Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey

R. Stuart Geiger*[1], Charlotte Mazel-Cabasse[1], Chihoko Cullens[1], Laura Noren[2], Brittany Fiore-Gartland[3], Diya Das[1], and Henry Brady[1]

[1] Berkeley Institute for Data Science, University of California, Berkeley
[2] Center for Data Science, New York University
[3] eScience Institute, University of Washington, Seattle

Last revised: 8 May 2018 (version 1.0)

Abstract: This report is based on a 2016 survey of members and affiliates of three institutes of data science at major U.S. research universities, focusing on career paths for data scientists within academia. After considering how our respondents define data science, we identify various activities, priorities, resources, and concerns around data science in academia, especially with respect to data science careers. We end by providing recommendations about how universities can better support an emerging set of roles and responsibilities around data and computation within and across academic fields.

Recommended citation: R. Stuart Geiger, Charlotte Mazel-Cabasse, Chihoko Cullens, Laura Noren, Brittany Fiore-Gartland, Diya Das, and Henry Brady (2018). *Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey*. Report. Berkeley, California: UC-Berkeley Institute for Data Science. URL: http://stuartgeiger.com/papers/careers-data-science-msdse.pdf

Contents

*Corresponding author: please contact stuart@stuartgeiger.com with any comments or questions

1 Introduction and research questions

1.1 Motivation: how is data science fitting into academia?

In recent years, "data science" has grown in both industry and across academic fields. As a term, data science became popular in the technology and business private sectors, with publications like the *Harvard Business Review* calling it "the sexiest job of the 21st century" (Davenport and Patil 2012). Today in academia, there is a similar enthusiasm over the promise of data science to transform both research and education. Universities around the world are increasingly developing courses, degree programs, institutes, initiatives, and schools specifically for data science. A portion of these academic efforts are intended to train students for careers as data scientists and data engineers outside of academia, where such skills are increasingly relevant across industries in the private sector. Academic data science has also been described as an emerging interdisciplinary field in its own right, providing a new paradigm of scientific inquiry that places concerns about data and computation at the center of scientific research (Fox and Hendler 2014; Hey, Tansley, and Tolle 2009; Mattmann 2013).

There has been much commentary and scholarship on careers in data science from a private sector perspective, including what universities ought to teach students so they can be successful in the private sector (Harris, Murphy, and Vaisman 2013; Manieri et al. 2015; Kim et al. 2016). In a different line of research and commentary, there has been increasing attention on academic career paths in general. In this study, we focus on career paths and prospects for those in academic data science. For those on a traditional or even non-traditional career path within academia, are positions in data science as attractive as they appear to be in the private sector? Or are the graduate students, postdocs, research staff, and faculty who call or consider themselves data scientists facing a somewhat different reception than their peers who have moved to the private sector?

In this report, we take an in depth look at this issue through a survey of members and affiliates at three institutes dedicated to data science at major U.S. universities: the UC-Berkeley Institute for Data Science, the eScience Institute at the University of Washington, and the Center for Data Science at New York University. We surveyed students, researchers, staff, and faculty with connections to these cross-disciplinary institutes, with 169 respondents completing our survey. More details about the survey distribution, response rates, and demographics are in the appendix. This survey research is also informed by an ongoing ethnographic and interview-based study of academic data science institutes that four of the authors of this report are also conducting in parallel with this research.

1.2 Previous work on academic careers across fields

While there have been no studies specific to the career paths of academic data scientists, many of the issues and concerns we identified are common across academic career paths in general. Previous literature has extensively surveyed and interviewed graduate students and early career researchers, generally finding that there is substantial uncertainty and ambiguity for those seeking academic positions (Russo 2011; Woolston 2017), with many "mixed messages" about if and how they should pursue academic careers. (Woolston 2015) Researchers have found that between 25% and 50% of graduate students become less likely to pursue academic research careers since starting their graduate programs (Roach and Sauermann 2017; Russo 2011). One 2011 survey found that 60% of U.S. graduate students are discouraged from seeking academic careers because they believe that traditional research careers are too competitive. (Russo 2011)

A joint report from the National Academies of Sciences and Engineering on "the postdoctoral

experience" found many issues across fields: the number of PhDs has been rising faster than the number of permanent positions; the proportion of researchers who take temporary positions (like postdocs) has been rising; and the median number of years researchers spend in temporary positions after their PhD has also been rising (a median of 3-4 years as of 2011). (Sciences, Engineering, and Medicine 2014) Previous research has tended to focus on more financial and economic factors, such finding stark pay gaps between post-PhD researchers who choose to take postdocs then enter non-academic positions, versus those who immediately take non-academic positions (Kahn and Ginther 2017). In our study, we include questions about financial and economic factors, but also find important issues around factors like intellectual freedom, the ability to work on interesting and meaningful problems, work/life balance, and long-term job security. We also contextualize these general issues with careers in academia to the specific domain of data science.

1.3 Research questions

Our primary and secondary research questions are:

- What is academic data science?
  - To what extent do various academics in our sample consider themselves data scientists?
  - How do self-identified data scientists define what it means to be a data scientist?
  - What activities and resources are important to academic data scientists?
- What kinds of careers do data scientists have and want?
  - What are the different kinds of career paths currently comprising academic data science?
  - What are the long-term career goals and priorities of academic data scientists?
  - To what extent are academic data scientists seeking or expecting tenure and other career structures for their work?
- What do different kinds of groups within academic data science need?
  - Can we cluster academic data scientists based on their goals and priorities?
  - How do experiences and goals vary across clusters and demographic groups?
  - What resources do academic data scientists say were helpful for their careers?
- What recommendations can we make for universities who want to support academic data scientists?

1.4 Summary of recommendations

- Academic data science involves a variety of new topics, roles, and activities (as well as novel combinations of established and new topics, roles, and activities), which are often not fully supported by traditional academic career paths. As there is no single model of what an academic data scientist does, universities should define and support a broad and diverse range of positions and career paths for data scientists, both within and across disciplines.
- For early career researchers, there can be substantial ambiguity and uncertainty about whether their academic institutions will reward and support long-term career paths for those who focus on topics, roles, and activities associated with data science. Universities should facilitate conversations within and across disciplines about formalized criteria and expectations for both tenure-track and non-tenure-track positions in and around data science.

- While salary is an important factor (with industry positions paying lucrative salaries), our respondents generally placed even more value on secure, long-term employment and intellectual freedom. Many academic data scientists are strongly seeking to avoid precarious positions, not wanting to be exclusively funded by "soft money." Universities should support tenure-track positions for data scientists, as well as long-term career paths (e.g. with 3-5 years of stability) for data scientists in non-tenure-track academic and research staff positions.
- Funding for research projects, computational infrastructure, and education/professionalization initiatives has been crucial to the success of academic data scientists, which should be continued and expanded. To further support academic data scientists, universities should also support a broad range of formal and informal training, mentoring, and professionalization initiatives. Such efforts should include both events that take place within and across disciplines and institutions, as some issues may be broadly applicable but others may be quite specific. Partnerships with industry-focused student career groups are also recommended.
- In working to support career paths in data science, it is important to work to support diversity and inclusion across many dimensions, including gender, race/ethnicity, national origin, class (including first generation college/grad students), and type of institution (e.g. large research-focused universities, four-year universities, and small liberal arts colleges).

2 What is academic data science?

2.1 Summary of findings

- There is no single universal definition about what a data scientist is and/or does, but respondents' open-ended definitions emphasized various aspects, which have a family resemblance to each other. The major themes we found are that data scientists are those who:
  - work with data, particularly data with high volume, variety, or velocity; and have skills in collecting, curating, and cleaning these kinds of data, in addition to analysis and visualization
  - develop and maintain computational resources (software and hardware) that support scientific data collection, curation, cleaning, analysis, and visualization
  - have specific combinations of expertise, such as computational, statistical, and domain or context-specific expertise; many define data science as working at the intersection of these fields
- We suggest that people have different definitions of data science because they have different ideas about positions and career paths in data science.
- Being a data scientist is not a binary status: respondents had differing degrees to which they did or did not self-identify as a data scientist, with a majority "somewhat" identifying as a data scientist.
- Self-identification varied substantially across fields: approximately 82-88% of those in statistics, mathematics, and the life, physical, and social sciences self-identified as a data scientist, while only about half of the computer scientists and engineers in our sample self-identified as a data scientist.

- Our sample population is highly interdisciplinary, with two-thirds of respondents working in multiple broad fields. Over a quarter of respondents' definitions of data scientists specifically discussed data scientists as those with expertise at the intersection of multiple fields.
- Positions in data science should reflect this interdisciplinary nature, such that both faculty and staff positions should support those who work in and make contributions across multiple fields.

2.2 Free response: What is a data scientist?

83 out of 169 respondents (49.1%) gave a response to the free response question "What is your preferred definition of a data scientist? (if you have one)." The answers varied widely, emphasizing various aspects of the roles, responsibilities, activities, skills, qualities, and disciplinary backgrounds – or stating that the term has no definition, as 5 respondents did. The mean length was 22 words, the median length was 18 words. We used an inductive and iterative grounded theory approach (Glaser and Strauss 1967) to identify themes and sub-themes in the responses and label each of the responses accordingly. These themes are non-exclusive, as definitions often included multiple themes.

Table 1 shows the identified themes and sub-themes, with an example definition from a respondent and the proportion of respondents whose definitions match the theme (out of all respondents who gave definitions). The three primary themes were that data scientists 1) work with data (in various ways); 2) develop tools, infrastructures, and/or methods for scientific research; and 3) have a particular disciplinary background or expertise. The sub-themes highlight different shared specifications within these themes. For example, while 73% of all definitions included some reference to working with data, only 10.3% of all definitions specified that data scientists work with data in an inductive or data-driven manner. We also had multiple definitions drawing on Drew Conway's venn diagram of data science as the intersection of domain expertise, hacking skills, and statistics/math skills (Conway 2013), as well as Josh Wills's oft-quoted definition — "Someone who can program better than any statistician and can do statistics better than any programmer."[1] Both of these definitions fall into our top-level theme of defining data science as an intersection of expertises.

While there has been much discussion and criticism of the term "data science" for not having a unified or formal definition, we do not see cause for concern in the differences our respondents gave. Disciplines, departments, and other established academic units have always struggled with defining and formalizing their scope and role, as well as reconciling internal differences between sub-fields. However, we find it constructive to consider that people in academic data science (and beyond) may have different definitions of data science because they have different ideas about positions and career paths in data science. If we see definitions in this way, these differences become more generative and productive, rather than incommensurable incompatibilities.

2.3 Data science as an interdisciplinary field

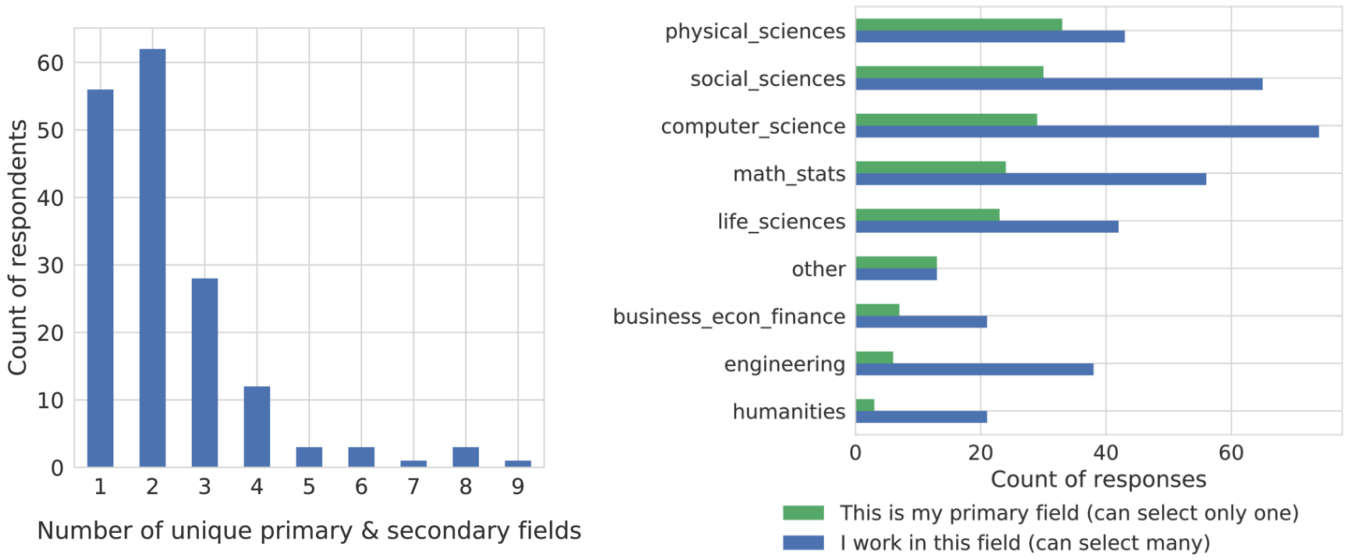
We asked respondents for both their primary field (where they could only select one) and the fields they worked in (where they could select many), listing a traditional grouping of academic fields that did not include data science. As Figure 1a shows, respondents were overwhelmingly working in multiple fields, with 33.1% reporting working in only one field, 36.7% working in two

[1] https://twitter.com/josh_wills/status/198093512149958656

Table 1: **Themes in definitions of data science**

Theme / *sub-theme*	%	***Example definition from survey***
Working with data	73.0%	"Anyone primarily working with data"
Working with data with high volume, variety, or velocity	29.5%	"Someone who extracts knowledge from large or messy datasets"
Working across a data workflow/pipeline	21.8%	"A data scientist can work in any part of the data chain from gathering, processing and QA/QCing data to machine learning/data analytics."
Doing analysis in an inductive or data-driven manner	10.3%	"Someone who does inductive quantitative research"
Developing tools, infrastructure, and/or methods for science	12.8%	"A data scientist not only gathers and/or uses data but also contributes to the development of more powerful tools and methods for gathering, processing, analyzing data."
Developing software tools	10.3%	"Someone who develops tools to use data to do science"
Disciplinary background / expertise	59.0%	
Computational	42.3%	"Somebody who applies the tools of computer science, programming, and data management to scientific problems in one or more domains."
Statistical	30.8%	"Someone who employs advanced statistical tools and machine learning to analyze data in a specific field or area or for a specific purpose."
Domain / context-specific expertise	30.8%	"Someone who uses large data sets to answer scientific questions and also has domain expertise"
Conway's definition	7.8%	"A data scientist has the right balance of domain expertise, computing and hacking skills."
Wills's definition	2.6%	"Someone who can program better than any statistician and can do statistics better than any programmer."

fields, and 30.2% working in more than two fields. Figure 1b shows the breakdown of each field by primary and secondary field. In terms of primary field, there was a roughly even proportion between the physical sciences (19.6%), social sciences (17.9%), computer science (17.3%), math and statistics (14.3%), and the life sciences (13.7%). However, the number of respondents working in various fields was much higher, such as 43.8% working in computer science, 38.5% working in the social sciences, and 33.1% working in math and statistics. Other fields that were less represented in terms of primary field were far more represented in secondary fields: only 3.6% selected engineering as their primary field, but 22.5% reported working in engineering; 1.8% selected the humanities as their primary field, but 12.4% reported working in the humanities.



