%
Applied Math	5%
Statistics	5%
Computer Sciences	4%
Astronomy	4%
Human Centered Design & Engineering	4%
Sociology	1%
Mechanical Engineering	0
Chemical Engineering	0

Table 14: **Gender identity of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Gender identity	Percent of respondents (N=197)
Male	52%
Female	47%
Self-identify	1%

Table 15: **Race/ethnicity of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Race/ethnicity	Percent of respondents (N=194)
White/Caucasian	85%
Asian	12%
Hispanic/Latino	5%
Other	2%
American Indian/Alaska Native	1%
Native Hawaiian/Pacific Islander	1%
Black/African American	0

Table 16: **Age in years of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Age in years	N	Minimum	Maximum	Mean
Age in years	177	23 years	90 years	42 years

Table 17: **Academic credentials of survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Academic credentials	Percent of respondents (N=203)
Doctoral	57%
Bachelor's	22%
Master's	20%
Professional degree	1%
Associate's	0
Alternative credential	0

Table 18: **Years in academia among survey respondents who completed the Software Tools, Environments and Support survey, Education and Training survey, and the Reproducibility and Open Science survey section**

Years in academia	Respondents N (%)
0-2 years	38 (19%)
3-4 years	22 (11%)
5-6 years	21 (10%)
7-8 years	14 (7%)
9-10 years	11 (5%)
Over 10 years	97 (48%)

Working Space and Culture survey demographics (N=169)

Table 19: **Familiarity with the UW DSE among survey respondents who completed the Working Space and Culture survey section**

Familiarity with UW DSE	Respondents N (%)
I had not heard of the UW DSE prior to taking this survey	88 (52%)
I have heard of the UW DSE, but I have never interacted or worked with individuals involved in the UW DSE	33 (20%)
I sometimes interact and work with individuals involved in the UW DSE	19 (11%)
I often interact and work with individuals involved in the UW DSE	13 (8%)
I always interact and work with individuals involved in the UW DSE	13 (8%)
Other	2 (1%)

Table 20: **Position at UW among survey respondents who completed the Working Space and Culture survey section**

Position at UW	Percent of respondents (N=168)
Research scientist/engineer	36%
Graduate student	25%
Tenured faculty	15%
Postdoctoral student	13%
Non-tenured faculty	6%
Other	5%
Research faculty	4%
Data scientist	2%
Undergraduate student	0

Table 21: **Professional identity of survey respondents who completed the Working Space and Culture survey section**

Professional identity	Percent of respondents (N=168)
Domain scientist	49%
Data scientist	36%
Method scientist	26%
Engineer	22%
Other	12%

Table 22: **Primary Department of survey respondents who completed the Working Space and Culture survey section**

Primary department	Percent of respondents (N=16)
Other	21%
Oceanography	11%
Biology	11%
Genome Sciences	10%
iSchool	9%
Computer Sciences	8%
Physics	7%
Biostatistics	6%
Astronomy	5%
Sociology	5%
Human Centered Design & Engineering	4%
Statistics	3%
Applied Math	2%
Chemical Engineering	1%
Mechanical Engineering	0

Table 23: **Gender identity of survey respondents who completed the Working Space and Culture survey section**

Gender identity	Percent of respondents (N=162)
Male	62%
Female	36%
Self-identify	2%

Table 24: **Race/ethnicity of survey respondents who completed the Working Space and Culture survey section**

Gender identity	Percent of respondents (N=158)
White/Caucasian	84%
Asian	15%
Other	3%
Hispanic/Latino	2%
Native Hawaiian/Pacific Islander	1%
American Indian/Alaska Native	0
Black/African American	0

Table 25: **Years in age of survey respondents who completed the Working Space and Culture survey section**

Age in years	N	Minimum	Maximum	Mean
Age in years	147	22 years	77 years	41 years

Table 26: **Academic credentials of survey respondents who completed the Working Space and Culture survey section**

Academic credentials	Percent of respondents (N=163)
Doctoral	52%
Master's	25%
Bachelor's	22%
Associate's	1%
Alternative credential	1%
Professional degree	0

Table 27: **Years in academia among survey respondents who completed the Working Space and Culture survey section**

Years in academia	Respondents N (%)
0-2 years	31 (19%)
3-4 years	22 (14%)
5-6 years	17 (11%)
7-8 years	12 (8%)
9-10 years	13 (8%)
Over 10 years	66 (41%)

Career Paths and Alternative Metrics survey demographics (N=38)

Table 28: **Familiarity with the UW DSE among survey respondents who completed the Career Paths and Alternative Metrics survey section**

Familiarity with UW DSE	Respondents N (%)
I often interact and work with individuals involved in the UW DSE	13 (35%)
I always interact and work with individuals involved in the UW DSE	11 (30%)
I sometimes interact and work with individuals involved in the UW DSE	8 (22%)
I had not heard of the UW DSE prior to taking this survey	2 (5%)
I have heard of the UW DSE, but I have never interacted or worked with individuals involved in the UW DSE	2 (5%)
Other	1 (3%)

Table 29: **Position at UW of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Position at UW	Percent of respondents (N=37)
Tenured faculty	38%
Postdoctoral student	19%
Non-tenured faculty	14%
Research scientist/engineer	14%
Graduate student	11%
Research faculty	5%
Data scientist	5%
Other	3%
Undergraduate student	0

Table 30: **Professional identity of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Professional identity	Percent of respondents (N=37)
Domain scientist	57%
Data scientist	49%
Method scientist	32%
Engineer	14%
Other	5%

Table 31: **Primary department of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Primary department	Percent of respondents (N=37)
Astronomy	19%
Other	14%
Biology	11%
Computer Sciences	11%
Physics	11%
Oceanography	8%
Human Centered Design & Engineering	5%
Applied Math	5%
Statistics	5%
Biostatistics	5%
Genome Sciences	3%
Chemical Engineering	3%
Sociology	0
iSchool	0
Mechanical Engineering	0

Table 32: **Gender identity of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Gender identity	Percent of respondents (N=32)
Male	69%
Female	28%
Self-identify	3%

Table 33: **Race/ethnicity of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Race / ethnicity	Percent of respondents (N=31)
White/Caucasian	87%
Asian	13%
Hispanic/Latino	3%
American Indian/Alaska Native	0
Black/African American	0
Native Hawaiian/Pacific Islander	0
Other	0

Table 34: **Average age of survey respondents who completed the Career Paths and Alternative Metrics survey section**

	N	Minimum	Maximum	Mean
Age in Years	28	25	68	40

Table 35: **Academic credentials of survey respondents who completed the Career Paths and Alternative Metrics survey section**

Academic credentials	Percent of respondents (N=34)
Doctoral	82%
Master's	15%
Bachelor's	3%
Associate's	0
Professional degree	0
Alternative credential	0

Table 36: **Years in academia among survey respondents who completed the Career Paths and Alternative Metrics survey section**

Years in academia	Respondents N (%)
0-2 years	2 (6%)
3-4 years	2 (6%)
5-6 years	3 (10%)
7-8 years	2 (6%)
9-10 years	7 (23%)
Over 10 years	15 (48%)

Appendix B: Evaluation Plan

Education and Training

Final Evidence Collection & Analysis Framework for UW Data Science Environment

Outcome/Output	Evaluation Questions	Data Sources	Data Interval	Evaluation Tools	Data Analysis & Interpretation Plan
Increase number of **- User-friendly tools** **- Data science courses & curricula** **- Pi-shaped educators**	What is the baseline state of awareness across UW faculty, students and staff of e-science education activities?	• Faculty/Department leader interviews (N=8) • Grad student/Post-doc interviews (N=5) • Faculty survey (N=50) • Grad student/Post-doc survey (N=50)	Feb - March 2015 Feb - March 2015	• Customized interview protocol • Customized surveys	• Constant comparison qualitative method to identify emergent themes and inform survey development • Descriptive statistical analysis of survey data • Mixed-method analysis of changes occurring over the established time period
Cross-department events and training	What are the strengths of data science education and training? What are the observed gaps in data science education across the UW campus?				

