Evaluation of the Moore-Sloan Data Science Environments

Final Report

February 2019

Prepared for:

Joshua Greenberg, PhD
Alfred P. Sloan Foundation

Chris Mentzel
Gordon and Betty Moore Foundation

Prepared by:

Luba Katz, PhD
Abt Associates

CONTENTS

Executive Summary

The goal of the Moore-Sloan Data Science Environments (MSDSE) program, established jointly by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation, is to create positive conditions for data-driven discovery at academic institutions. In 2014, the program funded three data science environments: the Berkeley Institute for Data Science (BIDS) at the University of California Berkeley, the Center for Data Science (CDS) at the New York University, and the eScience Institute (eScience) at the University of Washington. Each university received $12.5 million to support the environments for five years.

In 2015, the Foundations contracted with Abt Associates to conduct a 3-year evaluation of the MSDSE program. The study drew on several sources: review of MSDSE reports, websites, and policy papers; three site visits to each university; 132 interviews and two on-line surveys of participants and non-participants; and observations at three Data Summits. In addition, we examined the broader data science landscape by interviewing the leaders of 17 entities similar to MSDSEs about their efforts. Based on these sources, we reached the following conclusions:

The MSDSE funding strategy was effective. We concluded that the MSDSE program is accomplishing its goals and identified several design elements we believe played a positive role. First, the selection of grantees, which was based on the perceived commitment to the mission of institutional change espoused by the program, ensured that the energy to launch the environments was in place at the time of award and could be quickly harnessed. Second, the culture of partnership and experimentation adopted by the program facilitated mutual learning and growth. Third, the framework of agreed-upon working groups, each addressing a known "pressure point," enabled the centers to launch a cohesive set of programs and hold themselves accountable. Finally, a strong focus on the development of tools and practices, unusual for funders, was an insightful choice, as it highlighted the importance of this work at the universities where it is undervalued, and was embraced by the participants.

Physical space played a key role at MSDSEs. All three sites invested significant efforts in designing their spaces and learned from each other's experiences. The program demonstrated that attractive and centrally located space can raise the profile of the center on campus, draw participants, and facilitate collaboration.

Strong management and adequate staffing are important to the function of the centers. We found that having a critical mass of personnel to implement center programs helped reduce participant fatigue. PhD-level data scientists and research engineers were particularly well-suited for staffing the centers, as they had a broad range of interests and skills, tended to view their positions as longer-term, and were under less pressure to publish than postdocs. A combination of 4-5 data scientists who spend most of their time running programs and several postdocs/faculty/students contributing in a more limited manner is a staffing model that worked well for MSDSEs. Finally, an executive director position to help establish and manage the centers was an essential part of their success.

MSDSEs established promising career tracks for data scientists. One of the major benefits of the MSDSE grant was an opportunity to pilot positions which are difficult to fund at universities. Of the various tracks explored, in our view, CDS fellows, BIDS research scientists, and eScience data/research scientists were particularly good models. These staff played key roles in the environments and were highly satisfied with their experience. All sites were trying to sustain these positions beyond the duration of the MSDSE grant.

MSDSEs developed new programs that led to collaboration. Within a few years, each center established itself as a community hub by actively building bridges with departments and providing multiple entry points into the environment. While numerous programs were offered at each university, the Incubator and Data Science for Social Good programs at eScience and the XD working groups at BIDS appeared to be especially popular. A survey of participants in the two eScience programs revealed that many had maintained collaborations for at least one year.

MSDSEs were productive in research output. Over a three-year period, the centers reported 200 software products (virtually all publicly available) and 1,200 papers. In addition, they developed and disseminated various practices and tools to enable open and reproducible research.

Dedicated strategies may be necessary to maintain faculty involvement. We found that unless faculty had offices in the space and/or were supervisors of MSDSE postdocs, their involvement in center activities declined over time. Faculty interviews revealed that the lack of time and scheduling conflicts, not interest, impeded participation. Strategies to reduce attrition could include joint faculty lines, programs to protect faculty time (such as sabbatical leaves or teaching buy-outs), and rotating center leadership positions.

Universities included in the landscape review created unique entities from similar building blocks. We were unable to stratify 20 data science entities into distinct "families" based on staffing, programs, mission, or another characteristic(s). This suggested to us that universities are still developing an intuitional response to changes in the scientific enterprise brought about by big data, and that there is currently no consensus on how to organize a data science center.

We hope that our study offers helpful insights into the range of directions for supporting data science in academia.

Chapter 1: Introduction

Origins of the Moore-Sloan Data Science Environments

Technological advancements of the past 10–15 years have resulted in producing an unprecedented volume of scientific data. However, several institutional challenges remain to harnessing its full potential for scientific discovery. The Moore-Sloan Data Science Environments (MSDSE) program was established jointly in 2012 by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation (the Foundations) to empower data scientists and create positive conditions for data-driven science.[1-2] The goals of the program are three-fold:

- To develop and maintain collaborations between domain scientists (e.g., physicists, biologists, and social scientists) and methodologists (e.g., statisticians, computer scientists, and applied mathematicians)
- To establish rewarding and sustainable career paths for data scientists
- To contribute to the ecosystem of analytical tools and research practices to facilitate data-driven discoveries across scientific areas.

The selection of MSDSE grantees followed a multistep process. It began with numerous discussions between the Foundations staff and representatives from academia and industry about the unmet needs of data scientists and with institutions that might have the interest and capabilities to bridge these gaps. Based on these consultations, the Foundations developed a Letter of Interest (LOI) to begin a conversation with candidate institutions, which posed the following questions:[3]

- Who on campus is doing really innovative, compelling work methodologically; and developing and using new techniques?
- Where on campus are the pockets of strengths, and which domains on campus are building strength and buying into data-driven research?
- At the institutional level, what is the campus history and track record of innovation and of successful collaboration with other universities?
- Who has demonstrated engagement in the issue of data science and a commitment to advancing it (at the university leadership level as well as at the departmental and faculty levels)?

The LOI was sent to the vice presidents for research at 15 universities deemed competitive, who were asked to identify faculty representatives who would submit a response. A small group of advisors reviewed these responses, and narrowed down the initial pool of 15 to 6

---

[1] https://www.moore.org/initiative-additional-info?initiativeId=data-driven-discovery.

[2] The MSDSE program is one of three mechanisms (the "institution strategy") under the Moore Foundation's Data-Driven Discovery Initiative.

[3] Request for Letter of Interest in Crafting Solutions to the Data Science Challenges of the 21st Century. Moore and Sloan Foundations. November 2012.

universities that demonstrated the most authentic commitment to advancing data science.[4] Program staff from the Foundations then met in person with faculty, research staff, and the administration at these institutions. Three of the six universities stood out during these visits with their vibrant and collaborative cultures and were selected. The funded environments included the Berkeley Institute for Data Science (BIDS) at the University of California Berkeley (UCB), the Center for Data Science (CDS) at the New York University (NYU), and the eScience Institute (eScience) at the University of Washington (UW). Each university received $12.5 million for five years. Faculty from these universities met for a two-day facilitated workshop to design the environments; the resulting proposal laid out the framework for the MSDSE program.

Approach to implementing the MSDSE concept

The data science environments are organized around several "working groups," each focusing on mitigating a recognized challenge to data-driven science in academia. These included:

- Careers (renamed Career Paths and Alternative Metrics) – to create new career paths in academia for interdisciplinary researchers and for data scientists
- Education and Training – to offer both formal and informal training in data science skills at undergraduate, graduate, and professional levels
- Tools and Software – to enable development, hardening, dissemination, and use of software tools and infrastructure that support data-driven research
- Reproducibility and Open Science – to develop and promote practices designed to enhance data sharing, preservation, provenance tracking, and reproducibility of scientific workflow and data-intensive analyses
- Space and Physical Organization (renamed Working Spaces and Culture) – to design common spaces that encourage collaboration both within and between universities
- Evaluation and Ethnography (renamed Data Science Studies) – to understand the complex landscape within which data science is situated, and identify and evaluate best practices
- Methods (NYU only) – to put domain scientists in touch with relevant methods scholars, building collaborations that develop new methods.

While adopting this common blueprint, each MSDSE proposed its own staffing and activities for the working groups. The grantees also acknowledged the experimental nature of their approaches, leaving the door open for mid-course corrections. The program implementation phase began in fall 2013.

Abt evaluation

In 2015, the Foundations contracted with Abt Associates (Abt) to conduct an evaluation of the MSDSE program. Following a feasibility study, we developed a mixed-method evaluation approach to address the following questions:

[4] Interview with staff at the Moore and Sloan Foundations. June 2018.

- In what ways are the three data science environments different from and similar to one another? What are the relative strengths and weaknesses of different approaches?
- What challenges have the MSDSEs experienced and how have these been addressed? How have the environments evolved over time?
- To what extent are the MSDSEs accomplishing their stated goals?
- What institutional and cultural changes have occurred, and can any of these changes be attributed to the MSDSE funding?
- Are the successful elements of the MSDSEs sustainable?
- What are the characteristics of the data science programs established by other institutions and how do they differ from MSDSEs?
- Did the Foundations select the right strategies to achieve their stated goals?

To answer these questions, we reviewed MSDSE progress reports, proposals, policy papers, and public websites; visited each university annually for three years; conducted 132 individual and group interviews; attended three MSDSE Annual Summits; and implemented online surveys of participants in three MSDSE programs and in a Data Science Leadership Summit organized initially by Columbia University. We note that this report covers the period through May 2018, which marked the end of our data collection. We recognize that several important developments have occurred at MSDSEs in the past few months, which are not reflected.

The report is organized into 10 chapters:

- Chapter 1 – introduction to the program and the evaluation goals
- Chapter 2 – methods used in the evaluation
- Chapters 3–5 – organization, activities, and participant experiences at each MSDSE
- Chapter 6 – cross-MSDSE activities
- Chapter 7 – outcomes of the MSDSE incubator programs
- Chapter 8 – MSDSE-supported software and its role in developing careers
- Chapter 9 – review of the academic data science landscape in the United States
- Chapter 10 – summary of findings in the context of research questions.

Chapter 2: Methods

Site visits, interviews, and observations

We conducted site visits to each MSDSE university in February–March 2016, March–April 2017, and April–May 2018. The goals of the first visit were to learn about the centers and meet the participants. During the second and third visits, we focused on the role of the MSDSEs on campus, their evolution, emerging accomplishments, and challenges. During the third visit, we also discussed the strengths and weaknesses of the centers that emerged so far and the plans for sustainability beyond the end of the grant. Two Abt researchers spent 1-2 days at each site conducting individual and group interviews, observing the space, and attending events which happen to occur at the time.

We also conducted phone interviews with four groups: faculty members (N=26); fellows, postdocs, and data scientists/engineers (N=22 in 2017 and N=23 in 2018); software development leads (N=9); and leaders of non-MSDSE data science centers (N=23). In total, we completed 132 interviews (Exhibit 1). More than half of the MSDSE respondents were interviewed multiple times, either because they played several roles in the centers and/or because we were trying to capturing their views as the centers evolved. All interviews were conducted using semi-structured protocols and took 30–60 minutes. With the respondents' permission, we digitally recorded and transcribed the interviews, and used *NVivo* software to code and analyze the data. Finally, we attended three annual Data Science Summits hosted by MSDSEs, during which we observed the talks and breakout discussions, and interacted with the participants informally during breaks and meals.

EXHIBIT 1: RESPONDENT GROUPS INTERVIEWED DURING THE STUDY

Group	2015–16	2016–17	2017–18*
MSDSE PIs, executive directors, and managers	✓	✓	✓
Steering or Executive Committee at MSDSEs	✓	✓	✓
Working group leads at MSDSEs	✓	✓	✓
Administration at MSDSE universities		✓	
Non-data science environment researchers at MSDSE sites		✓	
Fellows, postdocs, data scientists/engineers at MSDSE sites	✓	✓	✓
Faculty members at MSDSEs			✓
Developers of MSDSE-supported software			✓
Leaders of data science entities at non-MSDSE sites			✓

**Note: data collection ended in May 2018.*

Analysis of program documents

Annual progress reports submitted by the MSDSEs in 2014–2017 included two parts. The first was Excel spreadsheets, containing data on (a) individual participants, such as their names, titles, fields, conferences attended/talks given, grants/awards, and outputs and (b) centers as a whole, including hires, transitions, institutional grants, and events. In addition, progress reports included narratives describing activities, accomplishments, and challenges. Much of these data were abstracted and presented as tables and charts below.

Review of the MSDSE software

Progress reports for 2015–2017 were used as the initial source of software products. After removing duplicates, multiple releases of the same tools, and minor items (such as snippets of code), the list contained 234 tools: 82 from BIDS, 86 from CDS, and 66 from eScience. For all items, we collected impact and usage metrics using the following sources: GitHub (contributors and releases), GoogleScholar (citations), Libraries.io (source rank), and Altmetric (attention score). Each software product was assigned a score from 1 to 3 to select 9 tools for more in-depth analysis. The scoring system was as follows: 1 – items with a low citations count (< 100) and/or low GitHub activity, zero publications, release date before the start of MSDSE grant; 2 – items with high citation count (100+) and/or high software update count (500+) and/or evidence of collaboration based on co-authorships; 3 – items that acknowledged support from the Foundations and/or included at least one MSDSE-affiliated developer, were well-documented, or had 50+ citations.