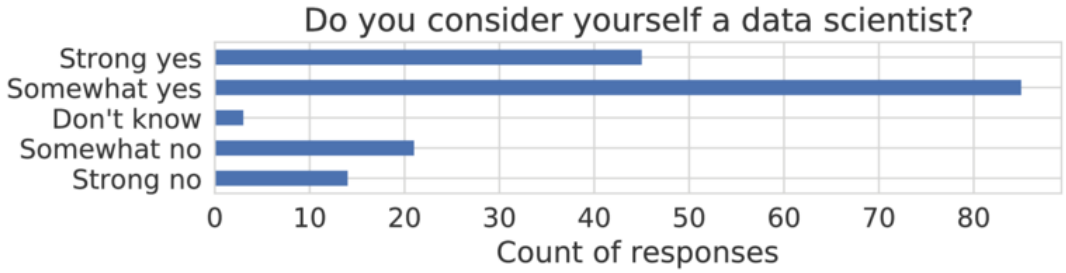
(a) **Number of fields**

(b) **Breakdown of primary and secondary fields**

Figure 1: **Respondents by academic field**

2.4 Identification as a data scientist

Respondents all had some formal relationship to one of the three institutes of data science, as we recruited using lists of those who had formal positions at, were funded by, or had some affiliation with the three institutes. However, not all respondents equally self-identified as data scientists. We asked respondents to what extent they identified as a data scientist, giving a five-point Likert scale of "strongly disagree" to "strongly agree" with a "don't know" option. While only 3 of 169 respondents chose "don't know", the distribution of responses was mixed, as Figure 2a shows. Approximately three quarters of our respondents considered themselves to be data scientists, with 50.6% selecting "somewhat agree" and 26.8% selecting "strongly agree". Figure 2b plots the self-identification as a data scientist by primary field, showing some interesting differences between fields. All of the respondents in the business, economics, and finance group either somewhat or strongly identified as a data scientist, while none of the respondents in the humanities group identified as a data scientist. Outside of the humanities, the fields with the lowest proportion of respondents who identified as a data scientist were engineering (50%) and computer science (55.2%). The remaining fields all had roughly equivalent proportions, between 82% and 88% identifying as data scientists.



(a) **Responses for "I consider myself a data scientist"**

(b) **Breakdown of identification as a data scientist by primary field**

Figure 2: **Respondents by academic field**

3 What activities do academic data scientists want to do?

3.1 Summary of findings

- There are various activities which our respondents almost uniformly valued highly. We recommend that academic data science positions across specialties, ranks, and academic/staff tracks should be structured to support and reward these activities, even if they are not a primary duty of the position:
  - conducting research for academic publications
  - advising students
  - collaborating outside of their institution
  - doing research in an open and reproducible manner
  - attending conferences
- There are many activities where respondents diverged on importance, indicating areas of specialization. Not all data science positions must to involve these activities, but it is important to have flexible career paths where they can become part of a data scientist's core duties. These are:
  - developing/maintaining computational resources
  - research consulting
  - writing grants

  - teaching short workshops and traditional semester/quarter-length courses
  - traditional academic service
  - managing a group or lab

- Existing career paths in academia are often segmented into activities typically performed by faculty and non-faculty. Many data scientists do not neatly fit into these existing career paths. A large proportion of our respondents want positions where they combine elements of faculty and non-faculty roles. Some of the clusters we have identified include:
  - Computational research data scientist: develop computational resources and do research work for academic publications
  - Teaching data scientist: teach short workshops and traditional semester/quarter-length courses
  - Consulting data scientist: do research consulting for academic researchers and advise students

- Those who strongly self-identify as a data scientist generally value developing computational resources, teaching courses and workshops, and research consulting more than those who do not. Otherwise, data scientists and non-data scientists generally place the same importance on activities.

3.2 Descriptive statistics about value of various activities

We asked respondents about "how much you value being able to do the following activities", specifically asking "in your ideal academic position, how important would it be for you to do the following activities?" There were seventeen different activities, including activities:

- that are officially a part of some academic positions (e.g. teaching courses, advising students, research work)
- that many in academic positions do even if they are not part of a job description or formal duties (e.g. attending conferences, developing computational resources, maintaining an academic social media presence),
- that are characteristic of certain styles of academic work (e.g. collaborating outside of one's own institution, making research open/reproducible), and
- that are not academic duties but instead reflect work/life balance issues (e.g. having/raising children, "the rest of life")

Respondents were given a five-point Likert scale ranging from "not at all important" to "very important," with an "I don't know" field at the end of the scale. Figure 3 plots the mean response for each activity, with responses coded as 5 for "very important," 1 for "not at all important," and "don't know" coded as 3.

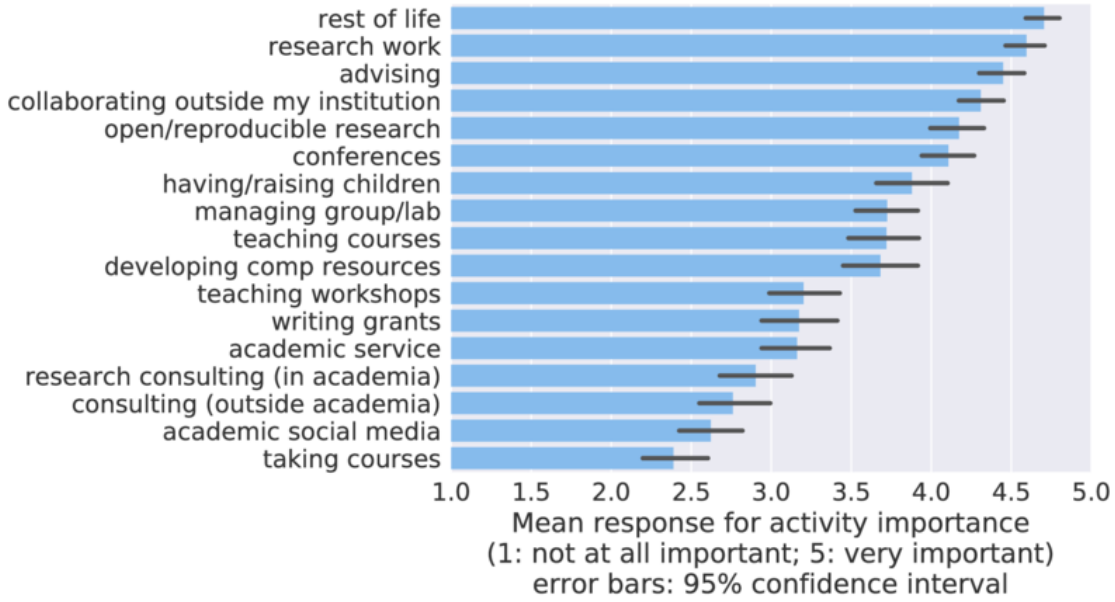


Figure 3: **Mean activity importance for all respondents.**

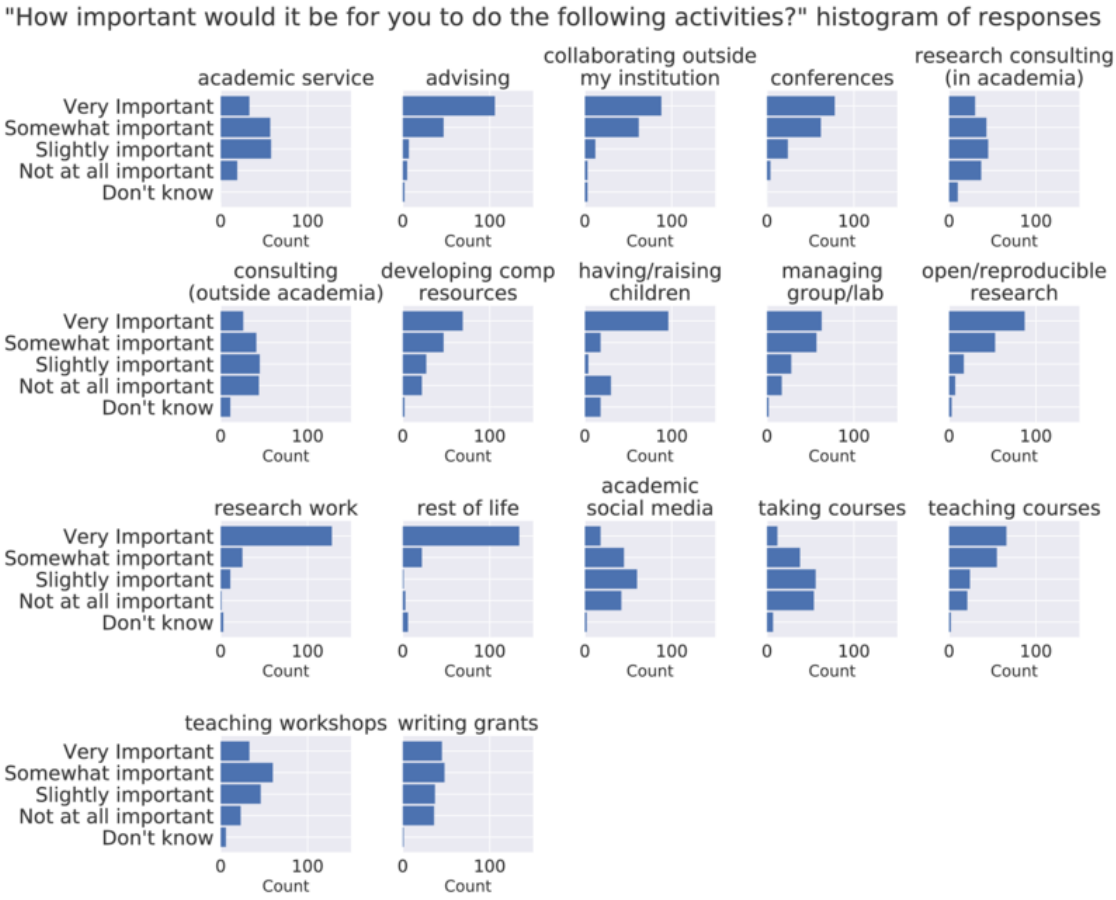


Figure 4: **Distribution of responses for activity importance across all respondents**

The highest valued activities and those with the smallest standard deviations were "the rest of life" (meaning that people valued their life outside of academia, capturing work/life balance), with a mean value of 4.70 (standard deviation of 0.73) and research work for academic publications, with a mean value of 4.60 (standard deviation of 0.86). Other activities with a mean response of over 4.0 were:

- collaborating outside one's own institution ($\bar{x} = 4.31, s = 0.95$)
- advising students ($\bar{x} = 4.45, s = 0.94$)
- making one's research open/reproducible ($\bar{x} = 4.17, s = 1.14$)
- attending conferences ($\bar{x} = 4.10, s = 1.12$).

These activities also had the lowest standard deviations, indicating more agreement between respondents than for other activities. These activities, with the exception of "the rest of life," can all be thought of as essential elements of building a highly visible and successful academic research career. Furthermore, it is important to support work/life balance across all academic positions and career stages.

The lowest valued activity was taking courses ($\bar{x} = 2.39, s = 1.33$), which likely reflects the career status of our respondents. Also less valued on average was maintaining an active academic social media presence ($\bar{x} = 2.62, s = 1.39$) and research consulting – for both other academics ($\bar{x} = 2.90, s = 1.47$) and those outside of academia ($\bar{x} = 2.76, s = 1.47$). These are often considered activities that are peripheral to a research-focused academic career, even though they may provide many benefits to an institution and the broader research community.

Figure 4, a stacked barplot showing the distribution of responses to this question, illustrates that activities like research consulting for academics are not uniformly less valued, but rather that there are fewer respondents who value it very highly: 18.2% of respondents stated research consulting for academics would be "very important" in their ideal position. Other professional activities with broad distributions (and higher standard deviations) include writing grants ($s = 1.56$), developing/maintaining computational resources ($s = 1.46$), and teaching courses ($s = 1.43$). **It is important to not forget about the subgroups who rate these activities as highly important to them, as they often perform valuable specialized roles that bring many benefits to academia, but may not be sufficiently rewarded and supported by traditional academic career paths.**

3.3 Factor analysis for activities valued

We then grouped respondents by their value of these activities, running a factor analysis seeking to cluster these activities, using R (R Core Team 2016) and the psych (Revelle 2017) package. We first included the ideal value of all activities except taking courses, having/raising children, and "the rest of life."[2] The factor analysis was directed to reduce dimensionality to two factors, and identified three distinct clusters of variables across two dimensions, as shown in the biplot in figure 5. The two factor analysis dimensions explain 28% of the variance between the responses across the 12 activities. The cluster aligned on the positive X axis mostly includes the highly rated activities related to a research career in the preceding section: advising students, teaching courses, managing a lab/group, and research work. The cluster aligned on the positive Y axis includes teaching workshops, research consulting for academics, consulting for public/private sector, attending conferences, maintaining an active social media presence, and collaborating outside of

[2]We excluded these two activities because they capture issues of work/life balance, which we expect to be a separate, orthogonal factor for data scientists choosing career paths than the combinations of various at-work activities.

one's institution. The cluster aligned on the positive diagonal of both axes contained writing grants, attending conferences, and traditional academic service.

The orthogonal nature of these two main clusters of activities is interesting and raises further research questions. While there is a clear clustering of activities that reflects traditional academic roles, there is no strong division in our sample between these two clusters. This is shown in Figure 5, which also plots the regression scores for all respondents. **This shows a wide distribution of people across all four quadrants, indicating a diversity of different sets of activities that people want to perform.** For example, there are people who want to exclusively focus on traditional faculty activities like research and teaching courses, while there are those who want to do this and tasks like developing computational resources, teaching workshops, and research consulting. Similarly, there are those who want to exclusively focus on these traditionally non-faculty tasks and not do tasks like research and teaching courses. We were surprised to the degree to which our respondents varied in the sets of tasks that they valued, and we believe that it demonstrates that there are various ways that people want to pursue their data science careers in academia.