Working Space & Culture

Final Evidence Collection & Analysis Framework for UW Data Science Environment

Outcome/Output	Evaluation Questions	Data Sources	Data Interval	Evaluation Tools	Data Analysis & Interpretation Plan
Number of data scientists and domain scientists working together on projects	What activities have helped catalyze collaborations between domain scientists and data scientists? Which of those activities have occurred in the Data Science Studio?	• Faculty (tenure-track and non-tenure-track)/Data scientist interviews (N=10) • Grad student/post-doc interviews (N=5) • Faculty/Data scientist survey (N=100) • Grad student/post-doc survey (N=50)	Feb - March 2015 Feb - March 2015	• Customized interview protocol • Customized survey	• Constant comparison qualitative method to identify emergent themes and inform survey development • Triangulation of data scientist, dept. leader and faculty interview data • Descriptive statistical analysis of survey data • Mixed-method analysis of changes occurring over the established time period
Inter-domain collaborations	What are the barriers to collaboration across domain scientists and data scientists?				

Software Tools, Environments & Support

Final Evidence Collection & Analysis Framework for UW Data Science Environment

Outcome/Output	Evaluation Questions	Data Sources	Data Interval	Evaluation Tools	Data Analysis & Interpretation Plan
Technique and technology adoption by scientific community	How are different types of techniques and technologies used by different populations on campus?	• Faculty/Data scientist interviews (N=10) • Grad students/Post-doc interviews (N=5)	Feb - March 2015	• Customized interview protocol • Customized surveys	• Constant comparison qualitative method to identify emergent themes and inform survey development • Triangulation of data scientist, dept. leader, post-doc and faculty interview data • Descriptive statistical analysis of survey data • Mixed-method analysis of changes occurring over the established time period
Adoption of data-intensive discovery	To what extent do researchers feel that there are important science questions they cannot pursue due to lack of appropriate software, methods, algorithms, or support?	• Faculty/Data scientist survey (N=100) • Grad student/Post-doc survey (N=50)	Feb - March 2015		

Career Paths & Alternatives

Final Evidence Collection & Analysis Framework for UW Data Science Environment

Outcome/Output	Evaluation Questions	Data Sources	Data Interval	Evaluation Tools	Data Analysis & Interpretation Plan
Awareness of need for additional career impact measures	How do UW institution/departments currently value/evaluate non-traditional products of career and scholarship?	• Department leader interviews (N=8) • Faculty/Data scientist survey (N=100) • Grad student/Post-doc survey (N=50)	Feb - March 2015 Feb - March 2015	• Customized interview protocol • Customized survey	• Constant comparison qualitative method to identify emergent themes and inform survey development • Triangulation of scientist, dept. leader and faculty interview data • Descriptive statistical analysis of survey data • Mixed-method analysis of changes occurring over the established time period
Recruiting of data scientists	What are the influences (emotional and other) on data scientists' career choices?	• Data scientist interviews (N=5) • Data scientist survey (N=50)	Feb - March 2015 Feb - March 2015		

Reproducibility & Open Science

Final Evidence Collection & Analysis Framework for UW Data Science Environment

Outcome/Output	Evaluation Questions	Data Sources	Data Interval	Evaluation Tools	Data Analysis & Interpretation Plan
Awareness of need for tools, code and data sharing/publishing in research labs/undergraduate education	What does reproducibility and open science mean to you and your peers? What do faculty, students, and scientists value about reproducibility and open science? What are the current practices and barriers to effective Reproducibility and Open Science code and data sharing?	• Department leader interviews (N=8) • Faculty/Data scientist interviews (N=10) • Faculty/Data scientist survey (N=100) • Grad student/Post-doc survey (N=50)	Feb - March 2015 Feb - March 2015	• Customized interview protocols • Customized surveys	• Constant comparison qualitative method to identify emergent themes and inform survey development • Triangulation of data scientist, dept. leader and faculty interview data • Descriptive statistical analysis of survey data • Mixed-method analysis of changes occurring over the established time period

Appendix C: Methodology

Nature of investigation

Using the UW Human Subjects Worksheet we determined that this formative evaluation study is not research as defined by the Common Rule. Therefore, this evaluation does not require IRB approval (Appendix D).

Although this preliminary report describes only our preliminary quantitative survey findings, this methodology section includes evaluation methods for both the quantitative and qualitative evaluation. We developed these data collections tools and methodologies to answer the evaluation questions identified by UW DSE leaders in the UW DSE theories of action and theories of change (Appendix E), and laid in the UW DSE evaluation plan (Appendix B).

Data collection and processing

1. Interviews

In early January 2015, data2insight worked with UW DSE working group leads to identify a purposeful sample of 26 key informants. UW DSE working group leads identified key informants to participate in interviews based on their knowledge and proximity to the UW DSE, or their potential to offer a valuable outsider's perspective. This purposeful sample included data scientists (N=6), junior faculty and librarians (N=6), postdocs and graduate students (N=5), and department chairs/senior faculty (N=9).

In late January 2015, data2insight conducted preliminary semi-structured telephone and in-person interviews with five (N=5) of the key informants identified for the purposeful sample. These individuals were selected by UW DSE working group leads for the value they could offer to the development of data collection instruments. Evaluators used preliminary interview data to inform the development of the survey and interview protocol (Appendix F).

In March 2015, data2insight evaluator, Lina Walkinshaw, conducted 16 in-person key informant interviews as part of data collection. Data2insight invited each key informant to participate via email; all individuals signed an electronic consent form prior to their interview. Data2insight sent weekly reminder emails to increase the response rate. These semi-structured interviews lasted approximately one hour, were audio recorded with permission, and transcribed for data analysis purposes. Audio recordings were deleted once transcription was complete. Following thematic analysis of the data, data2insight interviewed the remaining two key informants from the purposeful sample to clarify findings and answer questions that were raised as a result of data analysis.

UW DSE leaders provided feedback and gave final approval for all data collection instruments prior to their use.

2. Surveys

In February 2015, eScience Institute program management administered the UW DSE survey using Catalyst (UW web survey administration platform). eScience Institute program management

administered the UW DSE survey from March 2 - March 27, 2015 to one purposeful sample (N=58) and two random samples (N=1353). Data2insight evaluators set up weekly survey reminders to be sent through Catalyst from the eScience Institute program managers; eScience Institute Director, Ed Lazowska, sent one personal email to the purposeful survey sample to encourage participation. Data2insight offered all survey participants the option to opt in to receive survey findings.

Data2insight and the UW DSE Ethnography working group chose to administer the survey in three distinct distributions in order to shorten the length of the survey, and increase response rate. Across all three distributions we achieved 169 complete survey responses, plus an additional 42 responses for part one of the random sample distribution. These three distinct distributions are outlined below.

Purposeful sample

Data2insight worked with the UW DSE Ethnography working group to create the purposeful survey sample (N=58) from members of the UW DSE data science community. Participants were selected based on their familiarity with the efforts of the UW DSE, including:

1. All UW DSE evaluation interviewees (N=26)
1. All UW DSE Steering Committee and Executive Committee members (N=29)
2. All eScience Institute affiliated Data Scientists and Postdocs (N=16)
3. All UW DSE Working Group Leads (N=6)

Data2insight received these contact lists from the eScience Institute program managers in February 2015. Data2insight removed individual duplicates from these lists using Excel. The final purposeful sample included 58 individuals.

This purposeful sample received the complete version of the UW DSE survey. The complete UW DSE survey included prompts designed to answer the evaluation questions for five of the UW DSE working groups: Education and Training, Software Tools, Environments and Support, Reproducibility and Open Science, Career Paths and Alternative Metrics, and Working Space and Culture.

To encourage participation, purposeful sample participants were offered the opportunity to opt in to a raffle for two $50 Amazon gift cards. Thirty-three (33) out of 58 individuals completed the survey, achieving a 57% response rate.

Random samples

Data2insight worked with the UW Library Data Services Coordinator and the UW DSE Ethnography working group to identify the random survey sample. On behalf of the UW DSE, the UW Library Data Services Coordinator requested and obtained access to the UW Enterprise Data Warehouse (EDW). Using the EDW, the Data Services Coordinator queried based on parameters identified by data2insight and the UW DSE Ethnography working group.