Of the 234 products reviewed, 39 received a score of 2 or 3, from which we identified 2–3 items for each MSDSE that included at least one that was domain-specific and one that was applicable to multiple fields. We made every effort to exclude software authored by MSDSE participants already being interviewed for multiple other tasks to reduce demands on their time. This process resulted in the selection of the following tools: Librosa, TopoAngler, Carl, and ReproZip (CDS); AstroPy and Pulse2percept (eScience); and Viscm, Sncosmo, and Permute (BIDS). For each tool, we reviewed all publicly available information and interviewed the main developer about the rationale for creating the tool, the role of MSDSE funding in its development, and the benefits to respondents' careers.

Surveys

We used *SurveyGizmo* software to collect data from the participants in the UW Data Science for Social Good (DSSG) and Incubator programs, and in the NYU Seed Grants program.[5] A survey questionnaire was shared with the MSDSEs for input, programmed, and pilot tested. The survey was open for four weeks beginning on March 12, 2018. The survey sample included all project leads and student fellows who participated in the DSSG and Incubator programs in 2014–2017 (N=60 and N=23, respectively), and all PIs who participated in the Seed Grant program in 2016–2017 (N=16). We were able to achieve response rates of 60%, 87%, and 50%, respectively. Response frequencies were calculated for each survey item and χ^2 tests were performed to determine whether the differences in responses were statistically significant. Similar procedures were used in implementing a short survey of participants in the Data Science Leadership Summit held in October 2018.

Reports

We shared this report and the landscape report with MSDSE and non-MSDSE leadership, respectively, and made most of the suggested revisions.

[5] BIDS staff opted for an interview about the Machine Shop program instead of the survey of participants.

Chapter 3: Berkeley Institute for Data Science

In this chapter, we describe the MSDSE established at the UCB, including its physical space and leadership, participant experiences, accomplishments, and plans for sustainability.

Key findings:

- BIDS helped create the conditions to establish its new Division of Data Science and Information
- BIDS location in the main library has increased awareness of data science on campus and helped establish its lead role in interdisciplinary research and training for the new Division
- Domain scientists and methodologists at all career levels acknowledged the vibrant, intellectually diverse environment of BIDS, but were often uncertain as to how to integrate, contribute, or benefit from it long-term
- Faculty involvement at BIDS has been episodic and difficult to sustain
- BIDS made a fellowship program for postdoctoral scholars and graduate research students a priority, to extend its reach across campus and to support career growth for data scientists in academia
- Returns from the program to both the Institute and the fellowship participants have been mixed - for example, while fellows generally acknowledge a positive "water cooler" effect on their research from the program, some did not find it beneficial to their careers in academia
- BIDS made many contributions to the ecosystem of open source software for data science and analysis, especially scientific Python
- BIDS developed the infrastructure supporting foundational courses in Berkeley's data science degree programs
- Several promising training programs grew organically at BIDS, such as GraphXD, ImageXD and TextXD
- BIDS has not been adequately staffed to sustain programs like the XDs or to meet demand for its data science researchers and practitioners from the campus community

Space and leadership

BIDS is based in the Doe Library at the heart of campus. The large investment in renovation and the choice of popular location signaled a commitment to data science by the administration and elevated the status of BIDS. The open layout of the space, designed by BIDS staff, and its home within the library, were intended to encourage collaboration and community-building. While its location at the main entrance of the main library signaled strong support from campus leadership, it created some unanticipated challenges. Significant foot traffic - hundreds of people per day - in front of the workspace necessitated door monitoring and check-ins, which sometimes made the space appear less inviting. The library norm of being fully accessible was similarly sometimes at odds with BIDS' needs to use the space for closed convening events and for affiliated faculty and students' individual work spaces. Additionally, balancing the need for event space and work space has created tensions among BIDS affiliates. The MSDSEs experience with the space yielded many lessons on the importance on the design, both in location and intended use, which were shared with other sites.

For the first few years, the BIDS core leadership team included a faculty director (also the grant PI) and a non-faculty executive director. The inaugural executive director was responsible for monitoring activities of the working groups, onboarding and supervising fellows, allocating space, serving as a BIDS representative on campus, and preparing progress reports. Given this broad range of duties, it is not surprising that his departure mid-way through the grant was a major setback for BIDS. His replacement, hired six months later, planned to focus on the mission, strategic planning, fundraising, and developing/implementing standard operating procedures for all components of the center. The leadership team of BIDS was expanded in fall 2017 to include a chief research officer (CRO), whose main responsibility is to increase the research footprint of BIDS. At the time of our visit in May 2018, the CRO and the BIDS communications manager were focused on outreach to the Berkeley community to better understand how the center is perceived, clarify its mandate, and expand ties to the university.

All key decisions at BIDS have been made by an Executive Committee, which included as few as four and as many as 10 faculty as well as staff members, depending on the year. A few months prior to our last visit, the Executive Committee began inviting one senior and one junior scholar to its weekly meetings to make the governance process more participatory and increase transparency. This practice was viewed as a success by the committee members.

Participants

In addition to the leadership team, BIDS participants include faculty (called "senior fellows"), non-faculty fellows, and research staff. Non-faculty fellows is a mix of graduate students, postdocs, and research scientists who are co-funded by BIDS and another unit at Berkeley or elsewhere. These researchers are generally appointed to 1-2 year terms to help build bridges to the broader community and make BIDS more visible. We call them "junior fellows." The research staff are fully funded by BIDS and have the titles of computational fellows, research fellows, and research scientists. Funded for five years, these researchers are foundational to BIDS, playing key roles in developing tools, fundraising, running working groups, and leading most other activities at the center. We refer to them as "research scientists."

Based on progress reports, the size of the BIDS community more than doubled between 2015 and 2017, from 31 to 66 participants (Exhibit 2). The distribution by title was as follows: professors 8-24, data scientists 10-11, graduate students 4-12, postdocs 8-13, and professional staff 1-5. Between 10 and 14, depending on the year, identified as social scientists, and the rest as computer scientists (6-14), physicists (2-4), life scientists (4-10), geoscientists (1-7), engineers (1-2), psychologists (1-4), and mathematicians (5-10, Exhibit 3).

EXHIBIT 2: NUMBER OF PARTICIPANTS BY JOB TITLE INCLUDED IN PROGRESS REPORTS

Job title	2015	2016	2017
Assistant/associate/full professor	8	15	24
Data scientist/research scientist/software engineer	10	11	11
Graduate student	4	7	12
Postdoctoral fellow	8	13	13
Professional staff	1	3	5
Other	0	0	1
Total	31	49	66

EXHIBIT 3: DISCIPLINES OF PARTICIPATING RESEARCHERS INCLUDED IN PROGRESS REPORTS

Job title	2015	2016	2017
Mathematical sciences and statistics	5	9	10
Geosciences	1	3	7
Engineering	2	1	2
Computer and information sciences	6	7	14
Psychology	1	4	3
Social sciences	10	12	14
Life sciences	4	7	10
Physics and astronomy	2	4	4
Other	0	2	2
Total	31	49	66

Source: annual progress reports, individuals spreadsheet, metadata tab.

We interviewed most faculty and non-faculty researchers affiliated with BIDS to learn about their roles, benefits and challenges of participation, and career aspirations. Our findings are summarized below.

Faculty experience

Nature of participation

In the participant interviews conducted over the first two evaluation years, we heard that faculty involvement in BIDS was limited from the beginning and declined over time. To better understand the reason for this behavior, we spoke with seven faculty in spring 2018, and found that all but one became involved after the grant was awarded. When asked to describe the nature of their connection to BIDS, most faculty said that they supervise the fellows, attend talks and workshops in the space, and serve on hiring committees. Two faculty respondents also attended the annual Data Summit and participated in a working group (which they expected to be driven by research scientists). The faculty whose research had greater overlap with the mission of BIDS were more involved in its activities.

Benefits of participation

Most of the faculty interviewed highlighted the appeal of BIDS as interdisciplinary "intellectual crossroads," which enabled them to expand their research network at UCB. One faculty noted that BIDS had a "more eclectic" vision of data science than what is typical at academic computer science departments. For example:

> *You cannot underestimate the value of the social network.* (Faculty, 2018)

> *Because of BIDS' welcoming atmosphere, there has been this whole set of the possibilities for people who are in the social sciences to enter this world and discover this world.* (Faculty, 2018)

Support from BIDS led several faculty members to broaden their research programs. One respondent credited BIDS with turning software development in which he engaged from a "hobby" to a formal component of his academic work. He also was grateful for the guidance received from senior BIDS faculty during tenure review. Two faculty said that learning new

methods from research scientists and postdocs was a highlight of their experiences with BIDS, and that they subsequently used this knowledge in their work. One respondent noted that BIDS accelerated translation by bringing theorists and experimentalists together.

> *It is very common to have new discoveries made in machine learning that take 20 years to filter over to experimentalists, but because of things like BIDS that translation can happen very quickly. It is basically [because of] people sitting next to each other.* (Faculty, 2018)

Finally, many faculty praised the contribution of BIDS to undergraduate education.

Participation challenges

All faculty cited time constraints and competing responsibilities as main barriers to participation. One respondent appreciated the flexibility offered by BIDS:

> *I've never felt any pressure or anything to participate more than I have time for, so it's been really an incredibly positive experience.* (Faculty, 2018)

Some faculty expressed concerns for BIDS future, as it has not yet found a niche in the university ecosystem. We were also told that the BIDS mandate of "culture change" was unrealistic, given its relatively small budget and junior staffing.

Non-faculty staff experience

Nature and degree of participation

Research scientists and junior fellows reported that they spend most of their time on research projects. Nevertheless, they were also actively involved in numerous community-building activities, including designing and organizing workshops, seminars, and other events. It emerged from the interviews that the roles and expectations for these participants have not been clearly defined and/or articulated, which created some unease.

> *I think it is probably best described as unrestricted funding from the perspective of the fellows. I think the requirements are pretty minimal. You're supposed to show up, participate in the working groups, and sort of be a good community member.* (Junior fellow)

> *I think it's a bit different now because the leadership is trying to figure out "how do we get funding?" and "maybe we need the fellows to actually do something," but then they aren't telling us what they expect us to do.* (Junior fellow)

Benefits of participation

Similarly to faculty, the key benefit of BIDS mentioned by this group was being immersed in a diverse and dynamic environment. In addition, these respondents appreciated the flexibility to choose their projects and to focus on software development, which was seen as unusual for a traditional postdoc.

> *Probably the most useful thing about being a part of BIDS is hanging out with the other fellows and just hearing through osmosis what they are working on. There are at least two people who I got connected to who have played a fairly large role in my ongoing and future research who I connected to indirectly though BIDS.* (Junior fellow)

> *My favorite is by far these XD groups. The first one I went to was for Text XD, and I had never used text analysis. Now, not only do I have a project related to text analysis now, but we're*

trying to spin that off and turn it into a book, and I have international collaborators on this. It's an area that I had no knowledge of this before and it's really expanded my research horizon. (Junior fellow)

Participation challenges

One of the challenges experienced by non-faculty staff that consistently emerged in the evaluation was finding the right level of involvement in BIDS. For example, shortly after the center was launched, co-funded fellows felt over-taxed by their dual responsibilities. While the number of meetings they were expected to attend was consequently reduced, concerns about the level and nature of participation persisted through our last site visit, possibly exacerbated by the departure of the executive director, who was a *de facto* mentor for junior staff. A related challenge was the duration of the fellowship (two years), which was seen as two short to find one's footing at BIDS. Many respondents also spoke of limited faculty involvement.

I think one of the challenges is the fact we have two year fellowship. By the time that you gain enough institutional knowledge and get to know the people, and how things work, you're about to move on to the next stage of your career. (Junior fellow)

Some of the PIs that were really involved in writing the institute grant, I definitely expected them to wander into the space more often than they did. I would say that of the total number of faculty PI and senior fellows that appear on the BIDS website, maybe 15% were regular faces in the space, at most. (Junior fellow)

In the past a few senior fellows would come to the working group meetings. Now pretty much nobody comes… maybe we'll have one or two at the lunches on Thursdays. But it's usually the same faces. (Junior fellow)

While non-faculty staff perceived pressure from the leadership to participate in fundraising, they received little support to make it possible. We were told, for example, that BIDS did not provide mentorship in grant-writing, which some junior staff had expected. Research scientists noted that the junior- and transient-sounding title of "fellow" along with obstacles to obtaining a PI status made it more difficult to recruit graduate students to work on their projects and to apply for independent funding.

I am doing all of these projects, and BIDS and the university are very happy to point at my work and say, "isn't this really cool work," but I don't have that first class status as a faculty member that would just grease the wheels and make everything a bit easier, including getting grants. I know that if I was assistant professor somewhere a lot of those doubts would go away just based on the title alone. (Research scientist)

Finally, these participants felt excluded from setting a direction for BIDS. Possibly in response to this last criticism, the leadership at BIDS launched "all-hands" meetings to bring the community together, and began inviting junior fellows to its Executive Committee meetings. When we visited BIDS in May 2018, these activities were still new, and the fellows continued to feel disconnected from the decision-making process.