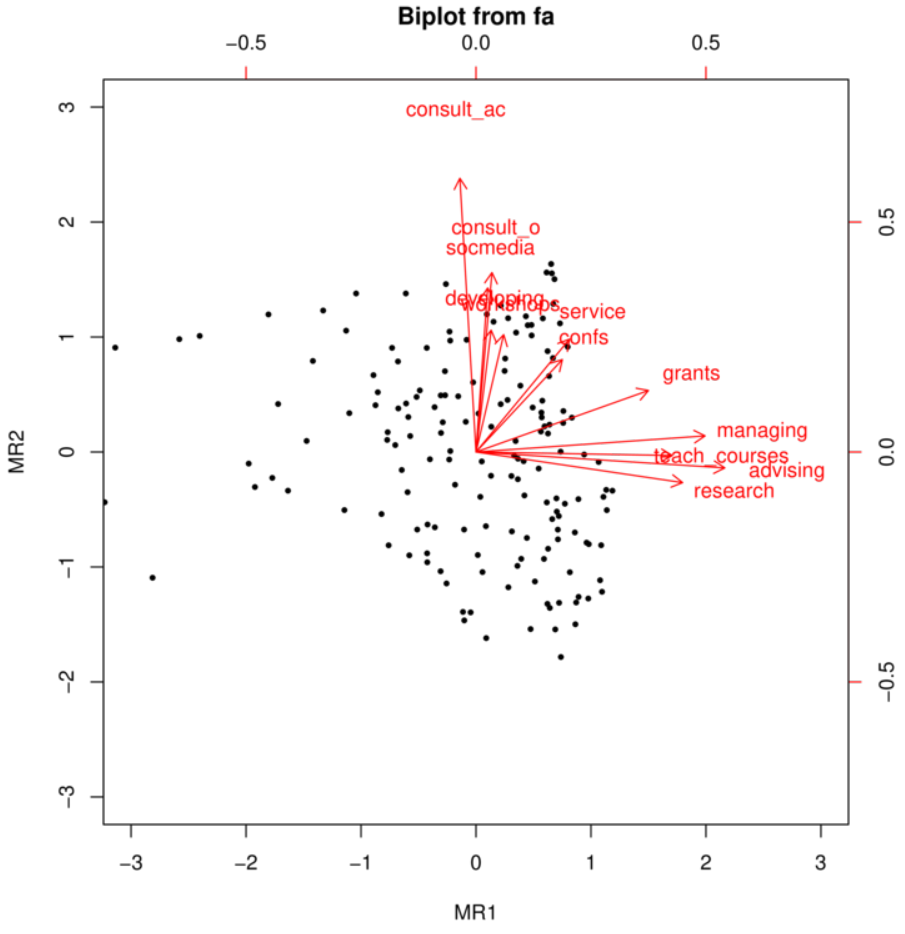


Figure 5: **Factor analysis for activity variables, with variable loading scores in red arrows and regression scores for respondents in points**

3.4 Activity importance based on self-identification as a data scientist

We also broke out activity importance responses by self-identification as a data scientist. As discussed in section 2.4, responses to the question "I consider myself a data scientist" of strongly disagree, somewhat disagree, and don't know were recoded as "No"; somewhat agree was recoded as "Somewhat yes", and strongly agree was recoded as "Strong yes." 50.6% of respondents were in the "Somewhat yes" category, 26.8% in the "Strong yes" category and 22.6% in the "No" category.

We broke this response down into the three levels after suspecting strong differences between the "somewhat yes" and "strong yes" groups. Figure 6 plots the difference in mean importance ratings for each activity between the Strong yes and Somewhat yes groups. Positive values indicate activities that are valued more by those who strongly identify as data scientists, negative values indicate activities that are valued more by those who only somewhat identify as data scientists. We ran individual linear regressions predicting the ordinal identification as a data scientist (1 to 5 scale) by the value of various activities.

Activities that are valued more by those who strongly identified as data scientists by a statistically significant measure include: developing computational resources (score=.15, p=0.021) and research consulting for academics (score=.13, p=0.039). For all other activities, there was no statistically significant difference between the Strong yes and Somewhat yes groups. Respondents who did not identify as data scientists placed a mean importance of 3.29 on developing computational resources, compared to a mean of 3.93 for those who strongly identified as data scientists and a mean of 3.71 for those who somewhat identified as data scientists.

As this survey is more exploratory and not based on testing a strict set of hypotheses, we advice against interpreting these regressions to strictly generalize and infer about broader populations. Given the number of multiple comparisons made in these 17 regressions, applying a Bonferroni correction means our p-values should only be interpreted as significant at the 0.003 significance level. We therefore cannot confidently infer that there are meaningfully significant differences in activity preferences between what those who do and do not identify as data scientists. This is likely because there is a broad diversity among data scientists in terms of what kinds of activities they would like to do. We do recommend that future work in this area focus heavily on the activities of developing/maintaining computational resources and research consulting in the context of data science careers.

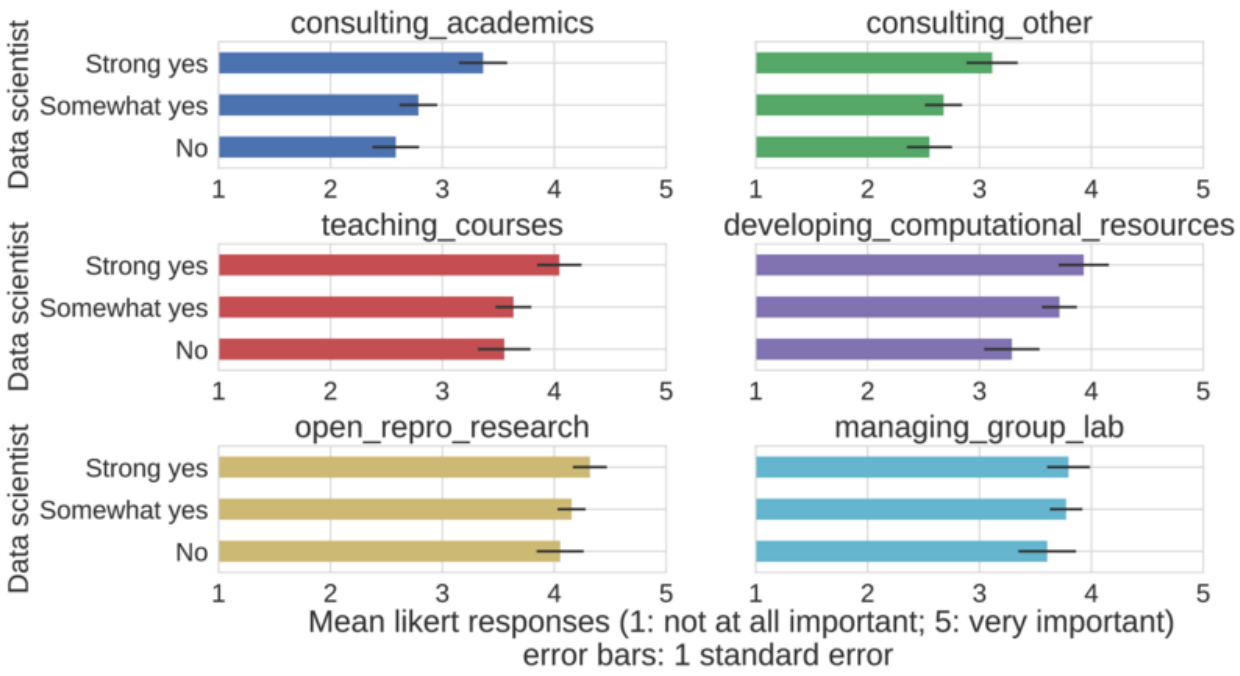


Figure 6: **Difference in activity importance based on self-identification as a data scientist.**

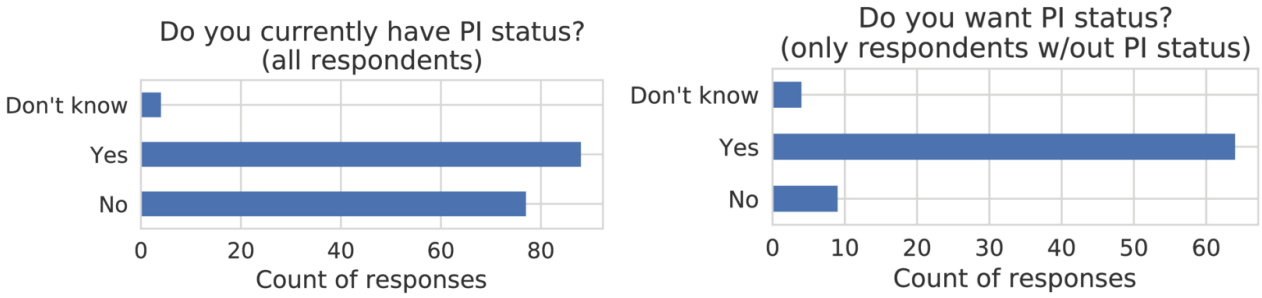
4 Resources and structures that academic data scientists need

4.1 Summary of findings

- Out of all respondents without Principal Investigator status, an overwhelming majority (82.7%) want PI status as part of their academic position. The ability to do one's own research is very important to almost all of our respondents. This suggests that the limitation, in some places, of PI status to ladder-rank faculty may discourage the development of some groups of data scientists.
- There are several resources that were uniformly highly valued by our respondents, and universities should make these available to those in various career paths across data science:
  - Computational resources (e.g. servers, cloud computing)
  - Digital library collections (e.g. publication databases)
  - Funds for research (including conference travel)
  - Ability & funding to hire students and/or postdocs to assist in research
  - Administrative staff
- Respondents also listed a wide variety of formal and informal structures that have been useful for their careers in academic data science, including: official and unofficial mentors, informal peer learning or discussion groups, cross-disciplinary institutes of data science, discipline-specific and cross-disciplinary training and education initiatives, and a wide variety of grant funding.

4.2 Principal Investigator status

We asked respondents if they currently had principal investigator status (specifying that those with PI status "can independently apply for and receive research grants from government agencies or foundations"). Eighty-eight respondents (52.1%) had PI status, 77 (45.6%) did not, and 4 (2.4%) did not know. We then asked if respondents "in your ideal academic position, would you be applying for and receiving research grants from government agencies and/or foundations as a principal investigator or co-PI?" Out of the 81 respondents who did not currently have PI status (including those who did not know if they did), 67 respondents (82.7%) would want PI status, 9 (11.1%) would not want PI status, and 5 (6.2%) did not know.

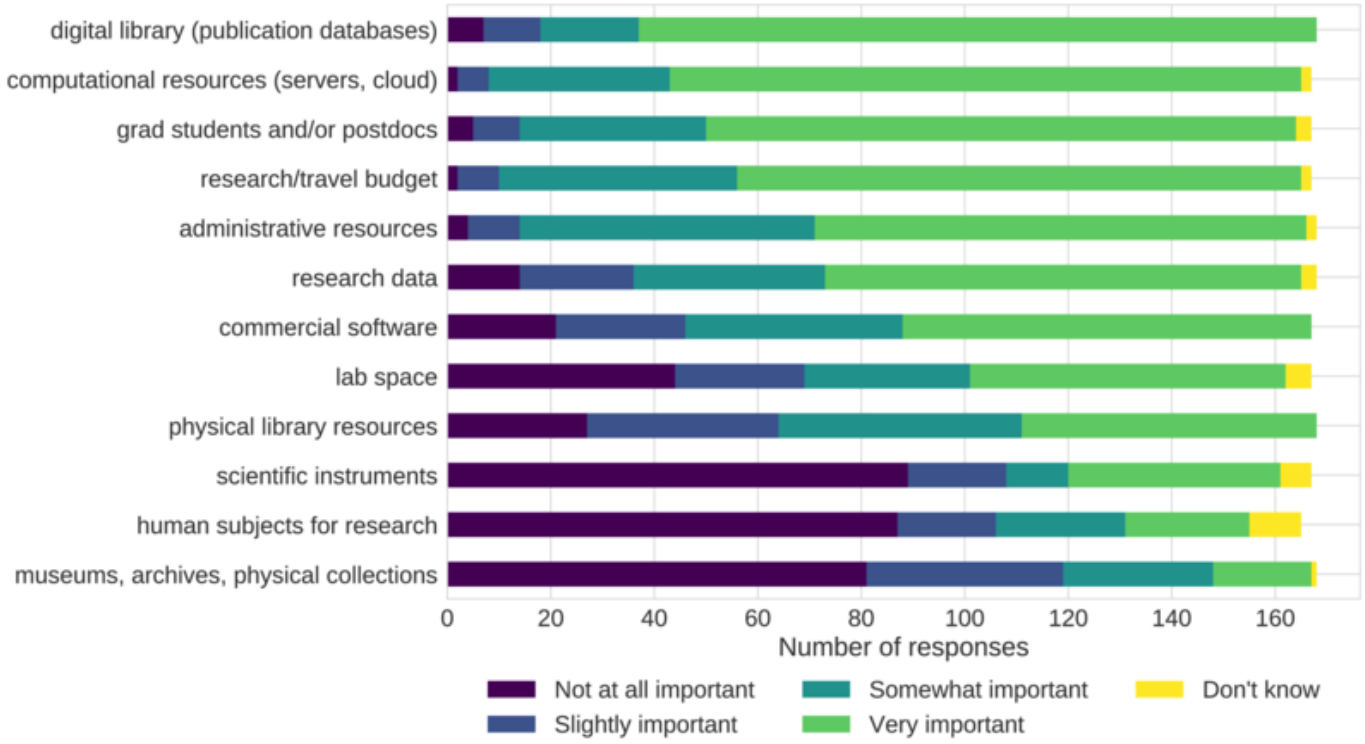


(a) **Current PI status** (b) **Ideal PI status**

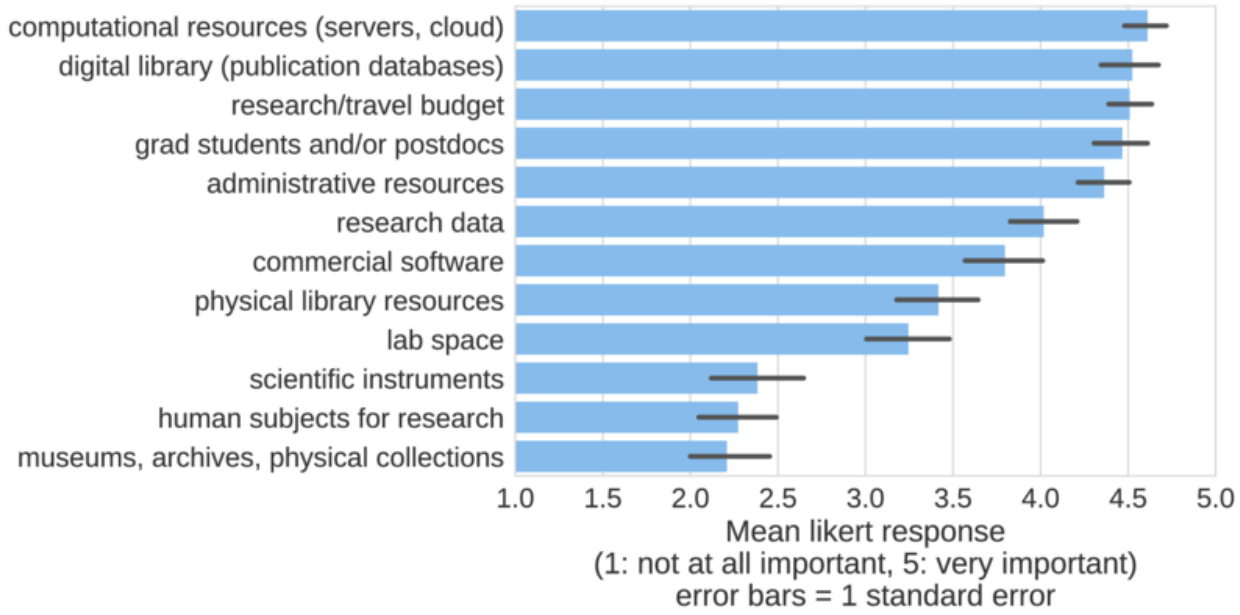
Figure 7: **Questions about principal investigator status**

4.3 Resources provided by universities

We gave respondents a list of 12 resources, asking how important it would be that "your employer provide the following to you?" Respondents were given a five-point Likert scale ranging from "not at all important" to "very important," with an "I don't know" field at the end of the scale. Figure 8a shows the distribution of responses for each resource and Figure 8b plots the mean response for each activity, with responses coded as 5 for "very important," 1 for "not at all important," and "don't know" coded as 3. The resources with the largest proportions of "very important" or "somewhat important" ratings were: computational resources like servers or cloud computing (94.0%), a travel/research budget (92.8%), administrative resources (90.5%), ability & funding to hire grad students and/or postdocs (89.8%), and digital library resources / publication databases (89.3%). On average, the lowest values were to discipline or field specific resources, but even these had larger proportions of "very important" or "somewhat important" ratings than their mean responses may indicate: museums, archives, physical collections (28.6%), human subjects for research (29.7%) and scientific instruments (31.7%).