These parameters included all non-emeritus Faculty (N=1,077), Graduate students (N=1,497) and Research Scientists/Engineers (N=387) from the following departments:

- Oceanography
- CSE
- Astronomy
- Physics
- Biology
- Sociology
- Applied Math
- Biostatistics
- Genome Sciences
- Human Centered Design and Engineering
- Statistics
- iSchool

The Data Services Coordinator sent all contact information to data2insight. Data2insight compiled the email lists into one master survey sample. The sample was randomized using a Microsoft Excel (Office 365) random number generator. From the randomized list, data2insight identified two random samples (N=225 each). Prior to survey administration duplicates were removed to ensure that the each sample was unique.

The **first random survey sample** received survey prompts from three of the five working group survey sections: Education and Training, Software Tools, Environments and Support, and Reproducibility and Open Science. This survey distribution was administered to the first random sample of 75 Faculty, 75 Graduate students, and 75 Research Scientists/Engineers as described above (N=225).

At two weeks, the first random sample survey response rate was 12%. Given this response rate, in order to obtain our target of 150 complete surveys, we administered this survey to an additional 237 Faculty, 236 Graduate Students, and 152 Research Scientist/Engineers (N=625). The final response rate was 21% (178/850).

The **second random survey sample** received survey prompts from two of the five working group survey sections: Career Paths and Alternative Metrics, and Working Space and Culture. This survey distribution was administered to the second random sample of 75 Faculty, 75 Graduate students, and 75 Research Scientists/Engineers as described above (N=225).

At two weeks, the second random survey sample response rate was 18%. Given this response rate, in order to obtain our target of 150 complete surveys, we administered this survey to an additional 98 Faculty, 97 Graduate Students, and 83 Research Scientists (N=278). The final response rate was 27% (136/503).

All random sample participants were offered the opportunity to opt in to a raffle for four $25 Amazon gift cards.

The data2insight evaluation team closed all surveys on March 27, 2015. Data2insight then downloaded all survey data from Catalyst in SPSS file format (.sav) for analysis. Data2insight labeled and merged data from the three surveys into one SPSS file for statistical analysis. Cross tabs were created for each prompt for the following groups by survey section:

- All respondents
- Random sample
- Staff (Scientists & Post Docs)
- Data Scientist
- Domain Scientist
- Purposeful sample
- Faculty
- Students
- Method Scientist

Document review

In January 2015, data2insight collected, compiled and reviewed background documents provided by UW DSE. These documents included:

- DSE annual reports submitted to the Moore and Sloan Foundations in November 2014
- UW DSE affiliated course, workshop, and bootcamp documents (e.g., course lists, workshop announcements, lecture slides) collected online and from UW DSE working groups in January 2015

Data Analysis

1. Quantitative Analysis

Evaluators conducted descriptive statistical analyses to determine the frequency of survey variables. Evaluators summarized quantitative findings using data visualization and written descriptions to answer working group evaluation questions.

2. Qualitative Analysis

After the completion of each interview, data2insight evaluators imported interview transcripts into Dedoose for coding and analysis. Evaluators developed qualitative codes to identify themes, ideas and concepts to analyze the qualitative data. Evaluators developed codes using grounded theory. Grounded theory is a qualitative data coding process that includes inductively discovering patterns and themes in the data, while also deductively applying an existing framework to understand the data (e.g., the UW DSE program map, and theories of action and change). When data fell into multiple thematic categories, we applied multiple codes.

Two data2insight evaluators each independently coded one transcript, using grounded theory to develop codes. The evaluators then met, compared, and reconciled their coding. After this first reconciliation, evaluators compiled all of the developed codes into a preliminary coding scheme. Evaluators next applied this preliminary coding scheme to a second transcript, and met again to reconcile differences in application and identify any additional emergent codes. After this second meeting the coding scheme was finalized, and one evaluator, Lina Walkinshaw, coded the remaining transcripts. Once the first evaluator has completed all coding, the second evaluator, Veronica Smith, performed quality assurance to ensure the codes were consistently applied across all transcripts.

After completing coding, evaluators used the coded data to perform a thematic content analysis. Content analysis is the overarching process of analyzing text to extract meaning.

Source document review consisted of tallying all the data science courses and eScience Institute educational activities offered through January 2015. Our review of course and educational offerings informed the semi-structured interview protocols, and provided context for understanding and analyzing the interview and survey findings.

Appendix D: IRB Human Subjects Worksheet

UNIVERSITY of WASHINGTON
HUMAN SUBJECTS DIVISION

WORKSHEET: Human Subjects Research

PURPOSE and INSTRUCTIONS

This worksheet provides **support** for individuals in determining whether an activity is human subjects research. It is completed and retained only when the activity is determined to be **Not Human Subjects Research**. The sections should be considered in the order presented. The term "data" refers to information of all types (information, records, specimens, recordings, photos, X-rays, etc.)

1. Are you a (select your position to reveal the correct Research Study Information box):

- (•) Researcher or Research Coordinator, etc. (not HSD Staff)
- ( ) HSD Staff Member

1. Research Study Information

PI Name:	Study Title:
Veronica S. Smith	UW Data Science Environment Year 1 Evaluation

END PART ONE

2. Research as defined by the Common Rule (45 CFR 46)

Check all that apply.

- ☒ A. The activity is an **investigation**.
  Investigation: A searching inquiry for facts, or detailed or careful examination.
- ☒ B. The investigation is **systematic**.
  Systematic: Having or involving a prospectively identified approach to the investigation, based on a system, methods, or plan.
- ☒ C. The systematic investigation is **designed** to **develop** or **contribute** to **knowledge**?
  Designed: the activity has a predetermined purpose and/or intent.
  Develop: to form the basis for a future contribution.
  Contribute: to result in.
  Knowledge: truths, facts, information.
- ☐ D. The knowledge the systematic investigation is designed to develop or contribute is **generalizable**.
  Generalizable: the data and/or conclusions are intended to apply more broadly beyond the individuals studied, or beyond a specific time and/or location, such as to other settings or circumstances.
- ☐ E. All of the following apply to the activity:
  - ☐ The activity is designed to contribute to the solution of social and health problems, or the evaluation of public benefit and service programs.
  - ☐ The activity involves the use of individually identifiable records from one or more of the following state institutions:
    - WA State Department of Social and Health Services (DSHS)
    - WA State Department of Corrections (DOC)
    - WA State Department of Health (DOH)
    - WA State Department of Early Learning
    - Any WA State Institution of higher education, including the UW and UW Medicine
  - ☐ The records pertain only to living individuals.
  - ☐ The records will be obtained and used without the informed consent of the person to whom the records pertain or the persons' legally authorized representatives.

Conclusion

All boxes are checked: the activity is Research as defined by the Common Rule. Proceed to Part 3.

If boxes A, B, C, and D are checked: the activity is Research as defined by the Common Rule. Proceed to Part 3.

If boxes A, B, C, and E are checked: the activity is not Research as defined by the Common Rule. However, IRB review is nonetheless required because of Washington State law RCW 42.48 concerning the research use of certain state records. Proceed to Part 6 to determine whether the activity is considered to be human subjects research by the Food and Drug Administration (FDA).

All other combinations of checked boxes. The activity is not Research as defined by the Common Rule. Proceed to Part 6 to determine whether the activity is human subjects research as defined by the Food and Drug Administration (FDA).

END PART TWO

3. Human Subject as defined by the Common Rule (45 CFR 46): Component 1

Check all that apply.

☐ A. The data are identifiable patient health information collected in Washington State that will be used for research purposes without the consent of the patient (whether living or deceased) or the patient's legally authorized representative.

Living: individuals who are alive according to applicable local and national regulations. With respect to specimens, data, and other information gathered without direct interaction with the individual: it is assumed that the individuals are living unless there is reason to think otherwise.

Washington State law: There is no Washington State law that defines "human subject". However, using identifiable patient health information collected in Washington State for research purposes, without the consent of the patient or the patient's legally authorized representative, requires IRB review (WA RCW 70.02.05). This applies to both living and deceased patients.

☐ B. The data are **about** individuals, and the individuals are **living**.