It would be great if our executive leadership realized that there is some untapped capacity and that we are feeling a little frustrated that we are not being included in developing the vision for the institute… It frustrates people who believe they have something here to give. (Research scientist)

Role of MSDSE in career progression

Many junior fellows interviewed were interested in tenure track faculty positions, but it was unclear whether participating in BIDS improved their academic job prospects. It is our impression that a BIDS postdoc was an asset for positions in more methodologically-focused departments, such as statistics or engineering. In contrast, some postdocs in the natural or social sciences told us that they had to justify or even downplay their involvement in the center. However, we note that the number of BIDS alumni at the time of data collection was too small to make any generalizations, and that many factors contribute to success in an academic job search. Furthermore, it is likely that academic culture is at least in part "responsible" for the challenges experienced by junior BIDS staff.

> *I think there is a little bit of a question around the role that BIDS is playing in our career paths. I think for some people it was an obvious plus to have on their CVs, but for me it was a little bit ambivalent…Unless you are in engineering or statistics, being in a data science institute looks a little strange, like "why weren't you doing your postdoc in a biology department?* (Junior fellow)

Research scientists were less certain about remaining in academia and some were using their time at BIDS to chart their next career steps. These staff were grateful for the support to do the work they enjoyed. Our respondents believed that the universities would eventually create attractive non-tenure track positions for people holding positions similar to them, but were not optimistic that this would happen before they looked elsewhere for employment.

> *There is an alternative reality where I would be in big trouble in terms of my career, but thanks to MSDSE and BIDS I am getting to have more time to do this work, to do it right, and actually succeed in it. The role is incredibly valuable to me. The difference in my career trajectory is pretty hard to overstate.* (Research scientist)

> *I think there is a degree of structural change going on in the academy, but I think that it's happening very slowly…Do these kind of positions of leadership that are not tenure-track faculty get created? If not, I'll probably end up going to work for some other non-profit, open source type of place.* (Research scientist)

Activities

Most of the BIDS activities were initially associated with six working groups established during the design phase. Over time, some of the groups became less active because they accomplished their original goals (e.g., Space and Culture); achieved an important result, but did not set new goals (e.g., Reproducibility and Open Science); or were incorporated into broader university-level data science initiatives (e.g., Education and Training). For the education and training initiatives, BIDS usually opted to participate as "good citizen," rather than bring them under the auspices of its own working groups. Nevertheless, BIDS fellows and staff continued to expend significant time on these efforts.

The Education Working Group was very popular and made an important contribution to the university by developing new undergraduate curriculum in data science. According to the BIDS leadership, education resonates with many faculty and staff because it is so fundamental to academic life.

> *The Education Working Group is very active, and it is different than the others. Education has been a neutral playground for all of these initiatives, from the undergraduate to the graduate programs, and workshops and boot camps. It's a place where people trade best practices and ideas. And people made sure that the whole new curriculum got started, so in that sense it may have had more impact on the university than anything else that we did*. (Leadership, 2017)

BIDS Careers and Alternative Metrics working group began by implementing surveys on career paths. The results were written as a report in 2018[6] and presented to the University of California system's Office of the President and to representatives from the 10 UC campuses. However, this impactful local work has come at the expense of maintaining effective collaboration with working group members at CDS and eScience.

The Reproducibility Working Group also attracted enthusiastic membership when first established. In collaboration with the other two MSDSEs, the members published a book of case studies on how to improve the reproducibility of research projects.[7] When this was accomplished, the group struggled to find a new project. One option considered was to begin offering reproducibility-related consulting services to the laboratories on campus, but the plan was abandoned due to lack of leadership and staff.

The remaining working groups were more limited in scale, with the participants drawn primarily from within the BIDS community. The Software Working Group came to be viewed as superfluous because BIDS participants were already actively working on tool development and did not need a mechanism to organize their efforts. Members of this group started to meet once a semester to brainstorm and prioritize ideas, which were implemented largely independently. The Space Working Group dissolved after BIDS moved to its new quarters, which it helped design.

Over time, BIDS shifted to the so-called "XD" (for cross-disciplinary) working groups which focus on particular data types or structures. This new family includes ImageXD, TextXD, VizXD, and GraphXD and brings together people from different disciplines to work on common problems in visualization or text analysis. The groups are typically organized around a training event, workshop, or seminar series. Exhibit 4 shows all activities and programs for each working group mentioned in the progress reports.

In addition to the activities organized by the working groups, BIDS hosted dozens of invited lectures and seminars. BIDS also enabled the organic growth of several discussion groups related to data science tool and techniques, which meet weekly or bi-monthly.

---

[6] RS Geiger, C Mazel-Cabasse, C Cullens, L Noren, 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. https://osf.io/preprints/socarxiv/xe823/

[7] J. Kitzes et al. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences.* University of California Press. First edition (October 17, 2017).

EXHIBIT 4: ACTIVITIES BY WORKING GROUP INCLUDED IN PROGRESS REPORTS

	2014	2015	2016	2017
Career Paths and Alternative Metrics		• Career paths survey • BIDS reference letters		
Education and Training	• DS Fair and BIDS launch event • Thunder talks • DS lecture series • BIDS tea • The Hacker Within the Berkeley chapter opens	• Software carpentry • Python boot camp • The Hacker Within • BIDS collaborative • First undergraduate DS course led by BIDS • Visiting Scholar Program • DS lecture series • DS fair • Data Points @Cal Series	• White paper on education best practices • Berkeley DS Meetup group	• Support for Discover Program
Software Tools, Environment, and Support	• Python for Science Bootcamp	• Data structure for DS Workshop • BIDS Machine Shop	• BIDS Machine Shop • A Field Guide to DS to capture existing best practices	
Reproducibility and Open Science	• Released Berkeley's Computational Environment platform	• Reproducibility Case Study Project • New statistics course which incorporates reproducibility	• Reproducibility case studies book	
Working Spaces and Culture	• Renovation of 190 Doe Library	• Office hours	• Office hours • White paper on space	
Data Science Studies	• Ethnography and Evaluation Working Group workshop • Oral history project	• Inclusion Initiative	• Critical data studies session at 4S conference • Algorithms in Culture conference • Interviews and observations of BIDS	• Algorithms and Culture conference • Guest lectures • Science without Borders partnership • TechWomen outreach
New groups			• ImageXD, TextXD	• ImageXD, TextXD, VizXD, GraphXD

Source: progress reports and renewal proposal narratives

Accomplishments

Research productivity and follow-up funding support

The number of publications by BIDS participants nearly doubled between 2015 and 2017, from 71 to 141 (Exhibit 5). Perhaps more notably, datasets and software tools were the second most frequently reported product, and their number also nearly doubled, from 25 to 47.

EXHIBIT 5: NUMBER OF OUTPUTS INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
Publications	71	103	141
Books	4	3	3
Preprints	6	9	24
Educational materials	3	0	0
Datasets and software	25	37	47
Other	8	31	44
Total	117	183	259

Source: annual progress reports, individuals spreadsheet, outputs tab.

BIDS researchers were also successful in fund-raising, winning 40–50 individual grants per year and nearly $83 million in 2017, a $30 million increase from the previous year (Exhibit 6). The federal government was the largest funder (contributing 35–40% of the total depending on the year), but BIDS also received support from nonprofits (23–30%), the university (20%), and industry (9–14%). Approximately 20% of the grants reported in 2016 were collaborative submissions.

EXHIBIT 6: FUNDING SUPPORT INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
N institutional grants	3	0	1
Total institutional funding	$8,900,000		
N reported individual grants	52	43	53
Total individual funding	Not reported	$52,615,062	$82,963,305

Source: annual progress reports, individuals spreadsheet, awards/grants tab, institutions spreadsheet, grants tab. Awards are attributed only to the year in which they were first awarded. The total value of the grant is attributed to the first year. The funding amount was not available for all grants.

Development of the ecosystem of tools and practices

As shown in Exhibit 5, BIDS reported between 25 and 47 datasets and software tools per year. These included contributions to the increasingly popular Jupyter notebooks as well as to the project called rOpenSci, a repository of open-source R software tools. Yet another example of products developed by BIDS staff is Binder,[8] a tool that enables scientists to create interactive, shareable computational environments. BIDS also plays a central role in the maintenance and continued development of many of the core projects in the scientific Python ecosystem. For example, members of BIDS serve as project leads, release managers, core developers, and members of the steering committees of NumPy, SciPy, Matplotlib, scikit-learn, scikit-image, and NetworkX. They also play key roles in community events such as the annual SciPy conference. Currently, BIDS leads a major effort to modernize NumPy, which is the core library underlying many scientific Python projects. The contribution of BIDS to software development is described in more detail in Chapter 8.

In addition to these tools, BIDS spearheaded the publication of the book, *The Practice of*

[8] https://mybinder.org.

Reproducible Research,[9] which describes 31 case studies drawn from all three MSDSEs. Each case study follows a common narrative structure, which lays out the project workflow and discusses the strategies for increasing reproducibility as well as common challenges. Fellows have published two other books, *Effective Computation in Physics: Field Guide to Research in Python* and *Elegant SciPy*, with O'Reilly.[10]

Finally, research scientists at BIDS organized and hosted a week-long "Docathon" – an event focused on software documentation, which led to a paper about the role of documentation in the open-source community. BIDS has organized and hosted numerous other developer sprints including a recent joint sprint for scikit-image, scikit-learn, and dask. In collaboration with the Data Science Studies WG, a paper studying the Docathon and documentation was subsequently published, which has received multiple awards.[11]

Institutional change

An overarching goal of the data science environment program is to facilitate changes at the grantee institutions to make them more hospitable to data science and data scientists. In the course of the evaluation, we looked for the emergence of these shifts, while being aware that they take time. One of the most frequently mentioned institutional impacts of BIDS is the new undergraduate course, *Foundations of Data Science*, known as "Data 8," which was developed with the contribution of BIDS staff and which is expected to change the way data science is taught at UCB. (The course was also adopted by other universities.) While acknowledging that formal education is not part of the MSDSE mission, BIDS participants and the senior university administration believed that the course increased the visibility of the center and helped build its interdisciplinary community.

According to many individuals interviewed, including the university leadership, BIDS catalyzed the formation of the new Division of Data Science and Information in spring 2017, which represents a major reorganization of UCB.[12] BIDS will be incorporated in this division as a research center,[13] although the practical details of what this transition means for BIDS, if anything, are currently unknown. Before the inclusion of BIDS was announced, some center participants expressed a concern that its unique character might be lost in the division. Finally, BIDS produced some positive results for possible career paths in data science. These include research scientist positions, which BIDS was looking to increase; joint postdoc positions; and a tenure-track faculty appointment for a BIDS research scientist.

---

[9] J. Kitzes et al. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences.* University of California Press. First edition (October 17, 2017).

[10] https://github.com/elegant-scipy/elegant-scipy

[11] R. Stuart Geiger, Nelle Varoquaux, Charlotte Mazel-Cabasse, and Chris Holdgraf. *The Types, Roles, and Practices of Documentation in Data Analytics Open Source Software Libraries*. Computer Supported Cooperative Work (CSCW) (2018): 1-36. https://doi.org/10.1007/s10606-018-9333-1

[12] https://news.berkeley.edu/2018/11/01/berkeley-inaugurates-division-of-data-science-and-information-connecting-teaching-and-research-from-all-corners-of-campus/.

[13] https://data.berkeley.edu/research/research-centers.

Evolution and sustainability

Within two years of its launch, BIDS established itself as a data science hub on campus. Many individuals interviewed, from graduate students to the most senior university leadership, believed that BIDS had been very successful in community-building, and that it would be a big loss to UCB if it faltered. However, some weaknesses of the BIDS model also emerged. We heard from multiple sources that its mission was unclear and, perhaps consequently, that it had not positioned itself for long-term sustainability. While BIDS was helpful to many members of the Berkeley community, its specific research contributions were difficult to articulate.

During our last visit in spring 2018, BIDS was actively working on boosting its research program, as evidenced by the creation of the new position of CRO. BIDS had tried to pivot in this direction for at least a year, but the departure of the executive director and delays in hiring his replacement temporarily put these efforts on hold. With both of these staff on board, BIDS has begun ramping up its research activities – for example, in April 2018 the center issued the first call for internal research proposals.[14] In the past year, BIDS has also become more focused on its longer-term mission and funding strategy. Both the executive director and the CRO saw fundraising as their key duty, and fellows and data scientists noted increased expectations from the leadership in this area. BIDS staff mentioned several current sources of support as potential avenues for longer-term financial stability. One was to continue and possibly expand the hosting of "free" or jointly funded postdocs. In addition, BIDS research scientists were becoming more involved in large and well-funded tool development projects such as Jupyter and NumPy, which partially covered their time. Finally, the BIDS cost extension proposal mentioned $350,000 in support from *Siemens* and *State Street* per year and a gift of $100,000 from an unspecified source, also expected to be renewed annually.[15]

[14] https://bids.berkeley.edu/news/bids-announces-2018-bids-call-data-science-research-projects.