(a) **Stacked barplot showing the distribution of responses for resource importance.**



(b) **Mean resource importance for all respondents.**

Figure 8: **Resource importance for all respondents**

4.4 What structures helped in your career?

Seventy-five out of 169 respondents (44.4%) gave a response to the question "Please list any institutional or organizational structures (e.g., mentoring networks, advising, grants, training opportunities, brown bags) that have helped support you in pursuing your ideal career path." Responses included a rich set of resources at various scales, spanning interpersonal relationships, local departments and research institutes, university-specific and cross-university programs and resources, discipline and field-specific programs and events, interdisciplinary collaborations and events, and national and international grant agencies. The mean length was 23 words, the median length was 10 words.

We analyzed these free response questions first by reading them by hand to identify common thematic elements. We then proceeded computationally, with Python functions that allow us to find the proportion of statements that contain at least one term from a set of terms. Our function is case insensitive and allows us to match on partial terms, such as having "thon" match "hackathon" or "datathon" This is useful in helping identify what proportion of respondents who gave an answer included a set of terms.

Of those who provided a response, 34.7% included some kind of term referring to advisors and mentors.[3] 34.7% included terms referring to meetings, conferences, and networking events.[4] 30.7% included terms that referred to ad-hoc or interdisciplinary training and education events or centers.[5] Finally, 44% of included terms referring to grants, funding, and foundations.[6]

We place particular importance on providing a wide range of mentoring, networking, and other types of events where graduate students and early career researchers can receive career advice. A 2017 *Nature* survey of graduate students across fields found that "PhD students are largely finding their own career advice online. Just one-third credited advice from a supervisor as a reason for their career choice" and only 20% credited career-specific training events or seminars. (Woolston 2017) While we have generally focused on academic career paths in data science in this project, we also stress that it is important to give graduate students and early career researchers opportunities to learn about and compare a broad range of career options, both inside and outside academia.

5 Career goals and priorities

5.1 Summary of findings

- In terms of career priorities, financial compensation is an important factor, but it is far from the only issue that respondents consider. On average, respondents in our sample:
  - place a higher importance on having secure, long-term employment, intellectual freedom, and "being around smart people".[7]
  - place less importance on having a lifetime appointment and high influence in the university.
- For many early career data scientists, existing academic career paths and structures do not generally provide much confidence and stability for doing the kind of work they want to do.

[3] Terms: ['mentor', 'profes', 'facul', 'advisor', 'adviser', 'coach']

[4] Terms: ['conf', 'network', 'event', 'brown bag', 'fair', 'summer', 'meetup', 'meeting', 'workshop', 'thon', 'hack']

[5] Terms: ['bootcamp', 'boot camp', 'workshop', 'carpentry', 'igert', 'hacker within', 'thw', 'training', 'dlab', 'd-lab']

[6] Terms: ['grant', 'fund', 'fellowship', 'mellon', 'nih', 'nsf', 'foundation', 'sloan', 'moore', 'award', 'big data hub', 'dse', 'national science foundation']

[7] We phrased "being around smart people" to capture a sentiment we have heard frequently expressed in our interviews with data scientists.

  - A majority of our respondents are strongly considering a position outside academia.
  - About two-thirds would not be comfortable with taking either an adjunct faculty position or a "soft money" position that is entirely dependent on grant funding.
  - Fewer than half of those in non-ladder rank positions believe that universities typically grant tenure to top candidates that do the kind of work that they do.
  - An overwhelming majority do not want a tenured position if it would mean they would not be able to do the kind of work they want to do.

- We identify two related but distinct factors: first, wanting to stay in academia versus take a position outside academia; and second, wanting a tenured faculty position versus a non-tenurable position.
  - Using a factor analysis, we plot non-tenure respondents on this two-axis grid. Our respondents fall into all four quadrants, indicating a broad range of situations, experiences, and priorities.
  - There are early career data scientists who want to stay in academia and believe that tenure is both beneficial and likely for them. There are also many data scientists who want to stay in academia but do not think tenure is beneficial and likely, as well as those who do thing tenure is beneficial and likely but do not want to stay in academia.
  - We place particular emphasis on doing further research and outreach to support the quadrant of those who want to stay in academia, but do not believe that tenure is worth it (or likely) in their case.

5.2 Descriptive statistics

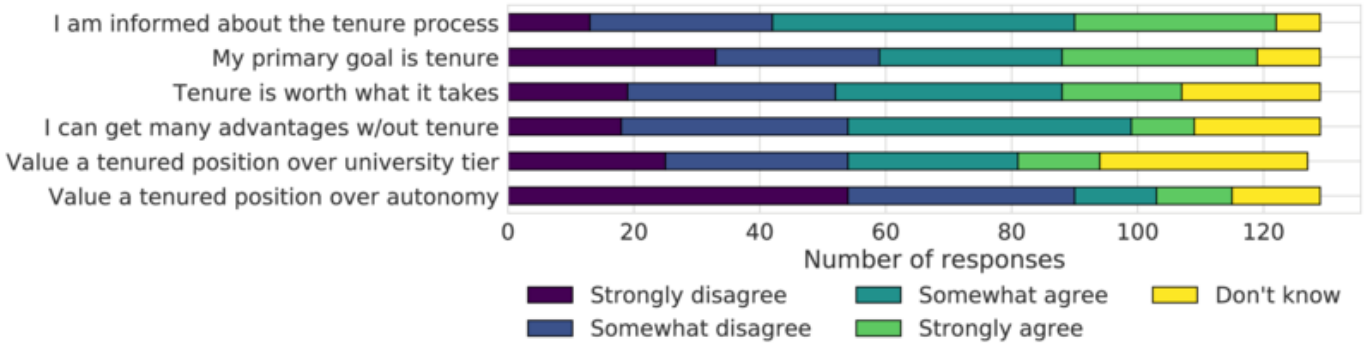
5.2.1 Career goals

We asked 15 questions about future career goals: 9 questions about general career goals, then 6 questions about tenure asked only to non-ladder-rank faculty. Figure 9a shows the distribution of responses for these questions on career goals for all respondents except tenured and tenure-track faculty. In general, these statements had wider distributions than other similar Likert questions: all but one question had a standard deviation greater than 1.25 (*tenure_not_worth_it*). The greatest variance was on the statements about whether their primary career goal was tenure ($\bar{x} = 2.99, s = 1.56$) and whether they were strongly considering a non-academic position ($\bar{x} = 3.14, s = 1.58$).

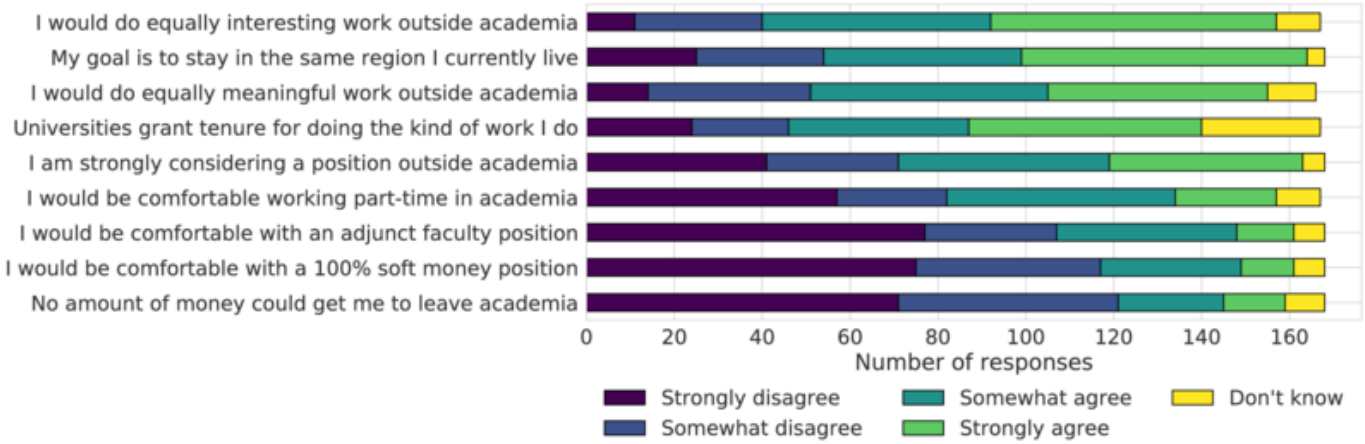
Respondents expressed the strongest disagreement with the statements that: no amount of money would get them to leave academia($\bar{x} = 2.17, s = 1.33$); that they would would pick tenure over being able to do the work they wanted to do ($\bar{x} = 2.17, s = 1.32$); that they would be comfortable with a 100% soft money position ($\bar{x} = 2.19, s = 1.37$); and that they would be comfortable with an adjunct faculty position ($\bar{x} = 2.30, s = 1.45$). Respondents expressed the strongest agreement with the statements that: they would do equally interesting work in a non-academic position ($\bar{x} = 3.78, s = 1.30$); they would do equally meaningful work in a non-academic position ($\bar{x} = 3.54, s = 1.35$); and they would strongly prefer to stay in the same region ($\bar{x} = 3.57, s = 1.51$). We interpret these results to mean that respondents generally wanted to do work that was meaningful and rewarding, and find academia to be a home for that kind of work — but they worried about life in a soft-money or adjunct position. Furthermore, most could be induced to leave academia with enough money, because they believe they could find interesting and meaningful work in a non-academic position.

5.2.2 Career priorities

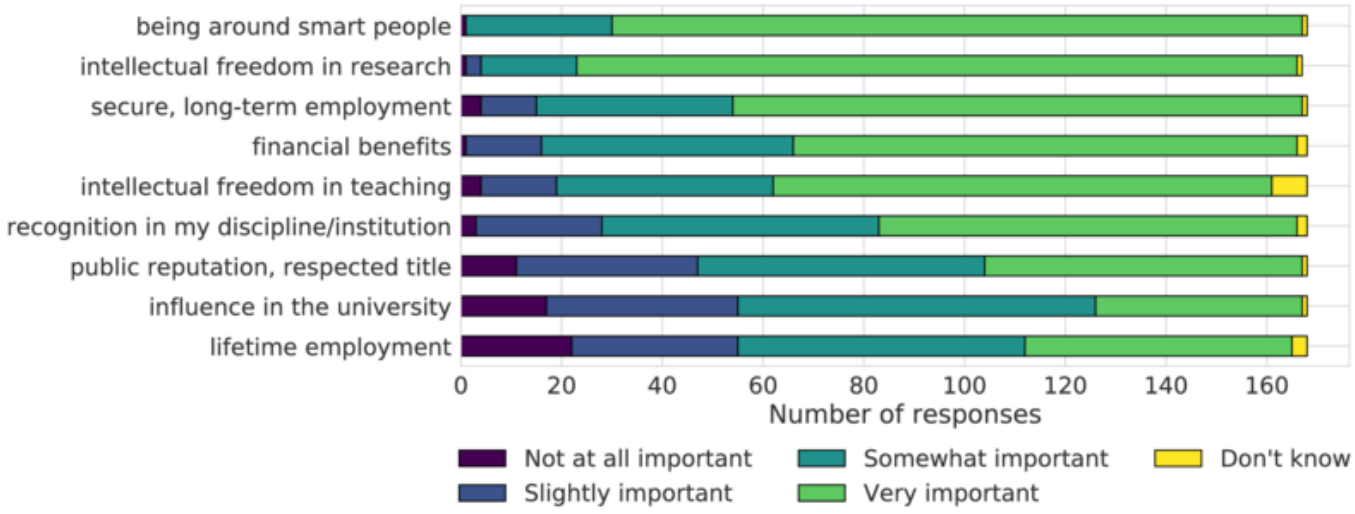
We also asked 9 questions about priorities, rating how important various criteria were in their ideal academic position. Figure 9c shows the distribution of respondent values. Some of the most universally important priorities were intellectual freedom in research ($\bar{x} = 4.80, s = 0.60$), being around smart people ($\bar{x} = 4.79, s = 0.50$), and long-term, secure employment ($\bar{x} = 4.46, s = 0.97$). The lowest priorities on average were influence in the direction of the university (($\bar{x} = 3.48, s = 1.34$) and lifetime employment ($\bar{x} = 3.51, s = 1.44$), although these also had the highest standard deviations. There are a subset of respondents who highly value these core components of tenure, but they are in general less important than other factors. We particularly call attention to the strong difference between the value of lifetime employment versus long-term, secure employment, which indicates that there are a substantial proportion of junior academics in our sample who do not necessarily need lifetime employment of tenure-track positions, but need more long-term security than post-PhD academic positions typically provide.