About: the data relates to the person. Asking individuals what they think about something (asking for an opinion) is almost always about the person. Asking for factual information, or other questions where the answers are expected to be independent of the person being asked, are generally not about the individual.

Examples:

- *A survey of elementary school teachers that asks them factual questions about class size, classroom features, and availability of classroom materials would generally not be considered to be **about** the teachers and would therefore not involve human subjects.*
- *A survey of elementary school teachers that asks them their **opinions** about the standard curriculum would generally be considered to be about the teachers.*
- *A researcher is developing a new user interface for a computer program. His research uses the "think aloud" method whereby he asks college students to verbally express their thought processes as they use the interface. Though the object of his interest is the interface, not the students, he is nonetheless collecting data **about** the students.*
- *Suppose you ask individuals, "How does your hospital respond to confidentiality breaches?" If you are seeking information about the hospital and are asking people who should know the answer, then the question is not about the person being asked. If you are seeking how often employees know the correct answer, then the question is about the person.*

Conclusion

If neither box is checked: human subjects are not involved, as defined by the Common Rule. Proceed to Part 6 to determine whether the activity is human subjects research as defined by the Food and Drug Administration (FDA).

If only box A is checked: proceed to Part 4.

If only box B is checked: proceed to Part 4.

If both boxes are checked: go to Part 4 to complete the determination about whether any federal human subjects regulations apply. However, regardless of outcome, the activity requires IRB review or exempt status because of Washington State law RCW 70.02.

END PART THREE

4. Human Subject as Defined by the Common Rule (45 CFR 46): Component 2

Check if "**YES**".

☐ The data are obtained through **intervention**.

Intervention: Physical procedures, or manipulations of the individuals or the individual's environment, that are performed for research purposes. Manipulations may be physical, social, psychological, or emotional. "Environment" includes an individual's social and virtual environments as well as physical environment.

*Obtain: Record in any fashion (writing, video, email, voice recording, etc.) for research purposes and retain for **any** length of time.*

Note: Individuals who "screen out" of a study because an intervention or interaction (such as a screening phone call or lab test) reveals that they do not meet the study eligibility criteria are still considered human subjects if they meet all of the criteria outlined in this Worksheet.

☐ The data are obtained through **interaction**.

Interaction: Communication or interpersonal contact between a member of the research team and the individual. Surveys - whether in-person, web-based, mail, email, phone, etc. - are an interaction between researchers and individuals.

*Obtain: Record in any fashion (writing, video, email, voice recording, etc.) for research purposes and retain for **any** length of time.*

*This includes so-called "**third party**" or "**secondary subject**" situations in which researchers obtain information about one individual through interaction with another individual. Example: a researcher is studying married couples where one spouse is the primary caregiver for the other spouse who has Alzheimer's disease and who is living at home. When the caregiving spouse provides the researcher with individually-identifiable private information about the spouse with Alzheimer's as a required part of the protocol, then the spouse with Alzheimer's is a human subject.*

Note: Individuals who "screen out" of a study because an intervention or interaction (such as a screening phone call or lab test) reveals that they do not meet the study eligibility criteria are still considered human subjects if they meet all of the criteria outlined in this Worksheet.

Conclusion

If one or both boxes are checked: human subjects are involved, as defined by the Common Rule. Proceed to Part 6 to see if the research is subject to FDA regulations.

If neither box is checked: proceed to Part 5.

END PART FOUR

5. Human Subject as Defined by the Common Rule (45 CFR 46): Component 3

Check if "**YES**".

☐ The data will be obtained about the individuals is **private** information.

Private information is defined as one or both of the following:

1. *Information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place.*
2. *Information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record or a residual medical specimen that is "leftover" from a health care procedure).*

Information: records, specimens, x-rays, photos, recordings and all other types of data

Publicly available data are not considered private.

Any information about the individuals that is collected specifically for the proposed research project through an interaction or intervention with the individual by the investigator or other member of the research team is considered to be private.

If permission is required to obtain information, then the information is usually considered private.
There are numerous "gray" areas in distinguishing "private" from "non-private". For example, there are some situations that are best considered "semi-private". This may include some behaviors, communications, and interactions that occur in electronic or social media. Also, a specific type of information may be considered private for one group of individuals but not for another.

☐ The private information is **identifiable**.

Identifiable: The identity of the individual is or may readily be ascertained or associated with the information by a member of the research team, OR a member of the research team could readily identify the individual through a constellation of the data variables.
Research team: Anyone involved in conducting the research. The act of solely providing coded private information is not considered involvement. However, individuals who provide coded private information are considered involved if they collaborate on other activities related to the research, including (but not limited to) (1) study interpretation, or analysis of the data resulting from the coded information; or (2) authorship of presentations or manuscripts related to the research.

Obtaining identifiable private information or specimens includes but is not limited to: Using, studying, or analyzing for research purposes identifiable private information of specimens that have been provided to researchers from any source, and recorded in any fashion for research purposes and retained for any length of time, or that were already in the possession of the researcher.

Note that the definition of "identifiable" is not the same as in the HIPAA regulations about health care records. Information that is considered an "identifier" by HIPAA regulations may not meet the federal

human subjects definition of "identifiable".

Some specific circumstances in which the information would not be considered identifiable:

- *The identifiers or the key to the identifier code have been destroyed.*
- *The research team has entered into an agreement with the holder of the identifiers or code key that prohibits the release of the identifiers or code key to the team members.*
- *When the data come from a repository or data management center: There are IRB-approved written policies and procedures for the repository or center that prohibit the release of the key to the team members.*
- *There are other legal requirements prohibiting the release of the identifiers or code key to the team members.*

Conclusion

If both boxes are checked: human subjects are involved, as defined by the Common Rule. Proceed to Part 6 to see if the research is subject to FDA regulations.

If one or no box is checked: human subjects are not involved, as defined by the Common Rule. Proceed to Part 6 to see if the research is subject to FDA regulations.

END PART FIVE

6. Human Subjects Research as Defined by the Food and Drug Administration

The UW may be involved in conducting only some components of an FDA-regulated study. When the components are limited to the following activities, it is UW policy that those research activities do not meet the FDA's definition of human subjects research when the UW-conducted activities meet ALL of the following criteria:

- The activities conducted by UW personnel are limited to any of the following:
  - Data analysis (whether or not the data are identifiable);
  - Analysis of specimens;
  - Accessing and providing medical records of participants.
- The UW is not the study clinical coordinating center.

Check if "**YES**".

☐ FDA definition of research. The activity involves any of the following (check all that apply):

- ☐ The use of a drug (whether approved or unapproved) in one or more living persons other than use of an approved drug in the course of medical practice.
- ☐ The use of a device (whether approved or unapproved) in one or more living persons that evaluates the safety or effectiveness of the device.
- ☐ Use of a test article regulated by the FDA (drug, device, biologic, etc.) to obtain data regarding subjects or control subjects that is intended to be eventually submitted to or held for inspection by the FDA.

☐ FDA definition of human subject. Either or both of the following are true:

- ☐ The research involves a living individual who is or becomes a participant in research, either as a recipient of a test drug, device (including in vitro diagnostics) or biologic, or as a control. The individual may be either a healthy individual or a patient.
- ☐ An individual on whose specimen an investigational device or control is used in the research, even if the specimen is anonymous.