[15] BIDS DSE proposal. August 2016.

Chapter 4: Center for Data Science

In this chapter, we describe the MSDSE established at NYU, including its physical space and leadership, participant experiences, accomplishments, and plans for sustainability.

Key findings:

- MSDSE funding helped shape the nascent CDS as it established itself as a research institute
- Faculty, fellows, and postdocs benefited from the rich intellectual environment of the center
- Faculty reported lack of time as the main barrier to participation, but some also experienced limited support for involvement from their home departments
- The fellows were satisfied with their duties and did well on the job market; some directly credited MSDSE with their success
- Participants published many papers, contributed to the ecosystem of tools and practices, and obtained follow-up funding
- CDS used its partnership with the NYU library to disseminate tools and practices to aid reproducible research and open science
- Due to the leadership change at CDS, the future of the MSDSE was unclear at the time of writing this report.

Space and leadership

NYU's MSDSE is physically and administratively based within the CDS. Initially located in a widely criticized temporary space, about two years ago CDS moved to its newly renovated, spacious quarters at 60 5th Avenue in Manhattan. The space, which was designed with input from UW and UCB and incorporated their experiences, is divided into a "quiet floor" for focused work and a "loud floor" for events and collaborations. The positive role of the space was highlighted in many interviews.

> *Space is really important at NYU because it is a campus in the city. It is less likely that you stumble into a colleague from physics and have a meaningful conversation with them, unless it just so happens that they are in the same building as you. CDS promotes that because now there is a place to go and you are more likely to run across people in other areas.* (Faculty member)

After some initial turmoil, the NYU MSDSE has been led by the same faculty executive director since 2015. In 2016, the MSDSE added a full-time program manager and an outreach coordinator to its staff. Unlike BIDS and eScience, the leadership team of this MSDSE does not include a non-faculty executive director, which was viewed by some participants as a weakness of the governance structure. Similarly to the other two centers, this MSDSE has a 12-member Steering Committee, which makes all key decisions. The NYU MSDSE and CDS share space, faculty, staff, and other resources, but CDS has its own director. Some sources referred to the

MSDSE as a research arm of CDS,[16] and in many conversations the two names were used interchangeably. For simplicity, we call the NYU MSDSE "CDS," unless the distinction is clear and bears on the narrative.

Participants

The CDS staff included faculty, data science fellows, postdocs, and research engineers. The MSDSE-affiliated staff come from various parts of the university, and the faculty of CDS proper (from the point of view of NYU) are all joint appointments. The center is particularly proud of its fellows program. Selected through a highly competitive international search, the fellows are hired for 1-3 year terms. Most fellows were fully salaried through the MSDSE, though one was funded fully through a faculty member's NSF grant and given the honorary title of MSDSE Fellow. In their level of independence, these researchers are more similar to assistant professors than postdocs. Postdocs are recruited both internally and from outside of NYU and co-funded with another entity at the university. This track was created to increase the visibility of the center and to build bridges; the CDS leadership told us during the site visit in May 2018 that the program had accomplished its goals and was being phased out. Finally, research engineers are fully or partially funded by CDS and MSDSE.

Based on progress reports, the number of MSDSE participants nearly doubled between 2015 and 2017, with faculty and postdocs increasing by the largest margin (Exhibit 7). The number of participants by title was as follows: professors 17-27, data scientists 3-5, graduate students 1-7, postdocs 10-19, and professional staff 1-2. An examination of the disciplinary focus of the MSDSE researchers revealed that Computer and Information Sciences were the most common participant fields, reported by 16-31 researchers depending on the year, followed by Social Sciences, reported by 7-16 researchers (Exhibit 8).

EXHIBIT 7: NUMBER OF PARTICIPANTS BY JOB CATEGORY INCLUDED IN PROGRESS REPORTS

Job title	2015	2016	2017
Assistant/associate/full professor	17	26	27
Data scientist/research scientist/software engineer	3	5	4
Graduate student	1	2	7
Postdoctoral fellow	10	13	19
Professional staff	2	1	2
Other	0	2	0
Total	33	49	59

Source: annual progress reports, individuals spreadsheet, metadata tab.

[16] For example, the NYU MSDSE webpage describes CDS (http://msdse.org/nyu/).

EXHIBIT 8: FIELDS OF PARTICIPATING RESEARCHERS INCLUDED IN PROGRESS REPORTS

Job title	2015	2016	2017
Mathematical sciences and statistics	2	4	2
Geosciences	0	0	0
Engineering	1	0	0
Computer and information sciences	16	20	31
Physics and astronomy	4	4	6
Psychology	1	2	2
Social sciences	7	16	14
Life sciences	2	3	3
Other	0	0	1
Total	33	49	59

Source: annual progress reports, individuals spreadsheet, metadata tab.

Faculty experience

Nature and degree of participation

We interviewed 10 faculty members about their level and nature of participation in CDS. Of these, half were involved with the center since the grant's planning phase; and the rest were either recently hired by NYU or were employed at NYU, but joined the center as supervisors of their postdocs. The faculty interviewed were involved with the center in multiple ways, such as attending talks or workshops (4), serving on committees or working groups (3 in each), participating in the Seed Grant program (3), attending the Data Summit (2), supervising postdocs (2), and using MSDSE-developed tools (2). Most faculty reported that their involvement had been steady or increased over time. Finally, several faculty said that NYU tends to protect junior faculty from administrative work. As a result, setting up and running MSDSE working groups and activities fell to senior faculty.

> *The senior faculty take the burden of community building and the junior faculty don't have enough institutional power yet to get institutional change…We definitely want to change things but you can't make institutional changes until you get to that level of power.* (Faculty member)

As noted in other data collection efforts, NYU was delayed in hiring a project manager and some faculty reported becoming overburdened by the administrative load early in the grant. Additionally, there was turnover in leadership at both MSDSE and CDS when each entity was getting established. These combined challenges led to some fatigue and attrition among a few early participants in this MSDSE. As one leader explained in 2017:

> *Like all places, we have issues with faculty churning. Most people get a month of salary and the commitment is a lot more than that. People who stuck around are heroes. They do this because they believe it is an important effort.* (Faculty member)

There was uneven awareness of these start-up challenges among faculty members who became involved later and excessive administrative burden did not appear to be a concern during the last site visit in 2018.

Benefits of participation

Virtually all faculty interviewed credited the MSDSE with extending their professional networks, both internally at NYU (5) and with the other MSDSEs (4). Several respondents also enjoyed the stimulating intellectual atmosphere of the Data Summit and described the weekly seminars organized by the fellows as a highlight of the MSDSE programming. Two faculty members are connected to the MSDSE through a postdoc; one of these said that that the tools created by the postdoc through an association with the MSDSE led to important scientific development in the laboratory. We also heard that spending time in the CDS space, either to attend events or to work, created a collaborative community that allowed faculty members to increase their productivity (5) and expand their research programs (3). Faculty reported that working in the space for varying amounts of time – from several hours a week to daily – led to collaborative projects both at their own MSDSE and with the centers at UW and UCB.

> *One thing that is kind of an intangible of the grant is that it has created a community of scientists across the three universities…I have had far more contact with Berkeley and Seattle than before. And when I go to those places I feel at home. I have a desk where I can sit at, I have people to hang out with, I have collaborations across the institutions, and so on.* (Faculty member)

Participation challenges

The greatest challenge reported by faculty (9 of 10), was constraints on their time. Most respondents viewed their administrative duties at MSDSE as additional to their departmental requirements, making it difficult to balance research and service. Some said that their chairs viewed MSDSE participation as taking time away from their departments. Nevertheless, the faculty continued to be involved in the MSDSE because they were personally invested in its mission and found the connections to other members of the community rewarding. Two junior faculty members stressed the difficulty of serving in formal MSDSE roles (e.g., on the steering committee or working groups) while pursuing tenure. Some faculty members believed that they were doing most of the work related to the center's administration and therefore should have access to discretionary funds to support their research and/or to buy themselves out of teaching.

Non-faculty staff experience

Nature and degree of participation

Fellows and postdocs at CDS focused primarily on their own research program. While not expected to participate in working groups, these researchers were encouraged to contribute to the center in other ways – by organizing seminar series or workshops, screening master's program applications, or teaching master's students. These staff characterized their responsibilities as both interesting and helpful to their careers without being burdensome. One fellow was willing to increase his contribution to the center to free up funds to hire more fellows. Some fellows appeared uncertain about the expectations for their role, particularly early on in the existence of the center:

> *We never really had a job description and so my offer letter said vague things about helping CDS, which was interesting when we were asked last year to fill out evaluation forms. We didn't know what criteria we were being evaluated on...* (Fellow)

Research engineers were generally responsible for providing support to other staff at and outside of CDS. Their specific duties varied from maintaining software to working on projects identified by the MSDSE leadership. Some research engineers also taught courses and developed/disseminated good data science practices.

> *I'm supposed to work with other scientists across the university to help them with their software engineering needs to build structures or implement algorithms and to train other researchers in software engineering if they want these skills.* (Research engineer)

Benefits of participation

All fellows, postdocs, and research engineers said that they benefited from the interdisciplinary environment of CDS. The fellows also highlighted the freedom to chart their own research path, which they viewed as both a rare opportunity and an important career benefit. Finally, access to NYU housing was viewed as a major benefit.

> *Having this experience and working with different types of people and on these different types of problems helped me shape myself as a strong researcher.* (Postdoc)

> *What was very useful and important about this fellowship, was that it was independent. The research I've done in the last few years, I was the lead and PI and that helps makes the case of my work as a reflection of what I will do in the next few years.* (Fellow)

Challenges of participation

Before the move, the biggest problem identified by the participants was the physical space, which was resolved to everyone's satisfaction. In contrast, mentorship-related challenges persisted through the course of the evaluation. Initially, mentorship was primarily informal, which worked well for researchers who were either very independent or felt comfortable approaching faculty if they needed help or advice. At the same time, some fellows and postdocs experienced problems identifying faculty mentors with similar interests who were willing to supervise them. The MSDSE leadership acknowledged the need to implement a more formal mentorship approach in the 2015 progress report, and all fellows joining CDS were assigned a mentor. However, the results appeared mixed:

> *We were assigned a mentor when we arrived. Some people had mentors in the same discipline, some were assigned some junior or senior professors doing work not directly related to what we were doing. There wasn't much communication about our responsibility as fellows with what to do with your mentor in particular.* (Fellow)

CDS leadership told us during our 2018 site visit that they tried to address this problem by selecting fellows and postdocs who were well-matched to existing faculty.

Another challenge mentioned by the fellows was inconsistent communication from the leadership and lack of participation in the governance of the center. Presumably to address these concerns, CDS began to offer monthly community breakfasts in 2017. These challenges were not mentioned during our site visit the following year.

> *I would have liked more communications between the MSDSE steering committee and Moore Sloan fellows…We have all this freedom and are expected to run our own research programs*

and collaborations, but within the university we are still postdocs and have no sort of institutional power at all. (Fellow)

Role of MSDSE in career progression

Fellows and postdocs at CDS were successful on the job market: all but one were able to secure a faculty position and one became a data scientist in industry. Some of the researchers directly attributed their success to MSDSE participation:

> *I would have never been considered or had had the research required for moving to* [university name redacted] *and having interdisciplinary work with computational methods without my fellowship.* (Fellow)

> *Whenever I told people I was working at the MSDSE, that helped people see me as a data scientist and that certainly helped me when looking for jobs.* (Postdoc)

Several postdocs mentioned that the interdisciplinary nature of their work raised some questions at traditional academic departments, but that they were able to articulate the benefits of their MSDSE participation. In some cases, the experience of being part of the center persuaded postdocs to stay in academia, at least for the time being. However, many researchers were content to move to industry if maintaining their focus on data science in academia became impossible, and some noted that the difference between the two tracks was exaggerated. The long-term career path for research engineers was less clear. These staff were not interested in a tenure-track position, but enjoyed the intellectual freedom and flexibility of the university. CDS leadership was working on developing funding models for more permanent research engineer positions, but we are uncertain about the results of these efforts. As of 2017, this MSDSE had a commitment from CDS, approved by the provost, to support two research engineers for the foreseeable future.[17]

Activities

The Methods Working Group, unique to the NYU MSDSE, and the Reproducibility and Open Science Working Group were the most active at this site, attracting people from within and outside of the center. In contrast, the Software and Education Working Groups did not gain much traction. The MSDSE leadership speculated that members of the Software Working Group were unable to find projects of common interest and lacked manpower for broader campus outreach. In contrast to BIDS and eScience, undergraduate education did not appear to be a priority for this center, and the Education Working Group remained small. The full list of activities by working group is shown in Exhibit 9.

[17] 2017 NYU MSDSE progress report.