(a) **Career goals for tenure (asked only to non ladder-rank respondents**



(b) **Career goals (asked to all respondents)**

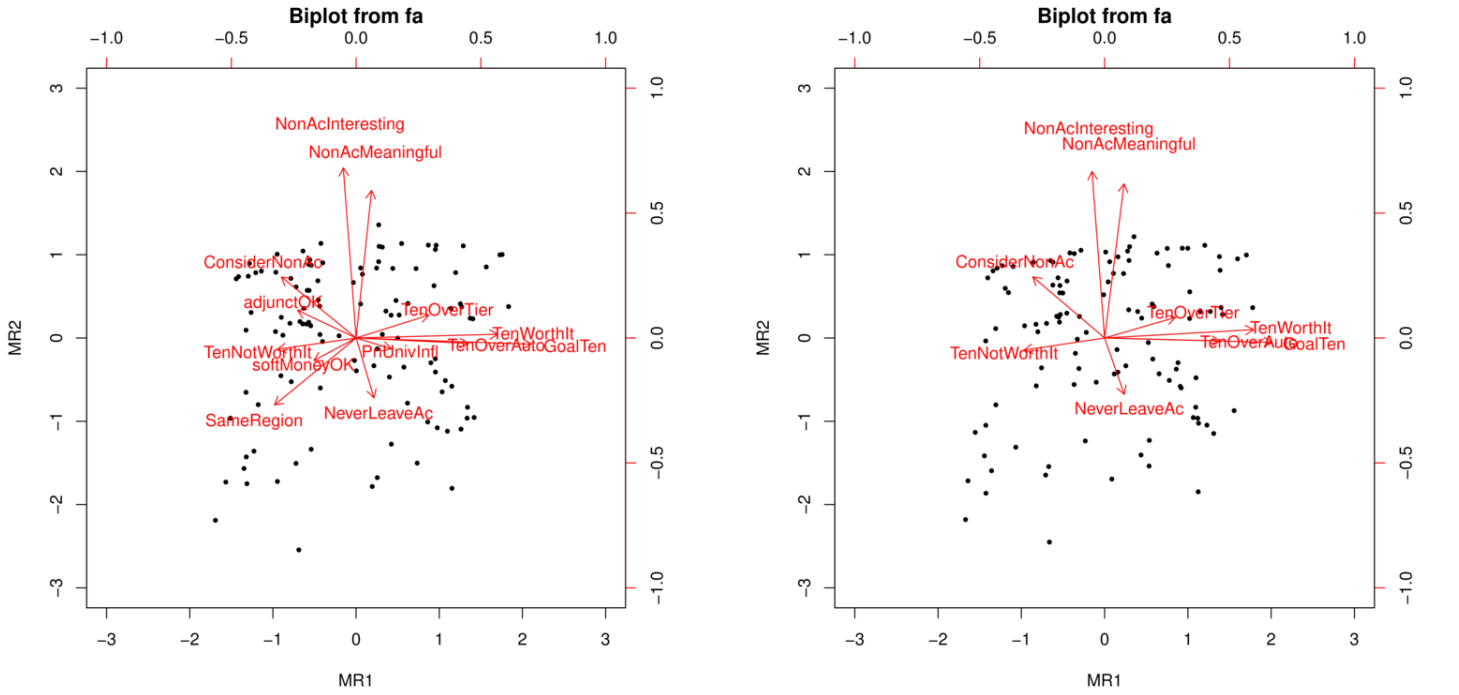


(c) **Career priorities (asked to all respondents)**

Figure 9: **Stacked bar plot of responses for career goals and priorities, showing distribution of values**

5.3 Factor analysis for career goals and priorities

To explore the clustering of respondents' career goals and priorities, we ran a factor analysis on a large number of goals and priorities, using R (R Core Team 2016) and the R psych (Revelle 2017) package. The two factors explained 31% of the variance in the set. We found two distinct dimensions (Figure 10a). The positive factors on the X axis are generally associated with valuing tenure, those negative on the X axis were associated with not valuing tenure. The positive factors on the Y axis are associated with going outside academia, while the negative factors on the Y axis are generally associated with staying in academia. However, there is much variance and overlap with this many factors, and so we ran a second factor analysis with fewer variables (Figure 10b). We excluded variables like wanting to stay in the same region that had more variance across these factors and are likely capturing issues that are orthogonal to these two factors. In this more focused factor analysis, the two factors explained 39% of the variance in the set. We then used these factor scores to cluster respondents into four groups:



(a) **Exploratory factor analysis with many goals and priorities**

(b) **Factor analysis with fewer goals and priorities, used for clustering respondents into groups**

Figure 10: **Factor analyses of goals and priorities, with variable loading scores in red arrows and regression scores for respondents in points.**

On the X axis, the factors positively associated with valuing tenure were:

- **GoalTen**: My primary career goal is to be a tenured professor
- **TenureWorthIt**: I believe that the advantages of tenure are worth what it takes to get tenure.
- **TenureOverAuto**: If I had to choose between taking a tenure-track position and being able

to do the kind of work I want to do, I would pick the tenure-track position (autonomy).

- **TenureOverTier**: If I had to choose between a tenure-track position at a lower-tier university and a non-tenurable position at a top-tier university, I would pick the tenure-track position at a lower-tier university.

On the X axis, the factors negatively associated with valuing tenure were:

- **TenNotWorthIt**: I believe I can get many of the advantages of tenure without going through the tenuring process.
- **ConsiderNonAc**: I am strongly considering a career path outside of academia. (this factor is also positively associated with the y axis, i.e. going outside of academia)

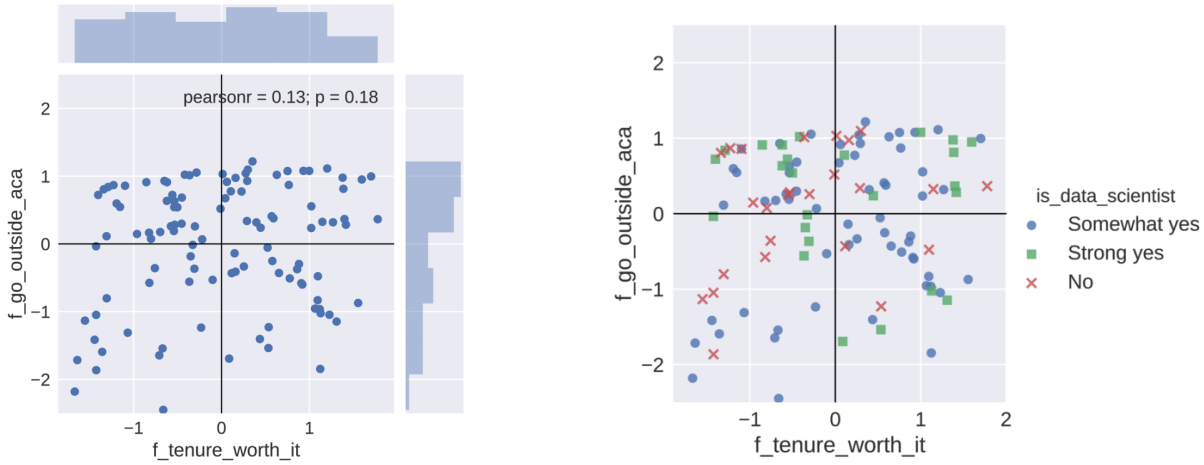
On the Y axis, the factors positively associated with going outside of academia were:

- **NonAcInteresting**: If I took a job outside of academia, I would do equally interesting and engaging work.
- **NonAcMeaningful**: If I took a job outside of academia, I would do equally meaningful work.
- **ConsiderNonAc**: I am strongly considering a career path outside of academia.

On the Y axis, the factors negatively associated with going outside of academia were:

- **NeverLeaveAc**: There is no salary or benefit package that could induce me to leave academia.

Plotting each non-ladder rank respondent by their regression score for the two factors shows a wide distribution across all four quadrants. The stacked bars across each axis in Figure 11a are histograms, showing that the factor for tenure worth it is relatively evenly distributed, while the factor for going outside academia is more concentrated in the positive direction.



(a) **Joint scatter and histogram plot of factor scores**

(b) **Scatterplot of factor scores, segmented by self-identification as a data scientist**

Figure 11: **Scatterplot of factor scores (Y axis: go outside academia versus stay in academia; X axis: tenure worth it versus tenure not worth it) for non-ladder rank respondents**

6 Portraits of academic data scientists facing challenges in career paths

In coordination with the ethnographic and interview-based research into academic data science — both at the three institutes surveyed and beyond — we created fictional composite portraits to illustrate how this two-axis mapping is useful for understanding the challenges that many academics in data science are facing. These portraits do not represent real individuals, but they are synthesized from real situations, issues, and concerns of many individuals in academic data science. These portraits represent the extremes of our grid and are intended to be useful for thinking about different kinds of problems and solutions.

- **Un**a: the **un**decided graduate student, who does not know what career paths looks like in or out of academia and how they might fit into them.
- **Inte**us: the **inter**disciplinary postdoc researcher, who publishes and teaches across many fields. Inteus wants to remain in academia in a tenure-track position, but she finds it difficult for a single discipline to recognize her work.
- **Ser**gei: the postdoc who provides much **ser**vice to others, through open source software development and research consulting for academic labs. Sergei is certain he wants to remain in academia, but doesn't know if he wants a position on or off the tenure track.
- **St**eph: the **st**aff researcher who has long been funded through short-term "soft money" positions. Steph is certain they don't want a tenure-track position and would like to remain in academia, but they want more stability, leadership roles, and a respected title.
- **Na**omi: the Ph.D student with **n**on-**a**cademic goals, who knows she wants an industry position. Naomi would love to build bridges between academia and industry, but she finds herself alienated by students and faculty when she seeks advice for non-academic positions.
- **Const**ance: the postdoc who wants a tenure-track faculty position, but has various issues and **const**raints that make it difficult for her to stay in academia, so she is seeing other careers.

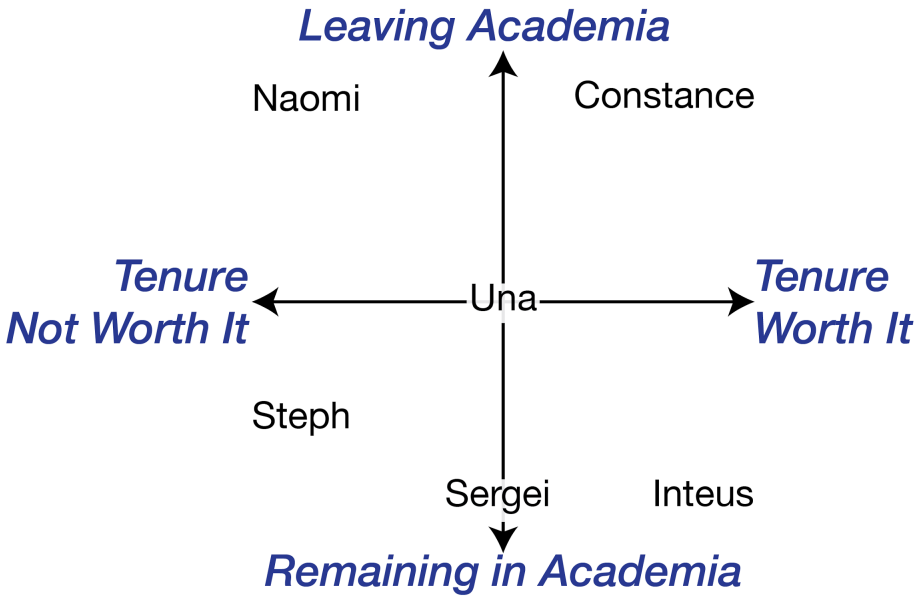


Figure 12: **Placing composite portraits on the two factor axes**

6.1 Inteus: the interdisciplinary researcher

Inteus' is a second-year postdoc at an academic data science institute, currently funded by a cross-domain research grant. She received her Ph.D from a social science discipline, but constantly collaborates not just across the social sciences, but all around academia. Her research — which leverages new computational techniques to analyze and visualize data at scale — has always been interdisciplinary, although she did not set out to be that way. Because of the scientific questions she was interested in, she always had to look beyond her own disciplinary silo to take her research to the next level. She also had to acquire computational tools and methods that are not traditionally associated with her discipline. She faced the amused indulgence, skepticism, and sometimes harsh criticism of her colleagues... until she found her way to an academic data science institute. Her efforts have paid off, and her publications record should make her a perfect tenure candidate for any forward looking academic institution. Yet her first year on the job market has proven otherwise. Because she has published across the social sciences, computer science, and statistics, her work hasn't been recognized by hiring committees in her discipline. Her contributions to open source software packages have also been completely ignored. To get a job next year, she figures she will have to play it safe and refocus her application to questions and frameworks traditionally associated with her discipline. Once she lands a job, she hopes she can convince her colleagues and tenure committee that her data science expertise is a great asset for her department and students.

6.2 Sergei: provides service to other academics

Sergei is a fourth-year postdoc in a physical science lab. While his Ph.D was in the physical sciences, he also gained many skills in the computational analysis of large-scale data pipelines. He also became deeply involved in open source scientific software projects now used across academia and industry. He could easily get a high-paying job in industry, but he knows that he wants to stay at a university — intellectual freedom is the most important value for him, and his spouse is adjunct faculty on the other side of campus. If you asked his advisor, she'd tell you that Sergei could have been a great candidate for the academic job market. He comes from a reputable lab, worked on innovative research questions, and made their lab and others far more efficient through his software. People around the world ask for his advice on what kinds of hardware and software to use for their work, and he routinely speaks at data science conferences. Yet this service work is becoming a source of tension between Sergei and his advisor. His advisor does sees the benefit of having a dedicated, computationally-savvy person like Sergei in the lab. But she worries that all this service work is taking him away from what she thinks should be his primary focus: publishing academic papers in their field. Sergei's advisor would love for him to have a career in academia, but it has been difficult to either find him tenure-track position or keep him funded in her lab. The current grant that funds Sergei's position (and thus indirectly supports labs around the world) is running out. Especially with a growing family, he is seeking a more stable, long-term position.

6.3 Steph: the staff researcher

Steph is a staff researcher who received their Ph.D over ten years ago in computer science, but always knew they did not want to go on the tenure track. Steph has worked on a variety of research projects around machine learning in both academia and industry, and they are currently funded 50% by a short-term grant for a life science lab and 50% by an academic data science institute. Steph cares about working on interesting and diverse projects they care about, solving seemingly impossible technical challenges, being surrounded by smart, dedicated people, and walking their dog twice a day. Until recently, Steph really enjoyed the flexibility that these short-term positions

gave them, which were a good way to work on the project-based approach that they're especially good at. And because they embrace new trends and technology long before everyone, they never really had to look long for a new job when funding or a contract ended. But after a decade, Steph is finding they have more to contribute to academia than just their machine learning expertise, as they find themselves informally mentoring, advising, and managing across the university. They are becoming more involved in shaping the direction of the institution around data science, and for the first time, Steph feels like they're in an institution they could call home. Yet because of how career paths and human resources work at Steph's university, their official position is a yearly-renewed contract with the title "assistant researcher" – something they say with a laugh every time they introduce themselves in meetings full of faculty and university leadership. Meanwhile, they are constantly headhunted by local industries.

6.4 Naomi: non-academic goals

Naomi is a fourth-year Ph.D candidate in statistics, with hopefully just one more year to go before graduating. Currently, she is funded by a data science fellowship that lets her spend time in an academic institute of data science while working on her dissertation. She often brings fundamental principles of statistics into conversations with the domain researchers who are using statistics to tackle a wide variety of problems. When she started graduate school, Naomi believed that high-quality, reproducible, statistically-sound science in all domains can improve people's lives and make the world better. She still lives by this statement, although she is not sure that working in academia is where she can have the most impact. She has also done an industry internship and found that she prefers the culture of industry to academia. Yet even though she has decided she's not going to pursue an academic career, Naomi plans to complete her PhD. But ever since she publicly discussed her decision to leave academia, she feels that people are distancing themselves from her. She has also found it extremely difficult to find useful resources on campus to make her application stronger and get the necessary professional training that she might need. In order to make her transition easier she would have liked to have more contact with alumni, or representatives of private or public sectors that have been known to benefit from hiring PhDs. In the future, she hopes to be able to build more bridges between academia and the rest of the world.

6.5 Constance: Academics with various constraints

Constance is in the quadrant of those whose goals strongly align with the duties and roles of tenured faculty, but are nevertheless are strongly seeking to leave academia. Constance's portrait is the most difficult to generalize, as those in this quadrant face a wide range of issues and concerns. For example, Constance could have a strong personal geographic constraint, not being able to relocate to a university that may offer her an otherwise ideal tenure-track faculty position. Constance could be a foreign Ph.D student facing issues over visas and residency, finding that companies are providing more certainty to her around these issues to her than universities are. Constance may also be a post-doc whose research focuses on a specialized sub-field, where new tenure-track positions rarely become available and where data science expertise is not generally valued. Given what extensive research has shown on gender equality in academia, Constance could be a graduate student looking for women faculty with successful work-life balances as role models — and finding that her otherwise ideal academic institutions provide little to no support around maternity leave, childcare, and sexual harassment. Or she could be tenure-track faculty who deeply enjoys the individual responsibilities of her position, but is facing deep structural issues in trying to break the glass ceiling and gain respect from her colleagues. Finally, Constance may be finding that univer-

sities will give her a tenure-track position she would enjoy, but that in order to do the research she wants to do, she needs access to cutting-edge computational resources and/or data sources that only a handful of major tech companies have.