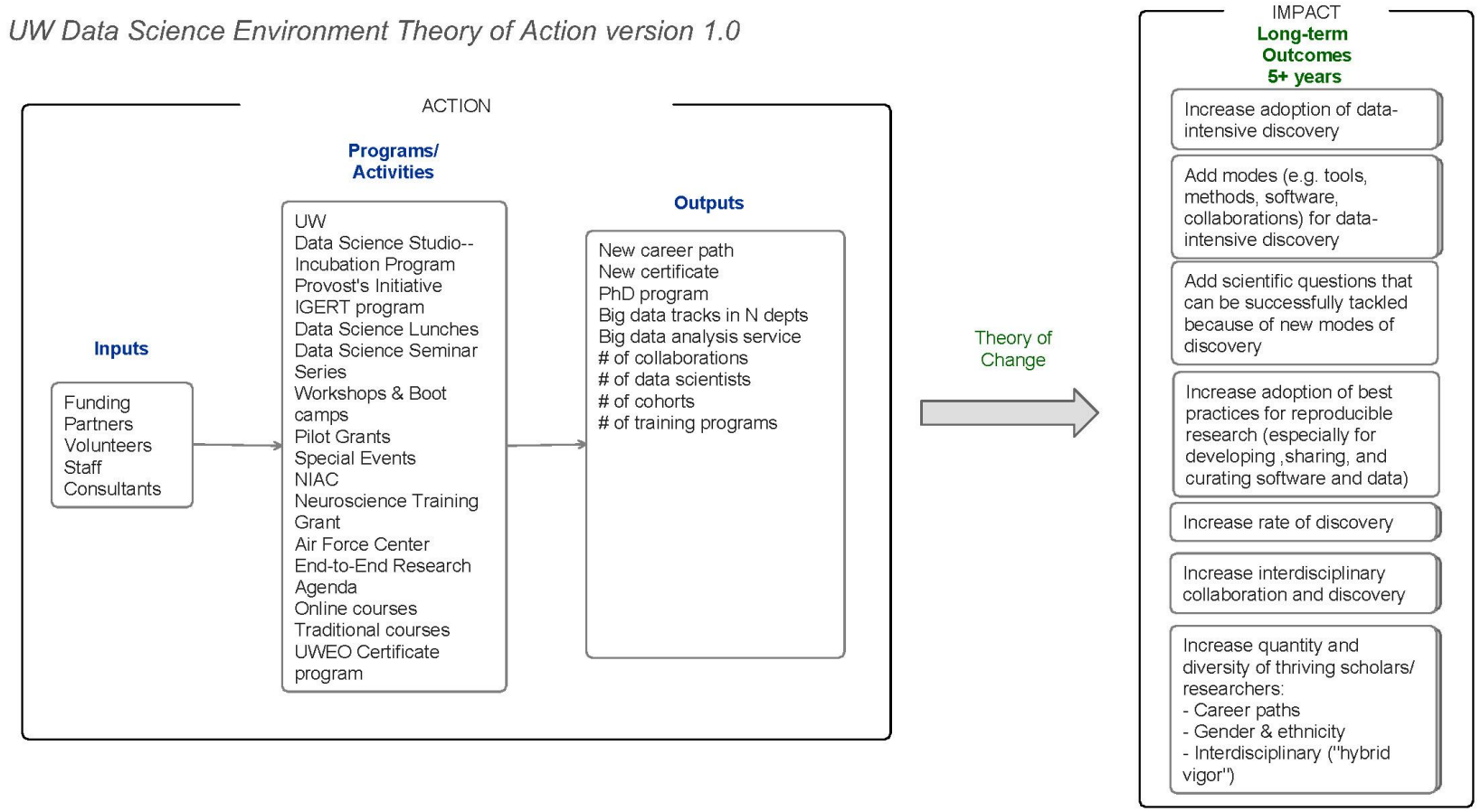
Conclusion

If both boxes are checked: the activity is human subjects research as defined by the FDA.

If no or one box is checked: the activity is not human subjects research as defined by the FDA.