EXHIBIT 9: ACTIVITIES OF WORKING GROUPS INCLUDED IN PROGRESS REPORTS

	2014	2015	2016	2017
Career Paths and Alternative Metrics		• New protocol for joint hires • Sponsored a meeting to form Text as Data Association	• Mentorship program for joint faculty • Outreach Coordinator position	• NYU housing for new fellows and postdocs
Education and Training	• Concepts and Categories talk series • Reading group in cognitive science • Software Carpentry • Weekly DS seminar series	• New classes in DS • Appointed MSDSE researchers as guest lecturers • Python tutorials • Astro Hack Week • Introduction to Text Analysis Using R Workshop	• Tutorials on GitHub, data management • Reproducibility office hours • Cosponsored DS and Social Science workshop • Cosponsored the Atlantic Causal Inference Conference	• PhD program launched • MS capstone project restructured • Creation of tracks within MS degree • Consolidated course offerings in DS • White papers based on hack weeks
Software Tools, Environments, and Support		• Incubator projects • Guest lectures to promote open-source tools	• Incubator projects	• Created Data Clinic • Released Data Polygamy and Urban Pulse software • Shadows of New York project
Reproducibility and Open Science	• Document on best practices in reproducibility • Developed ReproZip and noWorkflow • Redesigned reproducibility evaluation for ACM SIGMOD	• Created ReproMatch • Indexing software workshop • Inclusion of reproducibility in courses • Office hours • Talks and tutorials • Developed reproduciblescience.org and ReproZip	• Training on reproducibility and data management • Training modules available on GitHub	• Training on reproducibility and data management • ReproServer platform • Yadage and packtivity tools
Working Spaces and Culture	• Designed temporary space • Began design of permanent space	• Designed permanent space • Monitored space use • Installed and tested wormholes	• Moved into new space	• Space usage survey
Data Science Studies	• Community building exercises • Survey on definitions of ds/scientists • Course on data and society	• Collaboration network project • Weekly newsletter	• Weekly newsletter	• Critical Data Studies Thinking Group • Weekly newsletter • Syllabi for three new courses • Ethics module for the DS course
Methods	• Causality reading group • DS showcases	• Scikit-Learn tutorial • Text as Data seminar • DS showcases • DS seminar series • Causal Inference WG	• Text as Data seminar • DS showcases • DS seminar series • Causal Inference WG • Seed grants	• Text as Data seminar • DS showcases • DS seminar series • Causal inference WG • Seed grants

Source: progress reports and renewal proposal narratives

Accomplishments

Research productivity and follow-up funding support

Review of progress reports revealed that publications were the most frequently cited outputs at CDS (Exhibit 10). The number of publications more than doubled between 2015 and 2016, from 72 to 145, and remained similar in 2017 (N=148). Datasets and software were the second most common type of product reported: 53 in 2015 and 44 in 2016 and 2017.

EXHIBIT 10: NUMBER OF OUTPUTS INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
Publications	72	145	148
Books	0	2	6
Preprints	13	35	35
Educational materials	5	0	0
Datasets and software	53	44	44
Other	20	22	28
Total	163	248	261

Source: annual progress reports, individuals spreadsheet, outputs tab.

The number of grants awarded to CDS researchers increased from 27 in 2015 to 44 in 2017 (Exhibit 11), with the total amount of funding received in 2017 exceeding $65 million. The federal government was the main funder, at 40% to 60% of the total, depending on the year, but the center also received significant support from foundations and industry (18% and 27% of the total in 2017, respectively). The relative contribution of industry nearly doubled over three years, from 14% to 27%. Approximately 25% of the grants reported in 2016 were collaborative submissions.

EXHIBIT 11: FUNDING SUPPORT INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
N institutional grants	1	0	0
Total institutional funding	$744,189		
N reported individual grants	27	35	44
Total individual funding	Not reported	$31,709,963	$65,094,415

Source: annual progress reports, individuals spreadsheet, awards/grants tab, institutions spreadsheet, grants tab. Awards are attributed only to the year in which they were first awarded. The total value of the grant is attributed to the first year. The funding amount was not available for all grants.

Development of the ecosystem of tools and practices

The unique attribute of this MSDSE is its strong partnership with the NYU libraries, made possible by a staff member with a dual appointment, who offers tutorials and consultations related to reproducibility. Similar to the other MSDSEs, CDS staff also developed many software tools that can be used in a range of applications. These include ReproServer, which runs computations from a web browser; OpenSpace, which visualizes planetary features; and TopoAngler, which separates and reconstitutes MRI images. Finally, a MSDSE research engineer collaborated with other departments at NYU to create a tool for improving coordination across research groups called the Standard Cortical Observer. Additional examples of CDS contributions in this area are described in Chapter 8.

Institutional change

When asked for examples of institutional change that could be attributed to the MSDSE, the center's leadership described a successful process for hiring joint faculty. Many respondents also thought that the center was at least partially responsible for the growing enthusiasm for data science on campus, which manifested itself through an increase in the master's program applications, a growing attendance at the MSDSE events, and an interest in placing postdocs at CDS (in some cases with full funding support from departments). We also heard that more students and faculty across NYU interested in collaborations involving big data had increased, and that some of these collaborations led to joint papers and grant applications from researchers across departments. Finally, the university made a large financial commitment to support data science in the form of space, faculty lines, housing for fellows, and other resources.

Evolution and sustainability

Both CDS and MSDSE had a somewhat rough start with multiple leadership changes and inadequate space. However, the two entities joined forces to emerge as a vibrant research institute based in well-designed and spacious quarters; and staffed with high-caliber faculty, fellows, and research engineers. Multiple participants credited the MSDSE grant with this transformation. At the time of the site visit in May 2018, sustainability plans for the NYU MSDSE appeared to be linked to CDS. It was our impression that the revenue generated through the master's program, combined with some university funding, ensured that CDS was financially secure and could support all or most of the MSDSE staff and programs. However, we are uncertain whether these plans will remain in place with the new director appointed at CDS. The MSDSE also explored avenues for sustainability that were independent of CDS. These included hosting postdocs fully or partially funded by departments, co-funding staff positions with the library, identifying industry sponsors, petitioning the provost to fund seed grants, and applying for federal funding.

Chapter 5: eScience Institute

In this chapter, we describe the data science environment at UW. We begin with an overview of the space and leadership, followed by description of participant experiences and accomplishments, and conclude with a discussion of sustaining the center beyond the MSDSE grant.

Key findings:

- Established approximately five years prior to the MSDSE grant with $1 million per year of permanent funding from the Washington State Legislature, eScience is a mature center with a long-term sustainability plan
- MSDSE funding was essential to achieving the momentum and validation that led to significant additional investments in core data science infrastructure
- eScience created a promising career track for data scientists
- Faculty, staff, postdocs and students, from eScience and across campus benefited from the rich intellectual environment of the Institute
- Postdocs participate in working groups, teach software carpentry courses, and attend eScience events, but their primary focus is on research projects
- Most postdocs were successful on the job market or confident that the MSDSE experience would give them an advantage
- Participants published many papers, contributed to the ecosystem of tools and practices, and obtained follow-up funding
- eScience was crucial in ensuring that no individual department took over data science education, but rather that all contributed to teaching in a coordinated fashion.

The eScience Institute was launched in 2008 and at the time was staffed with three data scientists and a senior faculty director.[18] Over the next several years, the center secured a training grant from the National Science Foundation, some funding from the Moore Foundation, and several half faculty lines from the university, but these resources were inadequate to meet the data science needs on campus. The MSDSE grant was instrumental in enabling eScience to expand programs and reach maturity, but it did not fundamentally change its mission.

Space and leadership

In January 2015, eScience moved into its "WRF Data Science Studio", which was a former library space (similar to the BIDS space) renovated in 2014 with funds from the Washington Research Foundation (WRF). The open and modular layout of the space was intended to signal inclusivity to the university community and to foster collaboration. Many participants spoke about the importance of the Institute's physical location, which was intended to convey that it was a "neutral space" outside of traditional departmental boundaries. The efforts of eScience to be

[18] In 2008 the term "eScience" was commonly used instead of the term "data science"; the UW eScience Institute has chosen to stick with its original name rather than re-brand itself.

inclusive were successful, and the Studio quickly became too small to accommodate all events and visitors.

> *One thing we talk a lot about and I think has been verified, is that having a neutral space on campus is important. We're not viewed as part of the computer sciences department or another department in particular. There's this Switzerland effect, when you are outside of the departmental silos. People come here and are more likely to collaborate across disciplines than they might if they were all going to somebody's particular department.* (Site visit 2017)

The eScience core leadership team includes the faculty director, two executive directors, and a director of research, who are supported by the Executive Committee and the Steering Committee. In 2017, the inaugural eScience director stepped down. The new director was chosen in a systematic search from within the eScience community, and the transition appeared to be seamless. The executive director position is shared by two PhD-level scientists, who proposed this arrangement to the director; this unusual arrangement is viewed as highly successful. The number of full-time equivalents for this position ranged from 1 to 1.5, depending on the needs of eScience. Many participants interviewed said that the executive directors were very effective, and that their scientific training enabled them to credibly represent the center and understand staff needs and aspirations. Finally, the core leadership team includes the director of research (this position replaced an associate director), whose main responsibility is to supervise data scientists, but who appears to be involved in many other activities at the center.

All major decisions about funding and personnel are made by the Executive Committee, which is composed of faculty from various departments and the core leadership team. Several members of the committee are well-known faculty, which adds gravitas to eScience and increases its visibility on campus. We were told that the governing style of the center is very equitable and that the Executive Committee speaks with one voice. Finally, eScience includes a Steering Committee with an additional 15–20 faculty members. While initially helpful to broaden the reach of the center, the level of engagement in this committee has been declining and eScience is reconsidering its utility. The stability of engagement among core faculty and staff has been important to the success of eScience, as has been the pooling of diverse financial resources and their management by a broad-based Executive Committee.

Participants

Non-faculty participants at eScience include postdocs, data scientists, and research scientists, as well as PhD students funded by three non-MSDSE training grants. Postdocs are recruited externally for a two-year term, with the option of a third year based on need and performance. They are supported by a mix of funds from MSDSE and WRF, and are selected from among the applicants by a group appointed by the Executive Committee – one of many examples of the pooling of resources and their broad-based management. These researchers have two mentors, a domain scientist and a methodologist. eScience negotiated standard salaries and benefits for postdocs with participating departments to promote equality.

Data scientists are chosen through a national search and are fully salaried through eScience. These staff are foundational to the center and are heavily involved in all of its activities and programs. They also maintain a research program. To recruit and retain high-caliber data scientists, eScience offers higher salaries than typical for academia. In addition, data scientists

are hired into positions that give them a PI status to help apply for independent funding, and a portion of the funding they bring in is returned to them as a research stipend. Research scientists are joint with departments, on a 50/50 or 25/75 basis and are usually recruited internally, but in other respects are similar to data scientists (some respondents described them as interchangeable).

According to progress reports, the total number of participants in eScience reached nearly 100 in 2016 and remained at this level the following year (Exhibit 12). The number of staff increased from 7 to 12 for data scientists, from 15 to 18 for graduate students, from 13 to 19 for postdocs, and from 4 to 7 for professional staff. In contrast, faculty participation declined from 33 in 2015 to 24 in 2017.

EXHIBIT 12: NUMBER OF PARTICIPANTS BY JOB TITLE INCLUDED IN PROGRESS REPORTS

Job Title	2015	2016	2017
Assistant/associate/full professor	33	31	24
Data scientist/research scientist/software engineer	7	8	12
Graduate student	15	18	18
Postdoctoral fellow	13	15	19
Professional staff	4	7	7
Other	0	1	0
Intern	16	16	16
Total	88	96	96

Source: annual progress reports, individuals spreadsheet, metadata tab.

An analysis of self-reported participant disciplines revealed that Computer and Information Sciences were the most common, reported by 20–25 participants, depending on the year (Exhibit 13). This was followed by Mathematical Sciences and Statistics, Life Sciences, and Social Sciences, at around 15 in each category. Engineering, Physics and Astronomy, Geosciences, and Psychology were less common.

EXHIBIT 13: FIELDS OF PARTICIPATING RESEARCHERS INCLUDED IN PROGRESS REPORTS

Job title	2015	2016	2017
Mathematical sciences and statistics	16	12	13
Geosciences	9	9	8
Engineering	4	6	10
Computer and information sciences	23	25	20
Physics and astronomy	9	9	11
Psychology	0	1	1
Social sciences	14	17	16
Life sciences	13	15	16
Other	0	2	1
Total	88	96	96

Source: annual progress reports, individuals spreadsheet, metadata tab.

Faculty experience

Nature and degree of participation

Of the nine faculty interviewed, seven became involved in eScience after the MSDSE grant was awarded. Three of the nine reported reducing their participation in the center over time due to competing professional and personal responsibilities. The most common affiliation with eScience was through supervising fellows (6 of 9), of whom the faculty spoke extremely highly. About half of the respondents were involved in the working groups, although some no longer attended meetings due to teaching conflicts, and one participated in an incubator project. Three faculty members were on the Steering Committee and one was part of the search for the new eScience director. All faculty members spent some time in the Studio to attend events, hold office hours, and work on projects.