6.6 Una: The undecided first-year Ph.D student

Our final portrait is Una, the only one in the direct center of the grid, representing a deep uncertainty about both staying in academia and wanting a tenured faculty position. As a first-year graduate student, Una knows that she has plenty of time to make her decision, and is broadly seeking information about what academia is like — which she is already realizing is quite different than the impression she got from being an undergraduate. However, Una's challenge is that she is simultaneously overwhelmed by all the potential career paths, yet finding herself uninformed about what work and life would actually be like as a postdoc, as research staff, as a tenure-track faculty member, as an industry researcher, or any number of positions she might take 4-6 years from now. She has had a few ad-hoc conversations about careers with more senior grad students in her lab and her advisor, who mostly tell her not to worry about those issues yet — a deeply unsatisfying response for a goal-oriented person like Una. To support the many people in Una's position, it is important to have many opportunities for professional development and career guidance, giving early career researchers the information they need to make informed choices.

7 Perceived value of various activities for tenure

7.1 Summary of findings

- We identify four tiers of activities, based on how beneficial or detrimental respondents, on average, believe they are for tenure in their field:
    1. Seen as very important for tenure: core research contributions to one's primary field (e.g. publishing papers in one's field, writing grants)
    2. Seen as somewhat important for tenure: non-research faculty roles (e.g. advising, academic service, teaching courses) and publishing outside one's primary field
    3. Seen as neither beneficial nor detrimental for tenure: activities associated with data science (e.g. developing computational resources, doing open/reproducible research, teaching workshops, research consulting)
    4. Seen as actively detrimental for tenure (e.g. taking courses, having/raising children, and "the rest of life")
- On average, non-ladder rank researchers personally place a higher value on activities associated with data science more than they believe tenure committees in their field do, such as doing open/reproducible research and developing computational resources.

7.2 What do respondents think tenure committees value?

We asked all respondents to rate the 17 activities based on how beneficial or detrimental they were to tenure in their field. We also included additional qualities like receiving awards, and distinguished between publishing in and outside one's primary field. Respondents generally expressed similar responses across career stages, with all responses having a standard deviation below 1. The

only statistically significant difference between career stages was that ladder rank faculty on average did not rank "the rest of life" to be as detrimental to tenure as non-faculty by approximately half a point (score = .52, p=0.002).

Figure 13 plots the mean response for all activities, ordered from most to least beneficial for tenure. We found that the ordering reflected an interesting segmentation, which we have indicated with the horizontal lines in figure 13. The top rated activities are expected determinants of tenure: publishing in one's primary field, writing grants, and receiving awards. Secondary to this group were activities that are typically considered core faculty responsibilities, like attending conferences, managing a group/lab, teaching courses, publishing outside of one's primary field, and traditional academic service. Below academic service, we found activities that are often considered to be a different kind of service work prevalent in data science, including practicing open/reproducible research, developing computational resources, teaching workshops, and research consulting. Below this were taking courses, having and raising children, then "the rest of life," which received the lowest average scores.

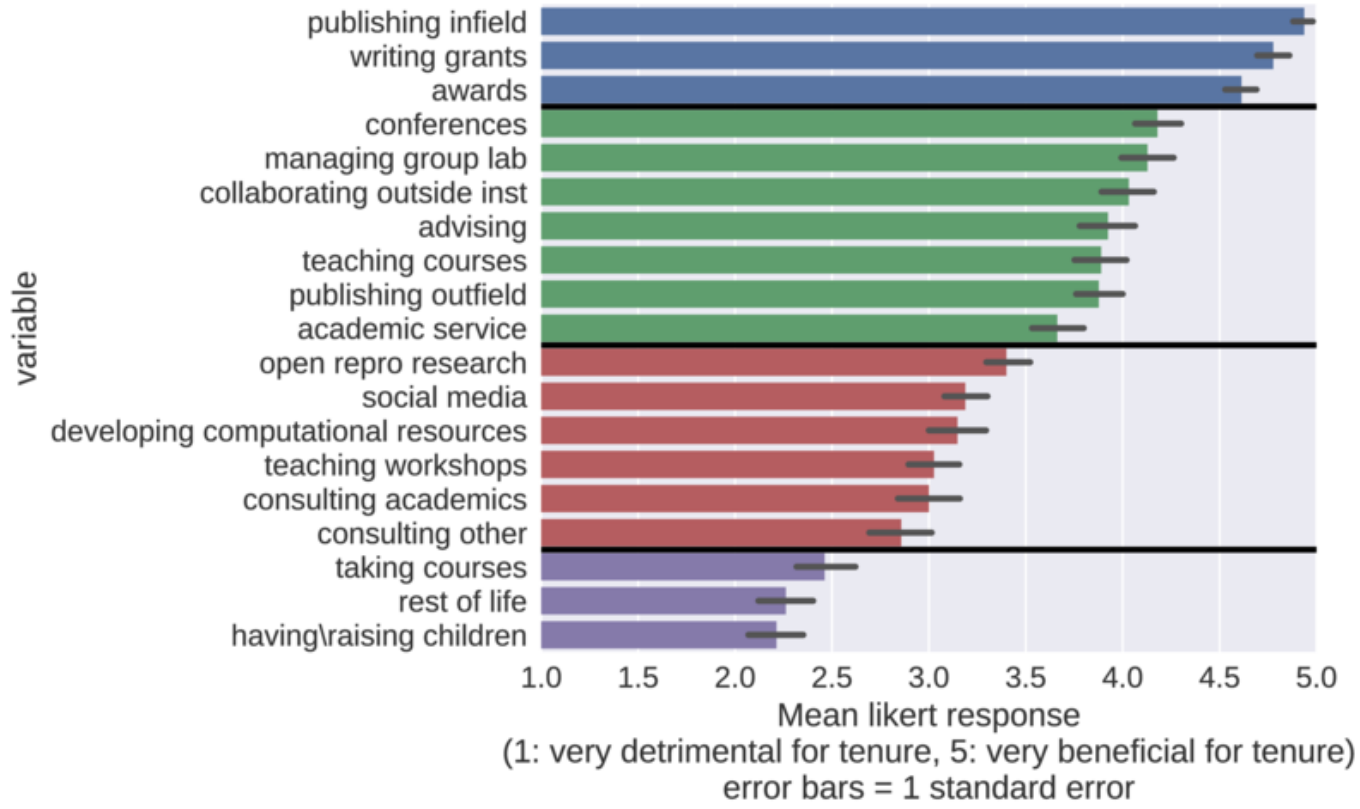


Figure 13: **Mean likert response for how detrimental or beneficial they believe each activity is for tenure.**

7.3 Tenure gap

We also calculated the difference between the ideal value that respondents' placed on an activity from how beneficial or detrimental they believe it is for tenure. This provides a 'tenure gap' rating for each activity, plotted in Figure 14, with a negative value indicating that the respondent values this more than they believe tenure committees do and a positive value indicating that the

respondent value this less than tenure committees do. This figure plots the mean tenure gap scores for all activities for non-ladder rank researchers. The largest negative values are "the rest of life" and having and raising children. This was followed by doing open and reproducible research and developing computational resources, which respondents generally personally value more than they believe tenure committees do. The activity with the highest tenure gap score was writing grants: respondents overwhelmingly believe that writing grants is important for tenure but do not personally value it as much.

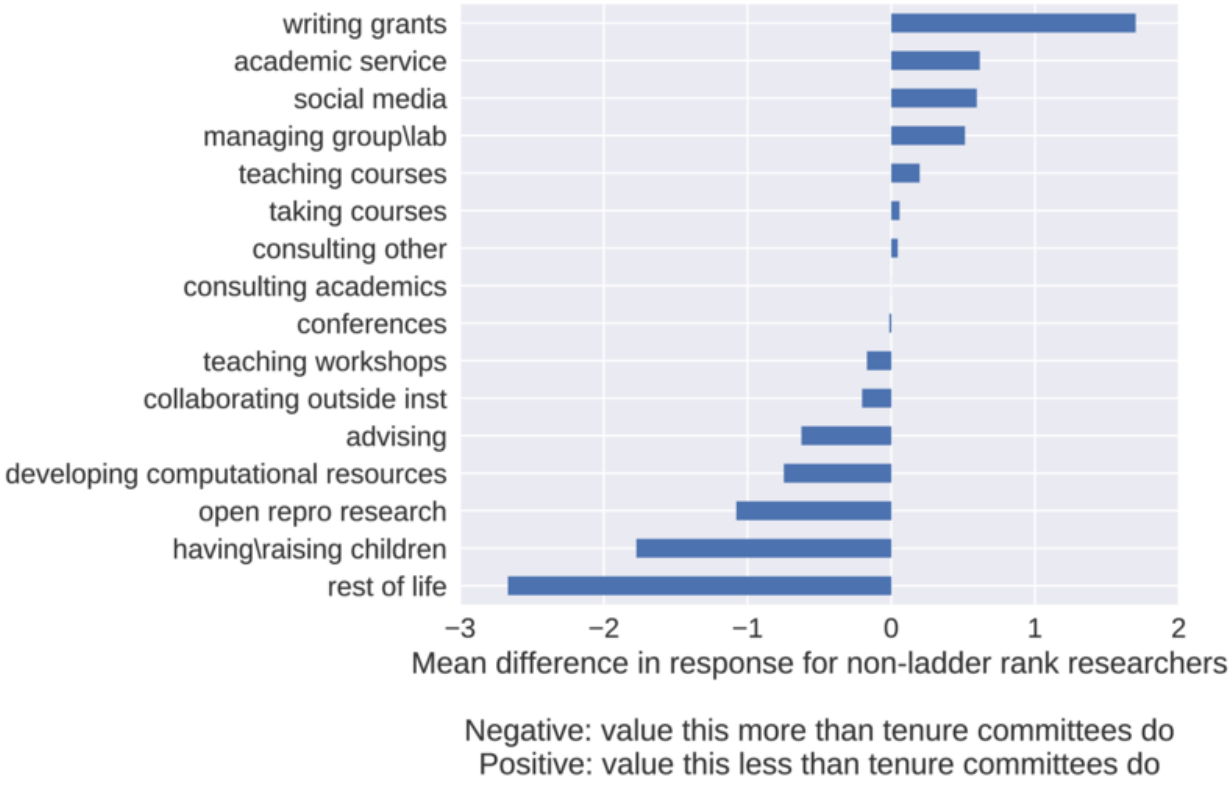


Figure 14: **Mean difference between how much respondents' value an activity and how detrimental or beneficial they believe it is for tenure.**

8 Goals, priorities, and career satisfaction by demographics

8.1 Gender

We analyzed the goals and priorities questions by gender, and Figure 15 plots the mean response for various selected questions on career goals and priorities. For conducting comparisons between groups, we only examined men and women; we did have other genders in our responses, but not in high enough numbers to conduct statistical comparisons. In terms of priorities, men and women expressed equivalent levels of importance for all questions except two. On average across all career groups, women expressed statistically significantly higher importance on positions with "influence in the direction of the university" (score=0.61, p=0.006). Women in non-ladder rank research positions (Ph.D students, postdocs, and research staff) also expressed higher average importance for positions with "public reputation / respected professional title" (score=0.64, p=0.043).

We did not find any statistically significant gender differences in the goals or career satisfaction questions, either in the entire sample population or only in non-ladder rank researchers.

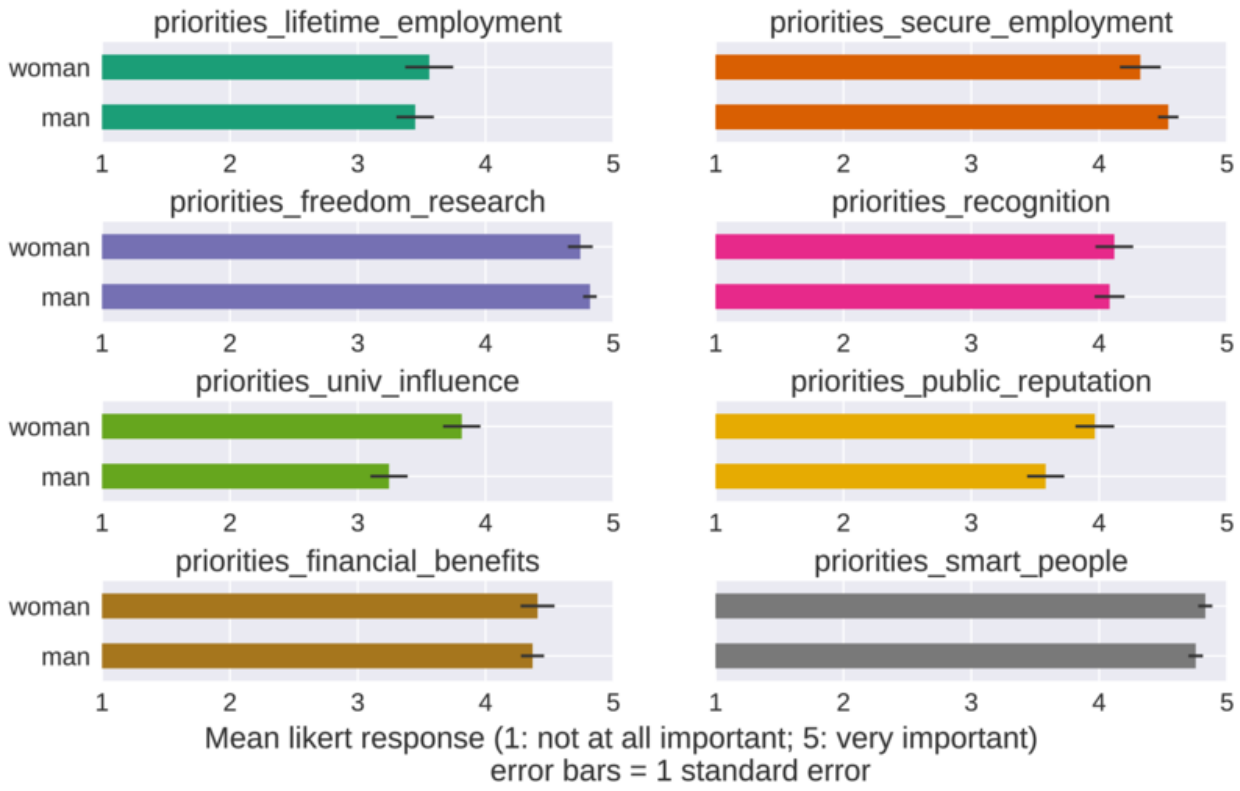


Figure 15: **Comparison of mean response values between men and women in non-ladder rank research positions.**

8.2 Race/Ethnicity

We also analyzed goals and priorities by race/ethnicity, using the binary variable of whether the respondent was a member of a minority (non-white) race/ethnicity or not, as we did not have enough numbers to run statistical comparisons between different groups. In terms of priorities, we found that non-ladder rank researchers from minority backgrounds placed a higher value on positions with a "public reputation / respected professional title" (score = 0.75, p = 0.032). We did not find statistically significant differences in priorities across all career stages, possibly in part due to the small numbers of minorities in faculty positions. In terms of goals, there were no statistically significant differences between minority and majority race/ethnicity backgrounds, except that among non-ladder rank researchers, respondents from minority backgrounds reported more agreement with the statement that "Universities typically grant tenure to top candidates for doing the kind of work I do" (in linear regression; score=0.70, p=0.039). We also did not find any statistically significant differences in reported career satisfaction, either in the entire sample population or only in non-ladder rank researchers.