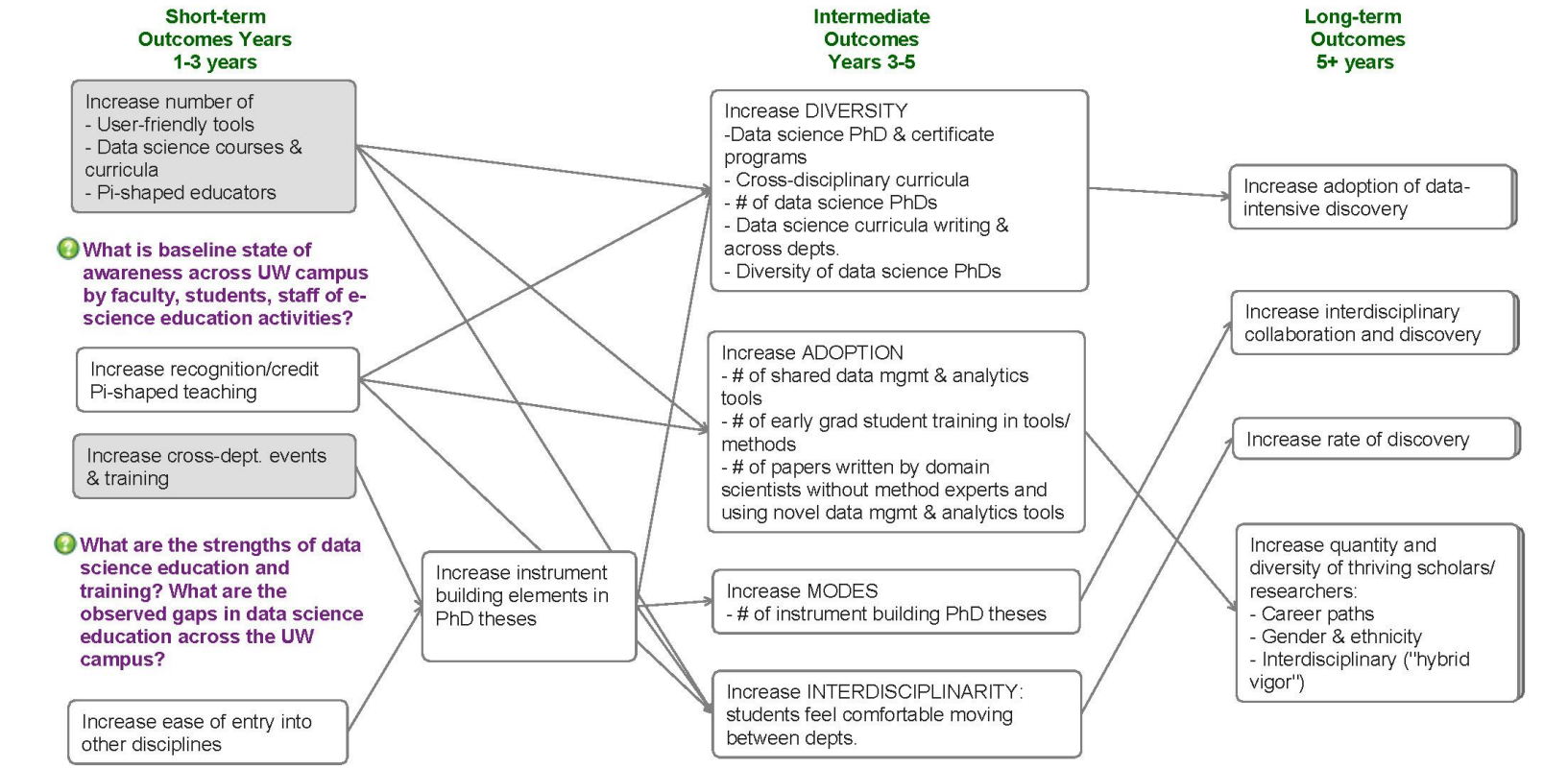
END PART SIX

Appendix E: Theories of Action and Theories of Change



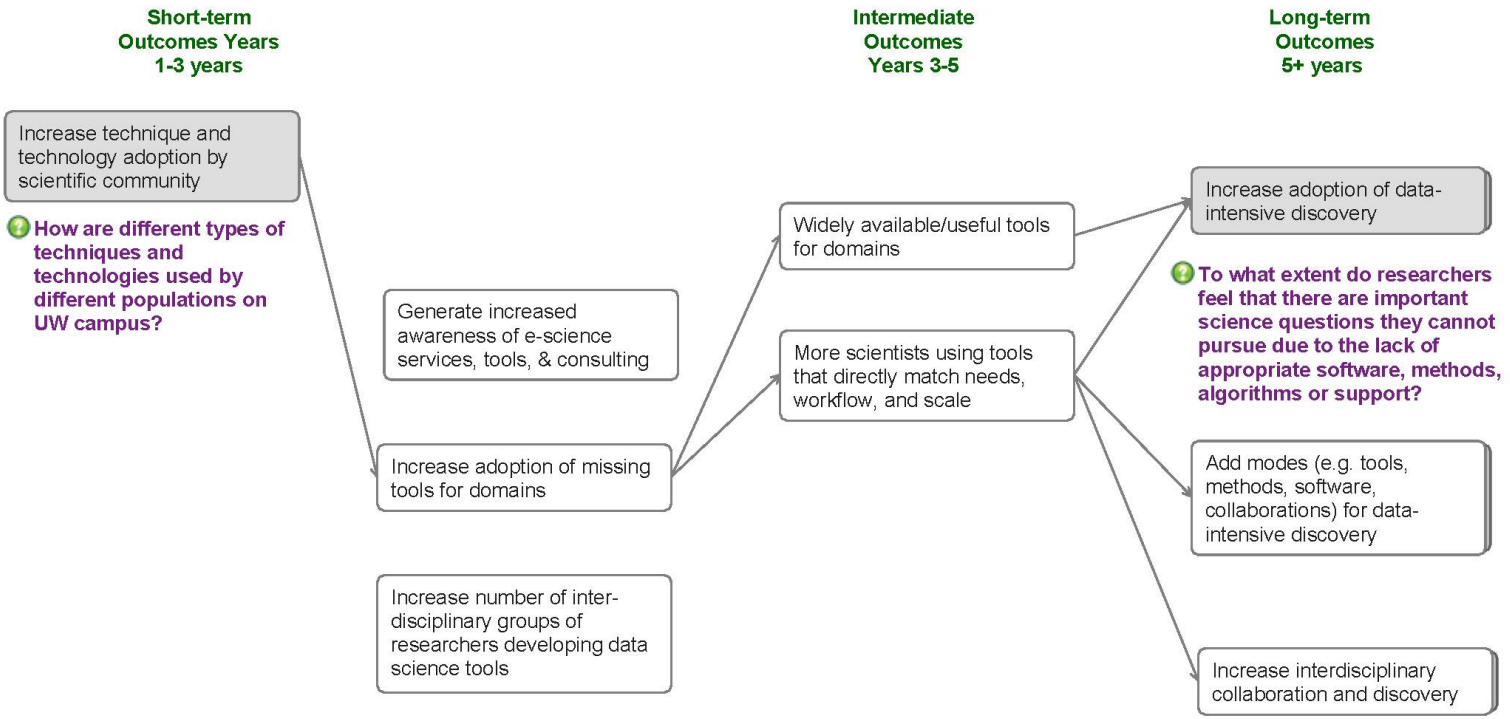
Education and Training Theory of Change

version 1.0



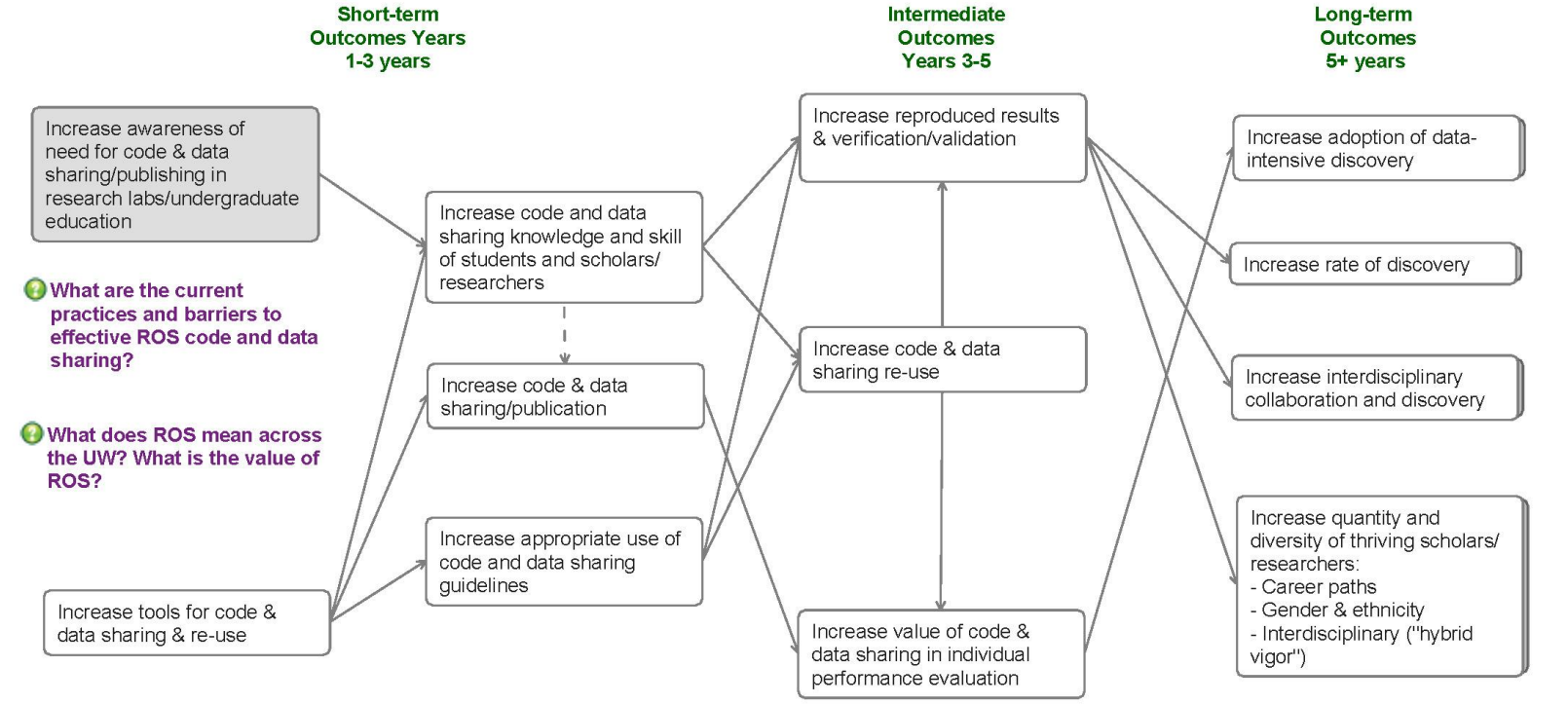
Software Tools, Environments, & Support Theory of Change

version 1.0



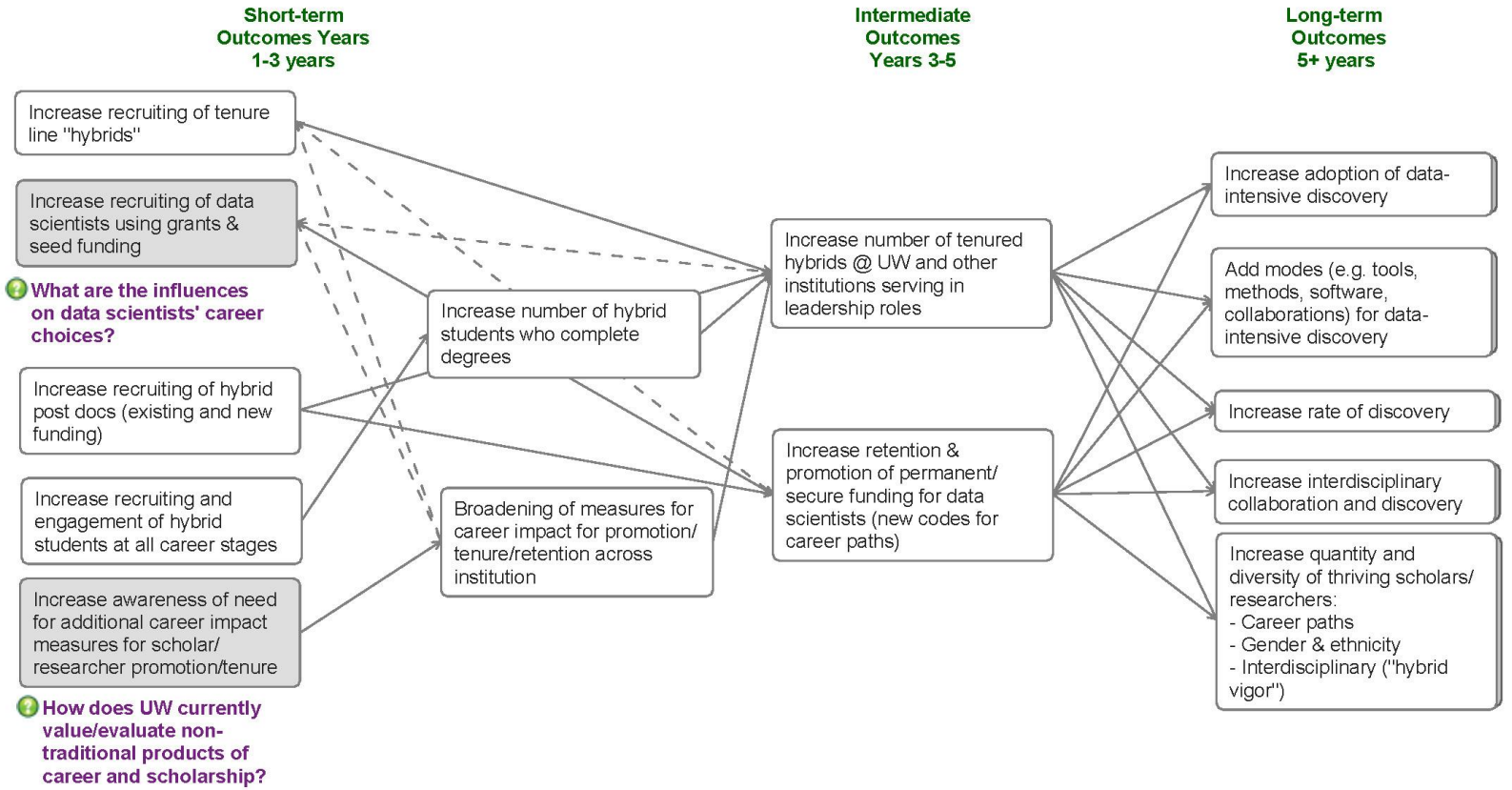
Reproducibility and Open Science (ROS) Theory of Change