Benefits of participation

The faculty cited several benefits of participating in eScience. The majority (5) reported that it fostered new collaborations, especially with postdoctoral fellows. Three faculty members said that eScience helped them recruit students to their laboratories and one even cited it as one of the reasons he came to UW:

> *Rather than me recruiting them into my lab, I'm recruiting them into a community that's broader than that and is kind of unique.* (Faculty member)

> *Would I have recruited postdocs who are doing machine learning and machine vision into my lab without myself having the chops in that space? I doubt it.* (Faculty member)

> *eScience…was very helpful, and not a small part of why I ended up at UW.* (Faculty member)

About half of the faculty members interviewed had changed their research direction as a result of their participation in eScience:

> *I think this project is going to be one of my main research focuses for the next five, ten, fifteen years. It's amazing to have the technical expertise brought by MSDSE fellows to help me get it off the ground, and to bring in a whole new set of technical skills to my lab.* (Faculty member)

> *Data science thinking became more central to the lab and has shifted the nature of the research we do.* (Faculty member)

Finally, the faculty who were involved in the Education Working Group said that the center provided a neutral and welcoming environment to develop new programs for students:

> *We would bring all these departments and discuss education in the Data Science Studio, not in computer science or in another department. So we were in a neutral space, a neutral working group where everyone's ideas were welcome, and this way we could come up with a UW-wide program.* (Faculty member)

Participation challenges

Almost all of the UW faculty interviewed (7) said that lack of time was the major barrier to participating in eScience. Two social scientists also told us that some of their departmental colleagues questioned whether eScience benefited their careers. To increase faculty involvement in eScience, the respondents recommended trying to further raise awareness of

the center among students and faculty, and to offer tangible incentives such as teaching relief or administrative support.

Non-faculty staff experience

Roles and responsibilities

Postdocs at eScience are expected to be part of the community, but have the flexibility to choose the nature and level of their contribution, which was appreciated. In interviews, postdocs told us that they participate in working groups, teach software carpentry courses, and attend eScience events, but that their primary focus is on their research projects. Data scientists are the backbone of eScience. They hold office hours, run the DSSG and the Incubator program, teach courses, lead some of the working groups, contribute to proposals, and organize hack weeks and other workshops. Research scientists have similar duties, scaled down to their level of affiliation with eScience. Data scientists and research scientists are also involved in research projects, which in some cases were launched during office hours or through incubators. Data scientists and research scientists told us that their number was sufficient to meet the needs of the center, while having the flexibility to choose the activities that matched their interests.

Benefits of participation

All data scientists, research scientists, and postdocs praised the diverse intellectual environment of eScience and enjoyed being part of the culture which valued non-academic contributions. Finally, one research scientist said that having half of his time covered through eScience offered the freedom to engage in more exploratory research and decreased the pressure to apply for funding.

> *I've probably learned more in the last 3 years than I have most of my career around computational methods, data science tools, working in the cloud …I mean, just sitting in this room, day after day, you get exposed to so much innovative work*. (Research scientist)

> *I liked the fact that there would be rewards for spending time on the production of tools that other people would use. That is such an unusual aspect of the data science environment. That is typically something that people don't care about and is counted against you, so to have that counted as a plus was fantastic* (Postdoc)

> *The advantage is really amazing. I am a soft money scientist, so to have 50% of my support here from eScience, which is contractual but more stable than the up and down world of writing grants all the time, has allowed me to be more exploratory and try out new things without worrying too much about it being linked to a specific grant.* (Research scientist)

Challenges of participation

Several postdocs told us that while they felt individually connected to eScience, they did not have their own community. Postdocs spend little time at the Studio, because the space is limited and can be noisy and in some cases is distant from their department.

> *The only expectation for us was to spend one day a week in the data science environment. That part didn't work out as well as I would have liked... It is a great space to meet people in. It is not necessarily a great space to sit down and do an entire day's work in* (Postdoc)

eScience made an effort to attract postdocs to the space by adding monitors and keyboards to all desks, opening private rooms to reservation, and offering weekly lunches and food before

seminars. The availability of easy-access desk space led some postdocs to work regularly in the Studio, but the lunches drew the largest regular attendance from both postdocs and students. Some postdocs questioned the utility of dual-mentorship model, as their methodological mentors often played a peripheral role in their research program. The leadership of eScience were aware of this problem and were taking a more active part in pairing postdocs with mentors.

Data and research scientists told us that because they were involved in many different activities to support the center, it was sometimes difficult to carve out time for their own work. Research scientists with joint positions experienced this challenge more strongly, especially if they were involved in the time-consuming DSSG program. Finally, some researchers appeared uncertain whether they were meeting the leadership's expectations. This led to greater focus on articulating the institute's strategy and to regular one-on-one meetings between the director of research and data scientists.

Role of MSDSE in career progression

All postdocs were interested in staying on an academic track, and only planned to move to industry if they could not secure a faculty position. Most had either been successful in finding a position of their choice or confident that the MSDSE experience would give them an advantage on the job market. However, a small number of postdocs interviewed either had experienced difficulties with an academic job search because of the interdisciplinary nature of their projects or were concerned that this might happen. As noted in the BIDS profile, these challenges probably reflect the academic culture rather than weaknesses of eScience.

> *I think my MSDSE postdoc was viewed positively. I think that the MSDSE name isn't known, necessarily, so what is perceived positively is the range of projects that you've worked on or the range of people that you have worked with. You do have a greater experience talking to people outside of your narrow subfield as a result of the broadness of the MSDSE cohort.* (Postdoc)

Most data scientists and research scientists had been planning to move to industry, but having the MSDSE positions allowed them to stay in academia. Some said that the source of support or title did not matter to them, as long as they could continue doing the work they enjoyed. However, some of these staff were concerned about their long-term prospects. We also hard a view that there was no clear path to promotion within the data/research scientist career track, even if the positions were made permanent. Finally, this group was interested in clearer job expectations and additional career guidance and mentoring, and some respondents were hopeful that these concerns might be addressed by the new research director.

> *If it weren't for this position I would not still be doing academic work, because I'm finding over time that I'm less driven by the normal academic goals and I'm really more interested in building open software tools that people use and that make their work more effective.* (Data scientist)

> *There was a period where I looked at the data scientist positions or the research scientist positions that are associated with MSDSE. But the funding scheme for that remains unclear to me in the five-year term. I was interested to see if that funding model would become clearer to me, but it remained less defined than I would have been comfortable pursuing.* (Postdoc)

More clarity on milestones would be good. One of the challenges with this kind of position, compared to faculty, is that faculty usually know what to have to do in order to say get tenure. But nobody ever took me aside for a conversation to say what I should do to keep my job in the long run. (Data scientist)

eScience hired a new director of research and I am hopeful. Mentoring for the data science and research scientists to help them figure out what to do strategically for themselves, their careers, it isn't something that is really addressed now, and it is hard because these are new jobs in academic research which means we need more mentoring not less. (Research scientist)

Activities

The working groups at this MSDSE are run by faculty and data scientists, with some assistance from postdocs. The groups are seen as helpful for integrating postdocs with the eScience community, raising awareness about eScience on campus, bringing in talent and ideas, creating a shared identity, and launching collaborations with other MSDSEs. Activities offered by eScience (Exhibit 14) were intentionally designed to accommodate people with a range of expertise, interests, needs, and desired level of engagement; the leadership of the center believed that this was key to its reputation for inclusivity and flexibility.

EXHIBIT 14: ACTIVITIES BY WORKING GROUP INCLUDED IN PROGRESS REPORTS

	2014	2015	2016	2017
Career Paths and Alternative Metrics	• Hired key personnel	• Fully staffed	• Created research scientist positions	• Leadership change • Established annual DS Career Fair • Established annual UW DS Summit, which became NW DS Summit
Education and Training		• Transcriptable graduate DS options approved in seven departments	• Undergraduate and graduate transcriptable DS options approved in additional departments • Established MS in DS • All curricula publicly available • Python tutorials publically available • New undergraduate graduate DS courses	• Established graduate DS option • Graduate and undergraduate transcriptable options approved in additional departments • Established a new series of pre-major undergraduate DS courses
Software Tools, Environment, and Support	• Software carpentry workshops • DS seminar series • Astro hack week • Python boot camp • Incubator	• Python seminar series • Summer seminar series • Software carpentry • DS workshop • Python boot camp • Cloud Day @UW • Incubator • DSSG	• Software carpentry workshops • Cloud usage tutorials • Python and other boot camps • Hack weeks • Community seminars • Incubator • DSSG	• Software carpentry workshops • Cloud usage tutorials • Python and other boot camps • Hack weeks • Community seminars • Incubator • DSSG

	2014	2015	2016	2017
Reproducibility and Open Science	• DS seminar series • eScience community seminar	• Reproducibility Badges project	• Seminar on reproducibility • Monthly Git/Github introductory class	• Community seminar on reproducibility • Monthly Git/Github introductory class
Working Spaces and Culture	• DS Studio opened	• Office hours	• Office hours • Student and postdoc weekly lunches	• Office hours • Coffee hour • Student and postdoc weekly lunches
Data Science Studies	• Developing mechanisms to assess job satisfaction	• Trace ethnography project	• Evaluation of DS courses and seminars	• Workshops at UW and UBC • Guest lectures
New groups			• Image XD	• Neuroinformatics WG

Source: progress reports and renewal proposal narratives

Accomplishments

Research productivity and follow-up funding support

We found that the eScience participants published 135 papers in 2015 alone, and that the publication output continued to increase, exceeding 200 in 2017 (Exhibit 15). The center also reported approximately 30 datasets and software products per year.

EXHIBIT 15: NUMBER OF OUTPUTS INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
Publications	135	131	201
Books	3	7	4
Preprints	9	5	22
Educational materials	3	0	0
Datasets and software	31	34	29
Other	16	16	15
Total	197	193	271

Data source: annual progress reports, individuals spreadsheet, outputs tab.

eScience also demonstrated a strong record of obtaining additional funding, which reached nearly $100 million in 2017 and included several institutional grants (Exhibit 16). Funding from the federal government accounted for approximately half of the total, but the center also received support from nonprofits (20–25%) and industry (12–18%). Approximately half of the grants reported in 2016 were collaborative submissions.

EXHIBIT 16: FUNDING SUPPORT INCLUDED IN PROGRESS REPORTS

	2015	2016	2017
N institutional grants	5	8	7
Total institutional funding	$12,687,817	$6,469,023	$14,702,757
N reported individual grants	55	39	53
Total individual funding	Not reported	$53,158,097	$82,708,964

Source: annual progress reports, individuals spreadsheet, awards/grants tab, institutions spreadsheet, grants tab. Awards are attributed only to the year in which they were first awarded. The total value of the grant is attributed to the first year. The funding amount was not available for all grants.

Development of ecosystem of tools and practices
Like other MSDSEs, eScience contributed to advancing open science and reproducibility. For example, the center's staff developed a tool called WideOpen to locate datasets overdue for publication and published guidelines for sharing software code in the prestigious and widely read journal *Nature Neuroscience*. In addition to their contributions to open science, eScience staff developed software packages with a broad range of applications, including simulating bionic vision, evaluating political events and trends, and detecting near-Earth asteroids. Additional examples of eScience software contributions are described in Chapter 8.

Career tracks and institutional change
eScience participants believed that one of their most important institutional accomplishments was establishing a viable career track for data scientists. While these positions predated the MSDSE grant, it enabled eScience to create additional support for these researchers and to increase their number to the point where they can staff all center programs while having the flexibility to choose how to divide their time. eScience also left a mark on student education: with considerable help from the Education Working Group, several departments created various undergraduate and graduate specializations in data science called "options." Finally, some departments began including data science as a research direction in their strategic plans.

Evolution and sustainability

As mentioned in the beginning of this chapter, eScience predated the MSDSE grant. One year after eScience was selected for the MSDSE award – and directly related to that selection – eScience received $9.3 million from the Washington Research Foundation for renovation, postdoc salaries, and faculty startup packages.[19] With this infusion of funds, eScience began a transition from a "bare bones" operation to a fully staffed center with a broad range of activities. Importantly, the MSDSE funding enabled eScience to experiment with its programs and staffing models, some of which proved very successful.

In a testament to its success, eScience recently secured additional funding from UW, which together with its state funding, totals approximately $1.75 million per year, an amount sufficient to cover most of its staff and programs at the present level. eScience plans to supplement this core budget so that it can continue experimenting with various components of the center. To raise additional funding, eScience is considering a shift from fully funded data scientists to co-funded research scientists. We were told that this model is not only more economical, it also helps build bridges and allows research scientists to find arrangements that best fit their career goals and research interests. The center's leadership also expected that its popular DSSG program would become financially self-sustaining with support from the partners.

[19] eScience progress report 2014.

Chapter 6: Joint MSDSE Activities and Learning

In this chapter, we discuss the added value of funding a cohort of three centers. While all centers received the same resources and were broadly organized around the same framework, each had the flexibility to implement its own staffing and programmatic models. To learn from these efforts in real time, the Foundations put in place several knowledge exchange mechanisms, which included joint MSDSE progress reporting, annual data summits, and regular phone calls between the funders and the centers.