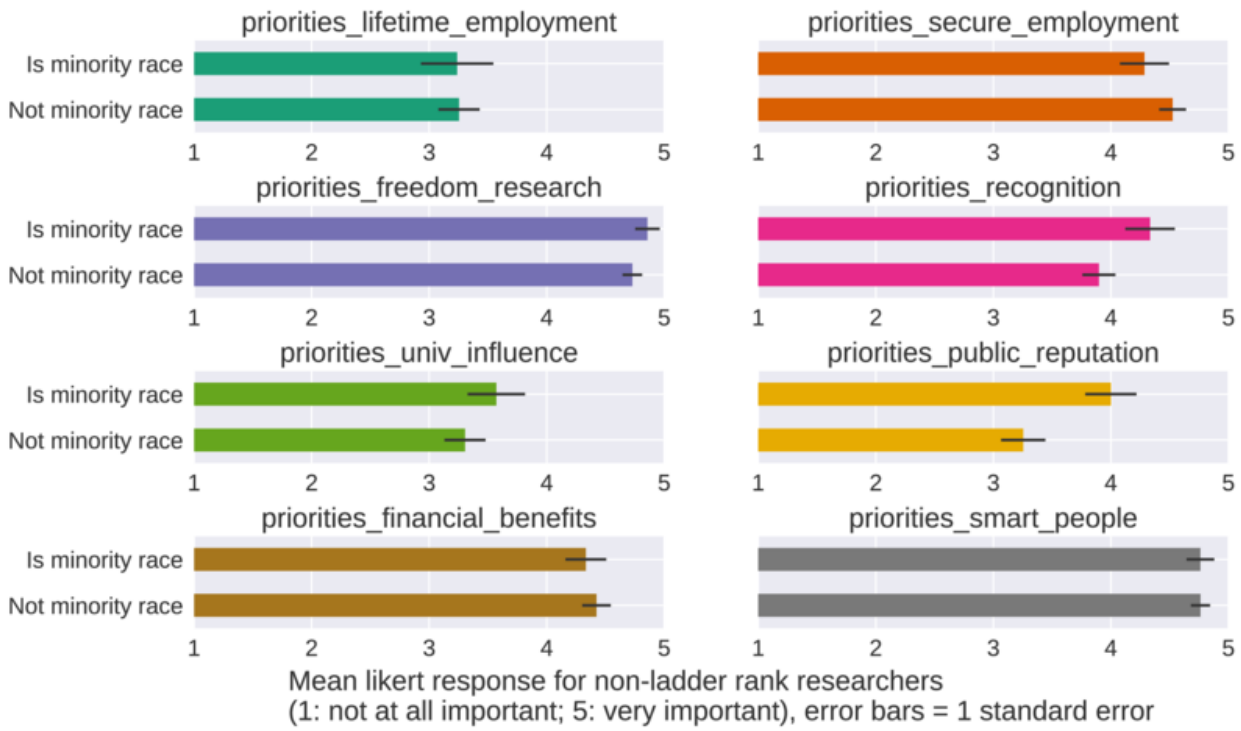


Figure 16: **Comparison of mean response values between minority and non-minority racial/ethnic backgrounds in non-ladder rank research positions.**

Figure 17: **Career satisfaction by career stage.**

8.3 Career stage and age

The most important variables we found for predicting career satisfaction were career status and age. Income career satisfaction had the highest disparity; in a linear regression, we found indepen-

dent statistical significance for both age (score=0.02, p=0.033) and ladder-rank status (score=0.62, p=0.012). This means that in the regression, every year older is associated with 0.02 points higher in income satisfaction, plus ladder-rank status is associated with 0.62 points higher in income satisfaction. This was also the case for satisfaction in publishing (age: score=0.026, p=0.004; is_ladder_rank: score=0.62, p=0.007). However, other satisfaction variables were exclusively associated with ladder-rank status and not age, including: overall career satisfaction (is_ladder_rank: score=0.58, p=0.002), satisfaction in teaching (is_ladder_rank: score=0.89, $p < 0.0001$), career advancement (is_ladder_rank: score=0.82, $p < 0.0001$), and skills learned (is_ladder_rank: score=0.54, $p < 0.009$).

8.4 On interpreting multiple comparisons

Like with section 3.4, our study was more exploratory and did not explicitly test a narrow set of hypotheses. We therefore advise against interpreting these regressions to strictly generalize and infer about broader populations. Given the large number of different regressions and multiple comparisons performed on many variables for gender, race/ethnicity, and age, the relatively small proportion of statistically significant findings could be more a result of random chance. At present, we cannot confidently infer that these demographic findings apply to a broader population of academic data scientists, but we present the scores and p-values to guide future work. However, these correlations do reflect issues of equity, status, and recognition for women and minorities that have been raised in previous literature on academic and professional careers (August and Waltman 2004; Oshagbemi 2000; Callister 2006; Sabharwal and Corley 2009; Bender and Heywood 2006).

9 Conclusion

9.1 Data science is multifaceted and brings value across academia

The methods, skills, approaches, techniques, technologies, infrastructures, values, and priorities that have collectively come to be known as "data science" bring much value to academia. Data science is in strong demand in across the private and public sector, and many universities are creating new programs and initiatives focused on teaching students what they need to be successful in a variety of careers outside of academia. While these educational efforts are important, the many faces of data science also bring substantial value to universities internally, particularly in terms of research. It is important to understand that data science is still in formation and differently understood by different people. Furthermore, data science will likely never be a unified monolith, but instead a broad network of many different kinds of people who contribute to the mission of the university in many different ways. Some academics in data science look like traditional tenure-focused academic researchers and have contributions that more easily fit into existing incentive and reward structures, but many do not. Furthermore, given there is ambiguity in who is considered a data scientist, there are many people who do work that support data science, but may not fit into some people's definitions of what a data scientist is.

9.2 Data scientists are making new kinds of contributions and scholarship in academia

It is important to emphasize that there are many kinds of contributions that academics can make to data science, and that those who make such contributions can have different positions, roles, and degrees to which they identify as a data scientist. In this survey, we have not directly focused on the kinds of contributions that academic data scientists bring — to the fields, to their institutions, to academia as a whole, and to data scientists across the private and public sectors. However, we

must emphasize that the forms and modes of scholarly contributions are also changing around data science. Existing academic incentive and reward structures often have difficulty with contributions outside of traditional research publications, and many of the issues and uncertainties we found are linked to whether these kinds of contributions will be valued and rewarded. In working to support career paths for data scientists, we must not forget that these new kinds of academic practice involve new kinds of contributions to their fields, universities, and beyond, such as:

- developing and maintaining software tools and packages that others in their fields (and beyond) rely on for their own data-intensive research
- developing and maintaining specialized research infrastructures that researchers at their institutions (and beyond) rely on for their own data-intensive research
- working with researchers to help them understand the right methods, tools, and infrastructures for their research projects (ranging from mentoring to research consulting)
- collaborating across fields and disciplines to develop methods, tools, and infrastructures that are broadly applicable and usable by many kinds of researchers and practitioners
- teaching shorter workshops that fill in the gaps of more formalized educational offerings
- mentoring students who have a broad range of backgrounds and interests in data science
- novel and traditional forms of academic service and "meta-research" in defining policies, priorities, standards, and best practices for data science, especially across fields and institutions

9.3 Institutional change

Institutional change is a difficult and complex issue in any large-scale organization, particularly in academia. Some kinds of academic units and positions are designed to be flexible to rapid change, while others are designed to maintain the long-term institutional priorities of academia. While this study has not involved an empirical study of institutional change around data science in universities, we do find that early career positions appear to be more flexible to incorporating new kinds of activities, priorities, and roles. In contrast, ladder-rank faculty positions appear to be less flexible, with more resistance and uncertainty about whether data science contributions and expertises will count for hiring and promotion. Within the context of a major research-focused university, it can be substantially easier to fund and hire graduate research assistants, postdocs, and short-term research staff to do activities requiring data science expertise than it can be to fund and hire a faculty member for the same reasons. Similarly, it can be substantially easier to get approval and funding to teach short workshops about data science compared to full courses, particularly those incorporated into a degree program. There can also be trade-offs between flexibility and stability in these institutional structures. The same qualities that make it easier to support early career positions and workshops for data science may also likely be the same ones that give them their short-term nature. Future work on career paths should keep these tensions in mind.

9.4 Directions and recommendations for future work

There is much future work in both studying academic career paths in data science, as well as working to support them. Our study was limited to a non-representative subset of data scientists within academia, as we surveyed members and affiliates at institutions explicitly organized around data science at three large, research-focused universities in the U.S. Future research on academic

career paths in data science should broadly survey across various departments, institutions, and countries, as there may be substantial differences. Given the many different definitions of data science and different levels of self-identification as a data scientist even in our subset of academic data science, we recommend that future research broadly recruit academics using a variety of keywords beyond "data science." Future surveys should also ask respondents to detail their current positions, duties, and time distributions, which we did not do in this survey. We also call for a wide range of interview-based and ethnographic research into academic career paths, taking a deeper and more situated look into many of the issues we have identified in this report.

We also find that many of the issues around career paths in academic data science stem from issues that exist across academia, not just in data science. As we reviewed in the first section, issues around faculty placement, promotion, the "postdoc purgatory," and priorities around non-tenure-track paths have been extensively studied and discussed across academia. Future work on career paths in academic data science — in both studying career paths and working towards supporting them — should examine previous historical changes to academic universities. We can learn much from previous cases of fields that have institutionalized (to varying degrees) in universities, particularly those that formed post-1945, including: computer science, bioinformatics, cognitive science, public policy, communication studies, cultural/area studies, Schools of Information, and the digital humanities. Both research and institutional change should also examine alternative career models in universities and other research institutions, such as how university libraries and the national labs[8] fund and support researchers and staff. Future work should also draw on the expertise of academic researchers who study institutional change across various contexts, such as organizational sociology, science and technology policy, and history, sociology, and anthropology of science and technology.

9.5 Summary of recommendations

(also in section 1.4)

- Academic data science involves a variety of new topics, roles, and activities (as well as novel combinations of established and new topics, roles, and activities), which are often not fully supported by traditional academic career paths. As there is no single model of what an academic data scientist does, universities should define and support a broad and diverse range of positions and career paths for data scientists, both within and across disciplines.

- For early career researchers, there can be substantial ambiguity and uncertainty about whether their academic institutions will reward and support long-term career paths for those who focus on topics, roles, and activities associated with data science. Universities should facilitate conversations within and across disciplines about formalized criteria and expectations for both tenure-track and non-tenure-track positions in and around data science.

- While salary is an important factor (with industry positions paying lucrative salaries), our respondents generally placed even more value on secure, long-term employment and intellectual freedom. Many academic data scientists are strongly seeking to avoid precarious positions, not wanting to be exclusively funded by "soft money." Universities should support tenure-track positions for data scientists, as well as long-term career paths (e.g. with 3-5 years of stability) for data scientists in non-tenure-track academic and research staff positions.

[8] For example, the U.S. National Labs, which are largely funded by the U.S. Departments of Energy and Defense.

- Funding for research projects, computational infrastructure, and education/professionalization initiatives has been crucial to the success of academic data scientists, which should be continued and expanded. To further support academic data scientists, universities should also support a broad range of formal and informal training, mentoring, and professionalization initiatives. Such efforts should include both events that take place within and across disciplines and institutions, as some issues may be broadly applicable but others may be quite specific. Partnerships with industry-focused student career groups are also recommended.
- In working to support career paths in data science, it is important to work to support diversity and inclusion across many dimensions, including gender, race/ethnicity, national origin, class (including first generation college/grad students), and type of institution (e.g. large research-focused universities, four-year universities, and small liberal arts colleges).

10 Acknowledgements

Many people have supported this project, including those who helped us formate our research questions, gave feedback on the questionnaire, helped us distribute the survey, helped us set up computational environments for analyzing our data, and gave us feedback on our study, our analyses, and this report at various parts of the process. We would like to thank (in alphabetical order): Aaron Culich, Ali Ferguson, Anissa Tanweer, Anna Jefferson, Carly Strasser, Cathryn Carson, Chris Hench, Chris Holdgraf, Daniel Turek, Ed Lazowska, Jonathan Dugan, Karthik Ram, Kevin Koy, Magdalena Balazinska, Marsha Fenner, Micaela Parker, Mik Laver, Nelle Varoquaux, Nick Adams, Olivier Philippe, Sarah Stone, Stacey Dorton, Tyler McCormick, Yuvi Panda [there are certainly more people!]. This project is also an expansion and reformulation of a previous unpublished survey conducted in 2015 by several of the authors of this report and Daniel Turek, who we greatly thank for previous work with us in this area. This project also would also not be possible without all our respondents who took the time to answer our survey. We also thank all the open source software developers who worked to produce the tools and infrastructure we used to conduct this research, which we cite in the appendix below. This work was directly supported by the Gordon and Betty Moore Foundation (Grant GBMF3834) and the Alfred P. Sloan Foundation (Grant 2013-10-27), as part of the Moore-Sloan Data Science Environments.[9]

11 Appendix

11.1 Survey overview

11.1.1 Distribution

We distributed the survey to 614 potential respondents across the Data Science Environments at University of California, Berkeley, New York University, and the University of Washington, Seattle. Respondents were selected in coordination with administrative staff at each DSE, who helped identify "core" and "peripheral" members. Core members were defined as those who are on staff at the DSE, have a fellowship or funding through the DSE, have a formal title through the DSE, or serve a designated role in the DSE. Peripheral members were defined as those who did not fit the core criteria, but had some kind of membership, connection, or affiliate status at the DSE, including those who were subscribed to a local events mailing list.

[9] https://msdse.org

We used the Qualtrics survey platform to host and distribute the survey. The survey was distributed over a one month period, synchronized to start approximately two weeks after the start of classes for the 2016 Fall semester for each university (which had different start dates). Two follow up e-mails were sent to respondents: one after 7 days and another after 14 days, and respondents had a total of 1 month from the time of the first e-mail to complete the survey.

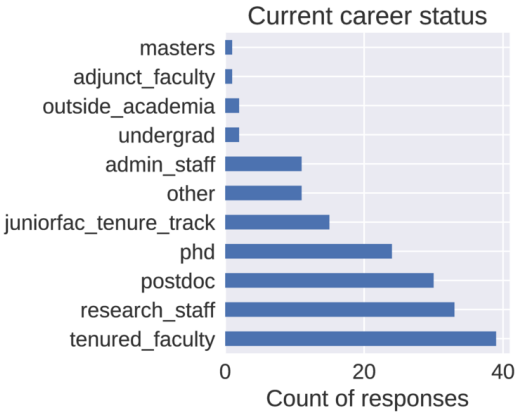
Out of 614 respondents who were e-mailed the survey, 226 respondents started the survey, and 169 respondents completed the survey, for a total response rate of 27.5% and a completion rate of 74.8%. For core members, 237 were e-mailed, 115 started the survey, and 98 completed, for a response rate of 41.3% and a completion rate of 85.2%. Peripheral respondents had a lower total response rate of 18.6% and a lower completion rate of 64.0%. In line with our IRB approval, we did not link e-mail addresses to responses and do not have a field in our dataset for core/peripheral status, but we did ask optional demographic questions about whether the respondent 1) was a member of a DSE, 2) received funding through a DSE, and 3) attended one of the three DSE annual summits.