version 1.0



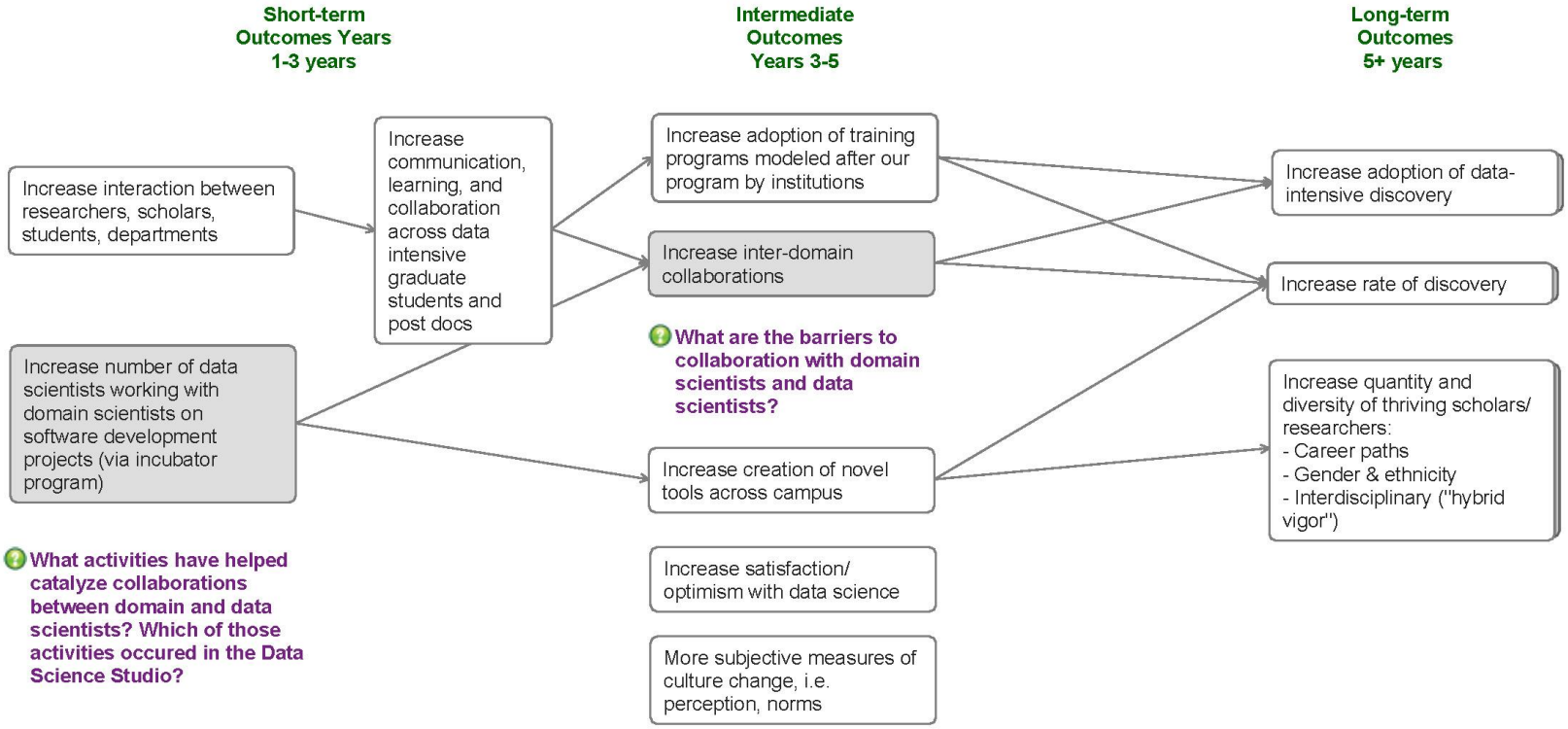
Career Paths and Alternatives Theory of Change

version 1.0



Working Space and Culture Theory of Change

version 1.0



Appendix F: Preliminary Interview Protocol

University of Washington Data Science Environment
Preliminary Interview Protocol

Thank you for taking the time to meet with me today. We are conducting the external evaluation for the University of Washington Data Science Environment (UW DSE). The purpose of this evaluation is to help UW DSE leaders answer key evaluation questions that will inform the development of the UW DSE. The evaluation questions we have been asked to answer focus on the five UW DSE working groups:

1. *Software Tools, Environments and Support*
2. *Education and Training*
3. *Reproducibility and Open Science*
4. *Career Paths and Alternative Metrics*
5. *Working Space and Culture*

We are conducting preliminary interviews with six individuals involved and familiar with the UW DSE in order to inform the development of surveys and interviews that we will be conducting February – March to answer the evaluation questions.

Our discussion will be recorded so we can transcribe it and capture the important themes. Do you have any questions for me before we get started? May I begin recording?

Introduction

1. What is your current position at the UW?
2. How are you related/connected to the UW DSE?

Software Tools, Environments, and Support

I have a few questions about terminology, and defining a handful of the common terms used in data science.

1. Can you please (a) define and (b) provide an example for each of the following terms/phrases:
    - Software tools
    - Software environments
    - Software support
    - Techniques and technologies
    - Methods
    - Algorithms
4. Which of these terms, if any, do you use interchangeably?
5. Are these terms used consistently across the UW Campus? If not, are there different groups that prefer/use different phrases to describe these elements of data science?
6. What research questions, if any, are you deterred from working on answering due to concerns about lack of available software tools, environment, or support?

7. How do scholars and researchers, in general, perceive data science, in terms of its potential to expand discovery going forward?

Reproducibility and Open Science

1. Could you please describe for me what 'reproducibility and open science' (ROS) means to you?
2. Do your peers have the same understanding of ROS as you? If not, what are the differences in your understandings?
3. What do you and your peers value about ROS?
4. What are the barriers you have experienced or observed to code and data sharing in research?

Education and Training

1. Could you please name the e-science education activities at UW that you are currently aware of or have participated in?
    - e.g., graduate and undergraduate education courses, Master's programs and certificates, seminars and weekly bulletins, and extra-curricular boot camps and workshops like the 'Software Carpentry Boot Camp' and the 'Astro Hack Week.'
2. Who in your network has expressed interest or participated in e-science education activities on campus?
3. What e-science education topics, if any, have you discussed with faculty, staff or students in the past 6 months?
4. What is essential to quality data science education (qualities and/or content)?
5. What skill and training gaps currently exist in data science education on the UW campus?