Key findings:

- The collaborative and flexible funding model helped each grantee develop unique environments that best fit their local context, while giving rise to a vibrant community of like-minded researchers.

Center development

By taking the unusual step of inviting the finalists to design their centers, the Foundations established the culture of partnership and mutual learning before the grants were awarded. We believe that this approach set the tone for the program and motivated the centers when they entered the implementation phase.

Our conversations with the participants and the review of progress reports revealed that the MSDSEs paid close attention to each other's staffing and programmatic efforts, and in some cases tried to replicate them. While not all of these attempts were successful, they helped the centers better understand their strengths and weaknesses, and to ultimately develop the strategies that suited their particular conditions. For example, CDS tried to imitate the eScience's Incubator program, but discovered that it was more labor intensive than they anticipated and ultimately opted for more manageable Seed Grants. The MSDSEs also influenced each other's staffing models. All three centers came to appreciate the role played by executive directors in managing the centers and BIDS inspired eScience to create a CRO position. Finally, MSDSEs learned from each other when designing their physical spaces.

Joint products

Not only did the centers influence each other in creating their environments, they documented some of these efforts for the benefit of the broader data science community. For example, CDS staff wrote a paper about designing space, which was informed by the experiences of all three universities and highlighted several important elements.[20] The MSDSEs also collaborated on a paper about creating institutional change in data science, which described their experiences and could serve as a guide for other institutions.[21] A more formal joint effort was a case studies book in reproducible research spearheaded by BIDS, a "how-to" with a potential to reach a broad audience.

---

[20] Laura Norén and David Hogg. *Data Science Space & Culture.* White Paper. March 2018.

[21] *Creating Institutional Change in Data Science: The Moore-Sloan Data Science Environments*. New York University, UC Berkeley, and the University of Washington.

Community building at data summits

Many respondents were enthusiastic about the annual data science summits sponsored by the Foundations, citing their role in providing a supportive environment for early career researchers and in spurring collaborations. This was consistent with our observations of the summits, which we found to be informative and engaging. We were also struck by participants' commitment to high ethical and scientific standards, and their recognition of the role of data in society. Finally, members of the MSDSE community were open and thoughtful about the strengths and weaknesses of their environments, and the challenges they might face if they chose to follow data science as a career path in academia.

Chapter 7: MSDSE Incubator Programs

In this chapter, we present our findings from the survey of internal grant programs at CDS and eScience and summarize the experience of managing a similar program at BIDS, which was shared with us in a key informant interview.[22]

Key findings:

- Many collaborations established through the programs at UW and NYU persisted for at least a year after the experience
- Participants gained a range of benefits from the experience
- Adequate staffing is necessary to manage these programs.

In March 2018, we conducted a survey of participants in the Incubator and DSSG programs offered by eScience and in the Seed Grant program offered by CDS (Exhibit 17). The objective of the survey was to investigate the benefits of participation and the persistence of collaborations.

EXHIBIT 17: INTERNAL GRANT PROGRAMS

Program	Goal	Years offered	Number of teams
Seed Grant (NYU)	To bring together data scientists and domain scientists to foster collaborations and generate new ideas.	2016 and 2017	14
Summer Research Projects (NYU)	To help integrate master's degree students into research and to help applied researchers at NYU complete their projects. This program replaced seed grants.	Summer of 2018	8
DSSG (UW)	The program brings together data and domain scientists to work on focused, collaborative projects that are designed to impact public policy for societal benefit.	Summers of 2015, 2016, 2017, and 2018	15
Incubator (UW)	To enable new science by bringing together data scientists and domain scientists to work on focused, intensive, collaborative projects.	Fall and spring of 2014; winters of 2016, 2017, and 2018	29
Machine Shop (UCB)	To build computational research solutions, while training students in the discipline of software engineering.	2016, 2017, and 2018	15

Source:
http://escience.washington.edu/get-involved/incubator-programs/data-science-for-social-good/
http://escience.washington.edu/get-involved/incubator-programs/
https://cds.nyu.edu/nyu-data-science-seed-grant/
https://cds.nyu.edu/research/initiatives/

[22] We invited BIDS to participate in the survey of Machine Shop participants, but were told that we would not be able to collect the data we were seeking.

Seed Grants at NYU MSDSE

The goal of the Seed Grant program at NYU was to foster collaboration between data scientists and domain scientists. The application process included two steps: participation in an "open dating" session to present the ideas, followed by a formal one–two page proposal within two weeks of this event. All proposals were evaluated by the Methods Working Group for impact, innovation, and scientific merit. The winning teams received up to $25,000 to cover a graduate student or postdoc to work on the project for one semester.

All respondents to the survey (N=8) said that they continued collaborations formed under the Seed Grant program and six of the eight were still working on the same project. Collaborations were most commonly described as "working together on projects" and "discussing ideas" (six respondents for each category). We also explored how the participants benefited from the experience. According to the survey, 3-7 of the eight respondents, depending on the item, learned about new methods/tools/software/datasets, scientific areas, or new ideas; 5-6 formed new collaborations with faculty or non-faculty researchers; four developed or improved tools; and three gained software skills (Exhibit 18). Notably, three of the eight made an important discovery and two of the eight had changed the direction of their work because of the experience.

EXHIBIT 18: BENEFITS OF PARTICIPATION

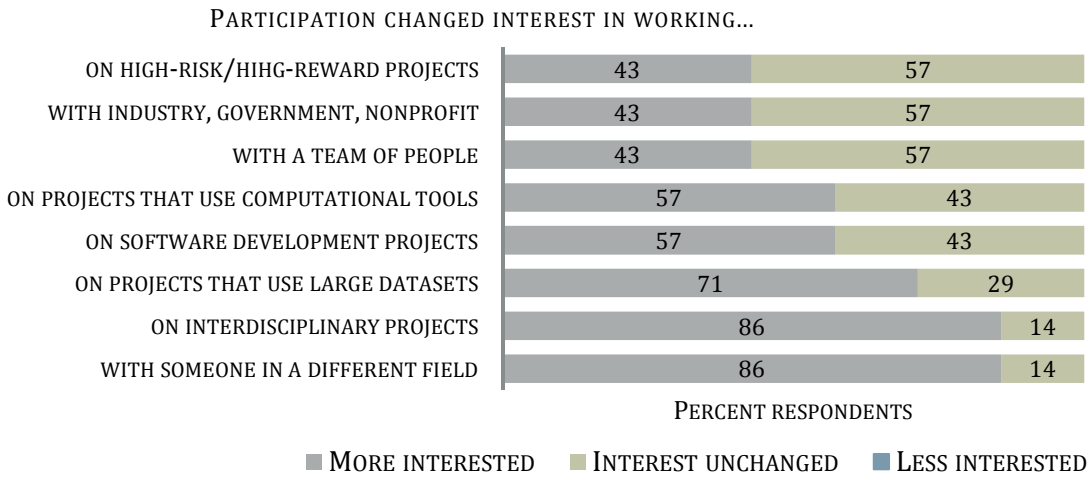
	Number of respondents (N=8)
Learned about new methods, tools, software, datasets	7
Learned about new scientific areas	6
Learned about new ideas	3
Learned about industry needs and resources	2
Formed new collaboration with faculty	5
Formed new collaboration with student, postdoc, data scientist	6
Formed new collaboration with industry, government agency, nonprofit	0
Strengthened existing collaborations with faculty, student, postdocs	2
Strengthened existing collaborations with industry, government agency	0
Gained or improved skills in software and other tool development	3
Developed or improved software, tools, methods, datasets	4
Changed direction of work	2
Changed direction of career	0
Made important scientific contribution or discovery	3
Use acquired knowledge and/or connections to obtain position	0
Published papers	2
Obtained funding to continue project	1
No benefits	0

Note: respondents could select all options that applied.

The survey also explored whether the experience has changed participants' interests. Exhibit 19 shows that 43% to 86% of respondents had become more interested in working on interdisciplinary/high-risk projects with a team of people, with

industry/government/nonprofits, and on computational projects. For the remaining, the interest was unchanged.

EXHIBIT 19: CHANGES IN INTEREST RESULTING FROM PROGRAM PARTICIPATION



When asked to suggest improvements to the program, respondents recommended more advertising, the possibility of a renewal, and more clarity about overhead spending.

In 2018, CDS replaced its Seed Grant program with a Research Internship program. The new program supports master's students to work with the NYU faculty for 12 weeks in the summer. Several reasons were offered for terminating Seed Grants. First, the concept of pairing methodological and domain PI was not viable, as most projects did not need a dedicated methodologist, and in any case the pool of CDS data scientists, who could serve as methodological PIs, was too small. In addition, it emerged that the computational needs of most NYU faculty were fairly basic, and could be met by the students enrolled in the CDS master's program, while simultaneously contributing to their professional development. As the number of master's students is relatively large, the program could be scaled up to meet the demand at the university. In its first iteration, the program received approximately 35 proposals, of which eight were funded.

DSSG and Incubator at eScience

The Incubator program, established in 2014 and offered every winter, supports collaborations between a domain science project lead (faculty, staff, postdoc, or graduate student) and an eScience data scientist. If selected via an internal call for proposals, the project lead must spend at least 16 hours per week in the WRF Data Science Studio working side-by-side with the data scientist. eScience has found that regular and consistent in-person interactions are key to the success of the projects. In 2015, eScience launched a related summer program called Data Science for Social Good (DSSG). In addition to project leads and data scientists, the DSSG teams include 4-5 students who work full-time on the projects over 10 weeks in the summer. The projects focus on societal challenges. Proposals are accepted from academic researchers, public agencies, nonprofit entities, and industry and evaluated based on the capacity for measurable outcomes, the methodological challenge, and applied social good dimension. The intent of both the Incubator and DSSG programs is that the sponsor of the project returns to their "home base"

with at minimum a solved problem and new expertise, and ideally having established longer-term collaborations.

All project leads and students who participated in the DSSG and the Incubator programs in 2014–2017 were invited to complete the survey. Exhibit 20 shows the distribution of respondents by title.

EXHIBIT 20: DISTRIBUTION OF SURVEY RESPONDENTS BY TITLE

	DSSG (N=36)	Incubator (N=20)
Faculty	3 (8%)	2 (10%)
Research staff	0	2 (10%)
Postdoc/fellow	1 (3%)	4 (20%)
Graduate student	24 (67%)	11 (55%)
Undergraduate student	4 (11%)	0
Other	4 (11%)	1 (5%)

Note: the titles of respondents who selected "other" are not reported because they can reveal their identity.

Half of the participants in the Incubator program and 31% in the DSSG program reported that they continued the collaborations formed during the project. When asked to characterize the collaborations, most respondents said that they were "working together on projects" (73–100%) and/or "discussing ideas" (55–80%, Exhibit 21). Co-mentoring, co-authoring grants, and becoming a mentor received two responses or fewer.

EXHIBIT 21: NATURE OF COLLABORATIONS AFTER PROGRAM PARTICIPATION

	DSSG (N=11)	Incubator (N=10)
Work together on projects	8 (73%)	10 (100%)
Discuss ideas	6 (55%)	8 (80%)
Co-mentor	2 (18%)	2 (20%)
Became a student/postdoc of the faculty on project team	1 (9%)	0
Wrote a grant together	1 (9%)	1 (10%)
Became a mentor to a member of project team	1 (9%)	1 (10%)
Co-teach	0	0

Note: respondents could select all options that applied.

Respondents who indicated that their collaborations had ended were asked why this occurred. Approximately half said that there was no obvious follow-up and/or that they were too busy to continue with the project (Exhibit 22). Lack of opportunity or interest were less frequently cited (10% and 26% for opportunity; 0% and 13% for interest, Incubator and DSSG, respectively). Most of the comments provided for the "other" option included some variation on "too busy" or "no opportunity," although one individual for each program indicated challenges working with the data science mentor or project lead.

EXHIBIT 22: REASONS FOR TERMINATING COLLABORATION

	DSSG (N=23)	Incubator (N=10)
No obvious follow-up	13 (57%)	5 (50%)
Too busy	11 (48%)	4 (40%)
No opportunity	6 (26%)	1 (10%)
Other	5 (22%)	2 (20%)
No interest	3 (13%)	0
Did not gain from/enjoy the experience	0	0

Note: respondents could select all options that applied.

The majority of respondents gained both methodological knowledge and domain knowledge: 94% and 100% for DSSG and Incubator, respectively, learned about new methods, tools, software, datasets, 67% and 79% about new scientific areas, and 53% and 47% about new ideas (Exhibit 23). Depending on the program and response category, 10–60% formed or strengthened collaborations with other researchers or organizations. Some respondents reported important professional outcomes, such as finding a mentor (11% for DSSG and 32% for Incubator), using knowledge/connections to obtain a position (19% for DSSG and 16% for Incubator), and changing the direction of either their work (6% for DSSG and 16% for Incubator) and/or career (11% for DSSG and 5% for Incubator). Finally, participants published papers (19% for DSSG and 42% for Incubator) and made important scientific contributions/discoveries (8% for DSSG and 5% for Incubator).