11.2 Respondent overview

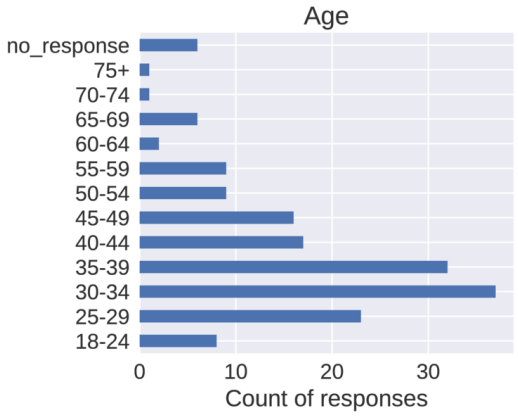
11.2.1 Demographics

Survey respondents ranged from undergraduates to tenured faculty, but as figures 18a and 18c show, the majority of respondents were non-ladder rank researchers: Ph.D students (14.2%), postdocs (17.8%), and research staff (19.5%). In the category of ladder-rank faculty, 8.9% of all respondents were tenure-track faculty and 23.1% were tenured faculty. 53.3% had principal investigator status at their institution. Institutionally, UC-Berkeley respondents are the most represented in the sample, both due to the larger number of individuals in the UC-Berkeley data science population and the higher response rate. Figure 18b shows the distribution of ages, which ranged from 18-24 to 75+, peaking at the 30-39 range. Figure 18f shows the distribution of gender, with 58.6% men, 35.3% women, 4.8% with no response, and 0.5% each for transgender and other.

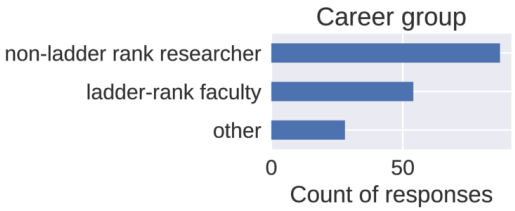
For race/ethnicity, we used the standard 2010 U.S. Census questionnaire, which allows selecting multiple options for race (including a 'multiracial' option) and a separate yes/no question for hispanics. Figure 18g shows the breakdown by options and figure 18h shows the breakdown by whether the respondent selected a non-white/minority race/ethnicity 74.0% of respondents reported only a white/Caucasian background with no Hispanic background, 18.3% reported a minority race/ethnicity, and 7.7% did not respond.



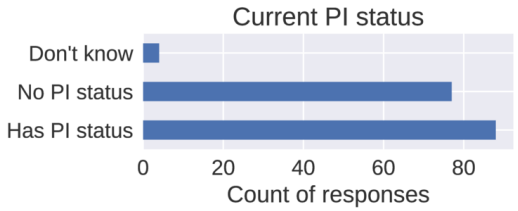
(a) **Career stage**



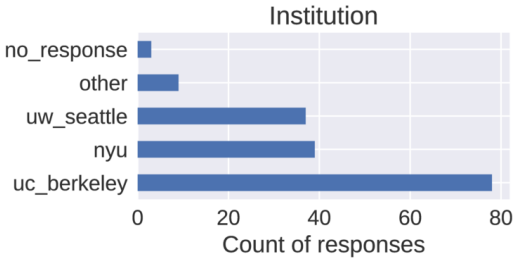
(b) **Age**



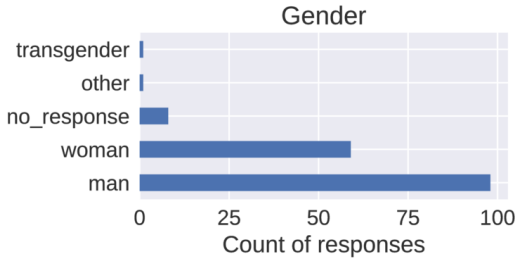
(c) **Career group**



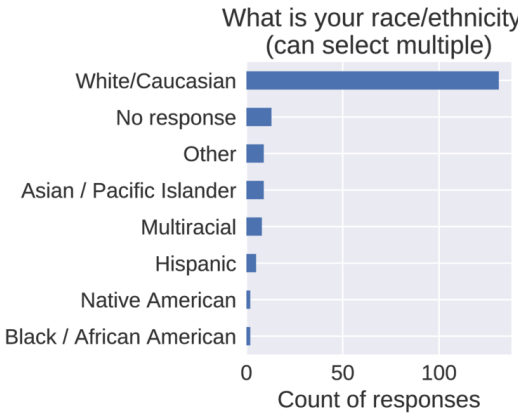
(d) **Has PI status**



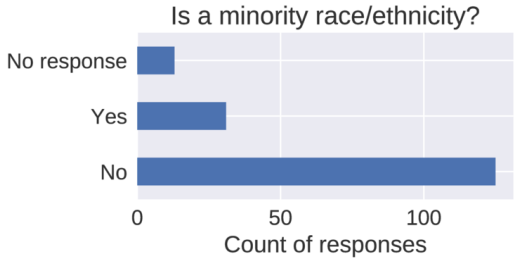
(e) **Institution**



(f) **Gender**



(g) **Race/ethnicity**

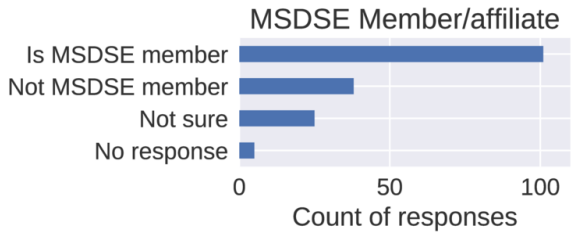


(h) **Is a member of a minority race/ethnicity?**

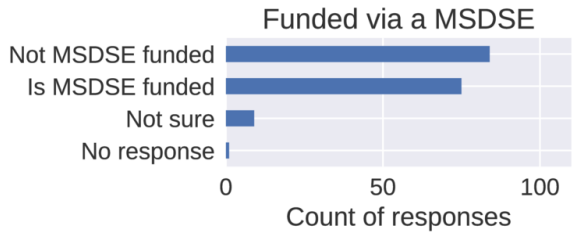
Figure 18: Demographics

11.3 Connection with Moore-Sloan Data Science Environment

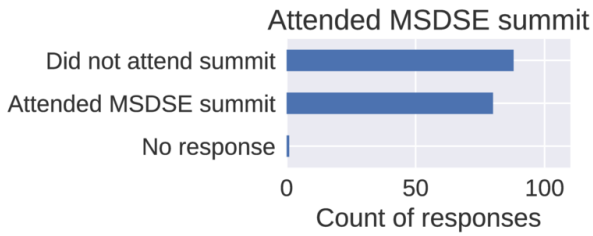
All respondents were recruited by individuals at a Moore-Sloan Data Science Environment, but many did not have formal affiliation or membership with a MSDSE. Figures 19a-19d shows three ways of representing a connection with the MSDSE initiative, including being a MSDSE member or affiliate (19a, being funded through a MSDSE 19b, having attended at least one annual MSDSE summit (19c. Figure 19d shows the number of respondents who answered yes to one of the three questions, with 65.7% having at least one formal connection to the MSDSE.



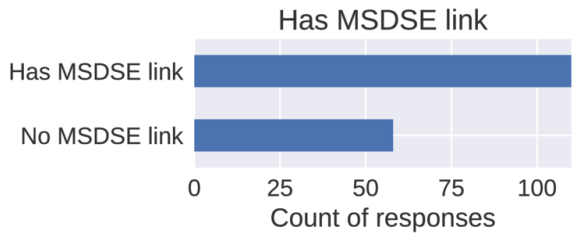
(a) **Member/affiliate of a MSDSE**



(b) **Funded through a MSDSE**



(c) **Attended at least one MSDSE annual summit**



(d) **Has MSDSE link (answered yes to one of the 3 questions)**

Figure 19: **Connection with Moore-Sloan Data Science Environment**

11.4 Software tools used

Qualtrics was used to design and administer the survey. This analysis was conducted in Python (Rossum 1995) and R (R Core Team 2016), using: Pandas dataframes (McKinney 2010) for data parsing and transformation; SciPy (Jones, Oliphant, Peterson, et al. 2001) and NumPy (Walt, Colbert, and Varoquaux 2011) for quantitative computations; Matplotlib (Hunter 2007), Seaborn (Waskom et al. 2014), and ggplot2 (Wickham 2009) for visualization; and psych for factor analysis (Revelle 2017). Analysis was conducted in Jupyter Notebooks (Kluyver et al. 2016) using the IPython (Pérez and Granger 2007) and IRkernel kernels.

References

August, Louise and Jean Waltman (2004). "Culture, Climate, and Contribution: Career Satisfaction Among Female Faculty". In: *Research in Higher Education* 45.2, pp. 177–192. ISSN: 0361-0365. DOI: 10.1023/B:RIHE.0000015694.14358.ed. URL: http://link.springer.com/10.1023/B:RIHE.0000015694.14358.ed.

Bender, Keith A. and John S. Heywood (2006). "Job Satisfaction of the Highly Education: The Role of Gender, Academic Tenure, and Earnings". In: *Scottish Journal of Political Economy* 53.2, pp. 253–279. ISSN: 0036-9292. DOI: 10.1111/j.1467-9485.2006.00379.x. URL: http://doi.wiley.com/10.1111/j.1467-9485.2006.00379.x.

Callister, Ronda Roberts (2006). "The Impact of Gender and Department Climate on Job Satisfaction and Intentions to Quit for Faculty in Science and Engineering Fields". In: *The Journal of Technology Transfer* 31.3, pp. 367–375. ISSN: 0892-9912. DOI: 10.1007/s10961-006-7208-y. URL: http://link.springer.com/10.1007/s10961-006-7208-y.

Conway, Drew (2013). *The Data Science Venn Diagram*. URL: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Davenport, Thomas H and DJ Patil (2012). "Data Scientist: The Sexiest Job of the 21st Century". In: *Harvard business review* 90.5, pp. 70–76. URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Fox, Peter and James Hendler (2014). "The Science of Data Science". en. In: *Big Data* 2.2, pp. 68–70. ISSN: 2167-6461. DOI: 10.1089/big.2014.0011. URL: http://online.liebertpub.com/doi/full/10.1089/big.2014.0011.

Geiger, R. Stuart et al. (2018). *Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey*. Report. Berkeley, California: UC-Berkeley Institute for Data Science. URL: http://stuartgeiger.com/papers/careers-data-science-msdse.pdf.

Glaser, Barney G. and Anselm L. Strauss (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Hawthorne, NY: Aldine de Gruytner.

Harris, Harlan, Sean Murphy, and Marck Vaisman (2013). *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. "O'Reilly Media, Inc.".

Hey, Tony, Stewart Tansley, and Kristin Tolle (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*. Redmond: Microsoft Research. ISBN: 0982544200.

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55. URL: http://ieeexplore.ieee.org/document/4160265/.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. URL: http://www.scipy.org/.

Kahn, Shulamit and Donna K Ginther (2017). "The impact of postdoctoral training on early careers in biomedicine". In: *Nature Biotechnology* 35.1, pp. 90–94. ISSN: 1087-0156. DOI: 10.1038/nbt.3766. URL: http://www.nature.com/doifinder/10.1038/nbt.3766.

Kim, M. et al. (2016). "The Emerging Role of Data Scientists on Software Development Teams". In: *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, pp. 96–107. DOI: 10.1145/2884781.2884783.

Kluyver, Thomas et al. (2016). *Jupyter Notebooks: a publishing format for reproducible computational workflows*. Ed. by Fernando Loizides and Birgit Scmidt. URL: https://eprints.soton.ac.uk/403913/.

Manieri, Andrea et al. (2015). "Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists". In: *Cloud Computing Technology and Science (CloudCom), 2015 IEEE 7th International Conference on*. IEEE, pp. 588–593.

Mattmann, Chris A (2013). "Computing: A vision for data science." In: *Nature* 493.7433, pp. 473–5. ISSN: 1476-4687. DOI: 10.1038/493473a. URL: http://dx.doi.org/10.1038/493473a.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56. URL: http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Oshagbemi, Titus (2000). "Gender differences in the job satisfaction of university teachers". In: *Women in Management Review* 15.7, pp. 331–343. ISSN: 0964-9425. DOI: 10.1108/09649420010378133. URL: http://www.emeraldinsight.com/doi/10.1108/09649420010378133.

Pérez, Fernando and Brian E. Granger (2007). "IPython: a System for Interactive Scientific Computing". In: *Computing in Science and Engineering* 9.3, pp. 21–29. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53. URL: http://ipython.org.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Revelle, William (2017). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.7.3. Northwestern University. Evanston, Illinois. URL: https://CRAN.R-project.org/package=psych.

Roach, Michael and Henry Sauermann (2017). "The declining interest in an academic career". In: *PLOS ONE* 12.9. Ed. by Joshua L Rosenbloom, e0184130. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0184130. URL: http://dx.plos.org/10.1371/journal.pone.0184130.

Rossum, Guido van (1995). *Python Library Reference*. URL: https://ir.cwi.nl/pub/5009/05009D.pdf.

Russo, Gene (2011). "Graduate students: Aspirations and anxieties". In: *Nature* 475.7357, pp. 533–535. ISSN: 0028-0836. DOI: 10.1038/nj7357-533a. URL: http://www.nature.com/doifinder/10.1038/nj7357-533a.

Sabharwal, Meghna and Elizabeth A. Corley (2009). "Faculty job satisfaction across gender and discipline". In: *The Social Science Journal* 46.3, pp. 539–556. ISSN: 0362-3319. DOI: 10.1016/J.SOSCIJ.2009.04.015. URL: http://www.sciencedirect.com/science/article/pii/S0362331909000500.

Sciences, National Academy of, National Academy of Engineering, and Institute of Medicine (2014). *The Postdoctoral Experience Revisited*. Washington, DC: The National Academies Press. ISBN: 978-0-309-31446-6. DOI: 10.17226/18982. URL: https://www.nap.edu/catalog/18982/the-postdoctoral-experience-revisited.

Walt, S. van der, S. C. Colbert, and G. Varoquaux (2011). "The NumPy Array: A Structure for Efficient Numerical Computation". In: *Computing in Science Engineering* 13.2, pp. 22–30. ISSN: 1521-9615. DOI: 10.1109/MCSE.2011.37. URL: https://arxiv.org/abs/1102.1523.

Waskom, Michael et al. (2014). *seaborn: v0.5.0 (November 2014)*. DOI: 10.5281/zenodo.12710. URL: https://doi.org/10.5281/zenodo.12710.

Wickham, Hadley (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

Woolston, Chris (2015). "Graduate survey: Uncertain futures". In: *Nature* 526.7574, pp. 597–600. ISSN: 0028-0836. DOI: 10.1038/nj7574-597a. URL: http://www.nature.com/doifinder/10.1038/nj7574-597a.

— (2017). "Graduate survey: A love-hurt relationship". In: *Nature* 550.7677, p. 549. DOI: 10.1038/nj7677-549a. URL: http://www.nature.com/doifinder/10.1038/nj7677-549a.