Working Space and Culture

1. Could you please describe the difference, from your perspective, between domain and data scientists?
2. How would you describe the current level of collaboration between domain scientists and data scientists on the UW campus?
3. What activities have you participated in at the UW that you feel have helped facilitate collaboration between domain and data scientists?

Career Paths and Alternative Metrics

1. What are the most common motivators for data scientists to work in academia?
4. What are the most common motivators for data scientists to work outside academia (e.g., Google)?
5. What is your sense of the general job satisfaction of data scientists currently working at the UW? Why?
6. What are the current career incentives for data scientists at the UW?
    - e.g., compensation, recognition, prestige, advancement
7. What are the most valuable data scientist products of career and scholarship at the UW?
    - e.g., publications, presentations, software development, cyber-infrastructure, technique and technology development

University of Washington Data Science Environment
Interview Protocol

Thank you for taking the time to meet with me today. We are interviewing approximately 25 key faculty, scientists and researchers, postdocs and graduate students to help University of Washington Data Science Environment (UW DSE) leaders learn more about the perceptions and awareness of, involvement in, and engagement with data science at the UW. What do you know about UW DSE?

The University of Washington eScience Institute was established in 2008 to help ensure that UW is a leader in both advancing the techniques and technologies of data-intensive discovery, and in making them accessible to researchers in the broadest imaginable range of fields.

In 2013, in partnership with the Center for Statistics in the Social Sciences, the eScience Institute secured a 5-year, $37.8M grant from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation - joint with NYU and UC Berkeley - to conduct a distributed collaborative experiment in creating what the Foundations referred to as "Data Science Environments" - conditions in which data-intensive discovery would truly thrive. The project launched a bit more than one year ago, at the end of 2013.

This interview is part of an independent formative evaluation study. Findings will be included in a report to UW DSE leaders. Interview questions will focus on five of the UW DSE working groups:

1. *Education and Training*
2. *Software Tools, Environments, and Support*
3. *Reproducibility and Open Science*
4. *Working Space and Culture*
5. *Career Paths and Alternative Metrics*

Thank you for agreeing to participate in this evaluation study and for signing the consent agreement. Do you have any questions about that? I would like to use an audio recorder to record your responses; we will use the recording to transcribe your interview. Once we have the transcripts, we will delete the audio recordings. We will summarize the themes that emerge from interview responses in our report to UW DSE leaders. While no participants will be identified by name in our reporting, because of the small select interview sample size (N=26), we cannot guarantee anonymity.

Introduction

1. Do you have any questions for me before we get started? May I start recording?
2. For purposes of this interview, the term ‘data science’ refers to the application of *computational methods* and theories to analyze and extract knowledge from data.
3. To begin, could you please describe for me your role / position at the University of Washington?

Software Tools, Environments, and Support

Next, I would like to ask you about your use of software tools, environments, and support.

1. What are the most common methods and algorithms you use in your work?
    a) How do you use these methods and algorithms in your work?
2. What are the most common methods and algorithms you have observed being used by others across campus? *Which of these do you feel is the most novel?*
3. What are the most common software tools and environments you use in your work?
    a) How do you use these software tools and environments in your work?
4. What are the most common software tools and environments you have observed being used by others across campus? *Which of these do you feel is the most novel?*
5. Which research questions, if any, are you deterred from working on answering due to concerns about lack of available software tools, environments, or support?
6. Which additional software tools, environments, and/or support, if any, would contribute the most to answering the scientific questions you are researching?
7. In your experience, how do scholars and researchers (in general) perceive data science in terms of its potential to expand discovery going forward?

Reproducibility and Open Science

In this section I have several questions for you about your understanding of, application of, and experience with reproducibility and open science.

1. What does **reproducibility** mean to you? What does it mean to your peers?
    a) How important is reproducibility in your work?
    b) What do you value most about reproducibility in science?
    c) How can UW DSE improve reproducibility in science?
2. What does **open science** mean to you? What does it mean to your peers?
    a) How important is open science in your work?
    b) What do you value most about open science?
    c) What limits, if any, should be put on open science?
    d) How can UW DSE contribute to the ethical practice of open science?

Provide a copy of the DSE ROS guidelines and ask the following questions:

1. Have you seen these guidelines before?
    a) *If no:* Please review them, and then I will ask you a couple of questions.
        i. Do you use these guidelines in your work?
        ii. Would you consider using them in your work?
    b) *If yes:* To what degree do you employ these practices in your work? Please provide an example.

2. How familiar do you think UW faculty, staff and research scientists across campus are with these ROS guidelines? How familiar do you think students are?
3. What other ROS guidelines or 'best practices' do you use in your work, if any?
4. What barriers, if any, have you experienced or observed to code and data sharing in research?

Education and Training

My first questions for you are about data science education and training at the UW.

1. Please describe the eScience Institute education activities at UW that you are currently aware of or have participated in. (e.g., graduate and undergraduate education courses, Master's programs and certificates, extra-curricular boot camps and workshops like the 'Software Carpentry Boot Camp' and the 'Astro Hack Week.')
    a) How have you learned about eScience Institute activities in the past 6 months?
2. Who in your network has participated in data science education activities on campus? Who has expressed interest?
3. What elements or content are essential to quality data science education and training?
    a) Based on your observations, please provide one example of these elements or content at UW.
    b) What elements or content are missing, if any, from data science education at UW?

Working Space and Culture

In this next section, Working Space and Culture, I have several questions for you about your collaborations with other scientists, and your use of the Data Science Studio.

To help frame our discussion, we want to share descriptions of the terms data scientist, method scientist, and domain scientist. For our purposes "data scientist" refers to someone who bridges the gap between methodology and application. A "method scientist" might contribute advances in statistics, machine learning, data management, or visualization, but not have much interest in or interaction with the scientific disciplines that utilize these methods. Conversely, a "domain scientist" might contribute advances in a disciplinary field such as astronomy or sociology, but not in data science methods. The data scientist contributes both to data science methodology and domain discovery.

1. In your opinion, what are the key differences between data, method, and/or domain scientists?
2. What are the benefits of collaborations across data, method, and/or domain scientists to answer research questions?
3. What are the barriers to domain, data, and/or method scientists collaborating?
4. What activities have you participated in at the UW that you feel have helped facilitate collaboration between domain, data, and/or method scientists?
    a) Please describe one activity from the last 12 months that facilitated your successful collaboration with a scientist outside of your field.
5. Have any of the activities that have facilitated collaboration been held at the Data Science Studio?
    a) How have you used the Data Science Studio thus far, if at all?

Career Paths and Alternative Metrics

In our last section of questions, Career Paths and Alternative Metrics, we have several questions for you about what motivates data and method scientists. To frame this part of the interview, we offer three categories of motivators to consider:

Intrinsic motivators are those related to motivation and satisfaction at work on a daily basis like feeling respected for your work, gaining a sense of achievement, or variety and change at work.

Extrinsic motivators are tangible rewards or conditions like being rewarded monetarily, having regular work hours, or having intellectual status.

Lifestyle values are associated with how and where you want to live and work, how you want to spend your time, and how you value money. For instance, you may want to live in a big city, or be active in your community, spend time with family, and/or save money.

1. Based on your experience, what are the most common motivators for data scientists to work in academia?
2. What are the most common motivators for data scientists to work outside academia (e.g., Google, Amazon, Twitter)?
3. In your opinion, what products of career and scholarship are most indicative of excellence in data science?
4. How are the careers and scholarship of data scientists at UW currently evaluated?
5. What challenges do data scientists face working at UW?
    a) What are the benefits of being a Data Science Fellow or Data/Research Scientist at UW?
6. What would be the most important career incentives or opportunities for UW to provide in order to better attract and retain research scientists/engineers and/or data/method scientists?

Thank you for participating in this interview. Your feedback is appreciated and the information you provided will be used to improve the UW DSE.

Do you have any questions or other comments? Would you like to receive a copy of the evaluation findings?