Several differences in the benefits between the two programs exceeded 15%, although only one of was statistically significant in chi-square tests. DSSG participants were more likely to learn about industry needs and resources ($p<0.05$) or to form new collaborations with other organizations (not significant). Incubator participants were more likely to strengthen existing collaborations, find a mentor, gain skills in tool development, and publish papers (all not significant).

Survey subjects who indicated changes to the direction of their work or career were asked to elaborate on their answer. The following responses were provided:[23]

Direction of work

Changed from experimental (lab-based) to computational (purely data analysis-based)
More reproducible, modular code in projects
From domain to data science
New research direction in the area of [redacted]

Direction of career/position

Increased interest in pursuing a data science role
Looking for a career in data science (industry) instead of academia
Switch into software development and data science
Working in technology industry compared to graduate school
I used development of the project in my proposal for current position
I developed skills that strengthened my job applications.

[23] These lists are not comprehensive.

Thesis won a national award and propelled me into positions I suspect would have been otherwise unattainable

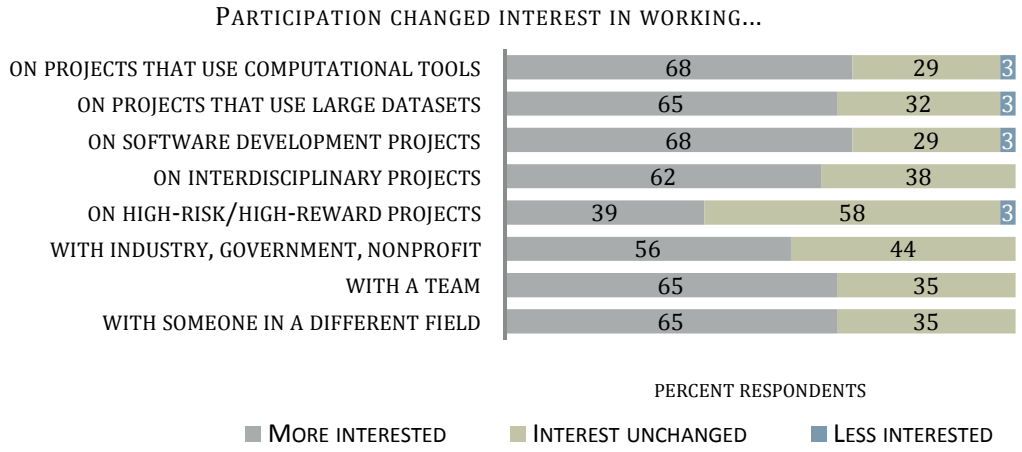
EXHIBIT 23: BENEFITS OF PARTICIPATION

	DSSG (N=36)	Incubator (N=19)
Learned about new methods, tools, software, datasets	94%	100%
Learned about new scientific areas	67%	79%
Learned about new ideas	53%	47%
Learned about industry needs and resources	61%	21%
Formed collaboration with faculty	39%	26%
Formed collaboration with student, postdoc, data scientist	39%	53%
Formed collaboration with industry, government agency, nonprofit	33%	16%
Strengthened collaborations with faculty, student, postdoc	31%	58%
Strengthened collaborations with industry, government agency	14%	11%
Found a mentor	11%	32%
Gained or improved skills in software and other tool development	81%	100%
Developed or improved software, tools, methods, datasets	72%	84%
Changed direction of work	6%	16%
Changed direction of career	11%	5%
Made important scientific contribution or discovery	8%	5%
Use acquired knowledge and/or connections to obtain position	19%	16%
Published papers	19%	42%
Obtained funding to continue project	14%	21%
No benefits	0%	0%

Note: respondents could select all options that applied.

The survey also explored whether the experience changed participants' interests. Exhibit 24 shows that 56–68% of respondents became more interested in working on computational, high-risk, and data-intensive projects; and on interdisciplinary teams. For virtually all the rest, the interest was unchanged. The results were similar between the DSSG and Incubator programs.

EXHIBIT 26: CHANGES IN INTEREST RESULTING FROM PARTICIPATION IN THE DSSG PROGRAM



Finally, 20% of the Incubator and 61% of the DSSG participants said that they would make a change to the programs. Suggestions for improvements included more mentorship and structure, a faculty stipend, time to socialize, a more intense initial training, a longer program duration, the possibility of follow-up funding, and several others.

Machine Shop at BIDS

BIDS staff opted out of the survey of its Machine Shop program, but agreed to an interview about their experiences.[24] We heard that the goal of the program was to connect research groups with computational needs with software development experts who can satisfy these needs. The ideal projects should be narrow in scope (take 6 to 12 months to complete) and result in a software product and a group of people who can continue its development. The immediate challenge of this model was to identify software experts willing to serve as mentors. BIDS was planning to engage as many of its own postdocs as possible, but only one or two were able to participate and several additional mentors had to be recruited outside of BIDS. In contrast, the interest among undergraduates was very high, but did not match their experiences in software development. So while BIDS managed to put together the first cohort of teams through a considerable effort, it was clear that the model was not sustainable. Ultimately, BIDS reduced the number of projects to three, which is manageable for one research scientist to oversee, and imposed more stringent selection on the students who were accepted.

BIDS staff also discovered that in addition to having the right team, it was important to find a project that is well-defined, interesting, built on previous work, and feasible to advance quickly. As an example, our respondent described an ongoing project to identify bee species based on veins in their wings. The first phase of the project had been described in a master's thesis, and the students in the program could train themselves by replicating this work. In addition, several mentors were interested and could share the supervision of students, and one of the mentors was actively collecting data to inform the algorithm. This project had all the components described above to be successful.

Our respondent commented that this type of program required several staff to run smoothly. He unfavorably contrasted BIDS with its two data scientists to eScience, where several staff are formally responsible for the DSSG and Incubator programs.

[24] Machine Shop: https://bids.berkeley.edu/research/bids-machine-shop.

Chapter 8: Contribution to the Ecosystem of Tools and Practices

One of the key goals of the MSDSE program is to contribute to the ecosystem of research-related tools and practices to enable data-driven discovery. We used publicly available information to trace more than 200 software products created by the centers in an attempt to capture their reach and impact. In addition, we interviewed the developers for nine of these tools to understand why they decided to get involved in this work and whether these efforts benefited their careers.

Key findings:

- The tools developed by MSDSE participants reflected the commitment to open science and reproducibility
- Researchers at different career stages benefited from tool development
- We were unable to determine the impact of the software tools unless they were published in peer-reviewed articles

Based on an expert interview and our own research, we concluded that the measures of influence and reach for software products are still under development and currently include the following:

- GitHub releases – the number of times a software package has been finalized for distribution (release) to end users on GitHub, an open-source tool repository[25]
- GitHub contributors – the number of people outside the core development team who have submitted proposed changes to the source code[26]
- SourceRank score in Libraries.io – a score assigned to open-source software packages based on attributes that tend to reflect a dependable package[27]
- GoogleScholar citations – the number of times a publication about a given software package has been cited by other publications[28]
- Altmetric attention score – a numerical measure comprised of a weighted count of indicators of online attention such as mentions in news sources, blogs, Facebook, Twitter, Wikipedia posts, policy documents, and patent citations.[29]

We collected these measures for 234 products reported by the MSDSEs and found that their values spanned a very large range. Consequently, we show the fraction of the tools for which

[25] https://help.github.com/articles/about-releases/.

[26] https://github.com/CoolProp/CoolProp/wiki/Contributors-vs-Collaborators.

[27] https://docs.libraries.io/overview.html.

[28] https://scholar.google.com/intl/en/scholar/citations.html.

[29] https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated.

the indicators were available, rather than the average values. Of the 234 tools, 92% had GitHub releases on average, reflecting the commitment of the MSDSE participants to open-source practices (Exhibit 27). Most of the tools were managed by a small number of contributors and had relatively few releases, although there were some notable exceptions (e.g., Jupyterlab issued 2,391 releases and Scikit-learn included 1,085 contributors). Only one-third of the tools, on average, had associated publications and only one-quarter had Altmetric scores.

EXHIBIT 27: STATISTICS ON THE DEVELOPMENT AND IMPACT OF MSDSE SOFTWARE

	Percent with indicator			Values across the MSDSEs
	BIDS (N=82)	CDS (N=86)	eScience (N=66)	Max value (tool)
GitHub releases	98	90	91	2,391 (Jupyterlab)
GitHub contributors	99	90	91	1,085 (Scikit-learn)
Libraries.io SourceRank	74	34	35	23 (IPython)
Google Scholar citations	17	44	42	1,460 (Visualization tutorial)
Altmetric score	12	31	33	1,489 (Code for visual-turing-tests)

Of the all the tools reported by the MSDSEs, we selected nine examples for more in-depth investigation, which included interviewing the lead developer and performing additional online searches. The results of these efforts are summarized below.

Brief description of the tools

Astropy is a Python library of common computational functions for astronomers. The development of this resource began before the MSDSE grant and its contributors number in the hundreds. One of these included a data scientist from eScience, who told us that the AstroPy project was a model for collaboration and tool sharing. The astropy core package was designed by a consortium of 44 astronomers in 2011. The project aims to support and encourage the development of open-source and openly developed packages in the Python computing language through a library that standardizes core functionality for astronomical software. As of 2018, the Astropy collaboration included 240 contributors and 20 package leads and maintainers, and has an ecosystem of 34 "affiliated" Python packages for functions related to astrophysics.

The Astropy project utilizes best practices in open science and open-source software by using a transparent, deliberative process to consider modifications and additions to the package and by maintaining the focus on the package being "developed for and (at least in part) by the astronomy user community." In acknowledgement of the fact that most astronomers are not trained in computer science or software engineering, the project has made use of data scientists and software developers to design code in keeping with good software practices. The core package includes 12 modules addressing the most commonly needed functions in astronomy and astrophysics, named in plain English to facilitate use. The software included in the library had 1.5 million downloads and over 1,400 citations, and had been used for over 900 projects.

Carl is a Python toolbox designed to be used in High Energy Physics (HEP) to carry out likelihood-free inferences in complex processes. It is designed for use in experimental particle physics, where observations are linked to complex simulations, and researchers would like to estimate values and their confidence intervals. The core developer of Carl is an experimental particle physicist affiliated with CDS. Carl was the by-product of the magnet explosion at the Large Hadron Collider, which led the developer to "look for" the particle in existing data instead

of collecting the data from the collider. The tool has been used in physics and in other fields such as genomics. Altmetric places this software in the top 5% of all research outputs.

Librosa is a Python library for audio and music analysis software. The lead developer was a CDS faculty member at NYU, who collaborated with two senior data scientists, one at CDS and another at eScience, to convert MATLAB scripts into Python. Music information retrieval (MIR) is an emerging field that is rapidly developing due to the rise of digital music services such as iTunes, Pandora, and Spotify. MIR spans topics at the intersection of musicology, digital signal processing, machine learning, information retrieval, and library science. Librosa is an open-source Python package that provides implementations of common functions used in MIR. The goal of the package was to ease the transition of MIR researchers, many of whom may be more familiar with MIR libraries in MATLAB or C++, into Python and to extend the reach of MIR techniques to a larger community of scientists. The tool has been used in at least 54 academic research projects.

Permute, available in both Python and R, is a package for implementing randomizations for a variety of experimental designs. The developer was a PhD student at BIDS, with the support of her BIDS-affiliate advisor and another student. Randomized experiments have come to be accepted at the "gold standard" in effectiveness testing in medical and now social sciences applications. However, this approach is often also used with data that do not meet the appropriate assumptions. The developers' goal was to enable researchers to make decisions about what kinds of nonparametric tests might be suitable for their experimental design and their data. The tool itself is applicable to a variety of experimental designs and for a variety of estimation problems within those designs. At the time of the research, Permute has been used primarily within the development team, with full release planned for summer 2018.

Pulse2percept is a software in Python designed to simulate the patient experience with retinal implant devices. The software was conceived by two eScience faculty, who collaborated with a postdoctoral fellow (the lead developer) and a data scientist to implement it. Hereditary retinal diseases affect millions of people. Several types of retinal prostheses are currently under development but, at present, none of these devices come close to restoring natural vision. Another gap in vision prosthetics is a lack of computational tools to capture a patient's experience without implanting a device. A team from eScience decided to address this problem by creating a simulation software package called pulse2percept. The tool has been used by researchers and artificial retina device manufacturers. A conference presentation of the software in 2017 received over 400 views on YouTube.

ReproZip is a Python tool for combining files into a single, portable package. It was developed by an all-CDS team, with a PhD student as the lead, supported by senior data scientists and other staff. The software package was designed to enable a researcher to pack all of the necessary data files, libraries, environment variables, and options – all of the dependencies needed to reproduce the experiment – on a second researcher's own machine or system. This tool has been used in both natural and social sciences applications and had 1,600 downloads.

Sncosmo is a Python library for supernova cosmology. Developed by a senior data scientist at BIDS, it became linked to AstroPy. Supernova cosmologists must conduct a series of computations in order to establish distances between bodies in the universe in order to understand their evolution and extinction. The usual work of these scientists is to observe the

phenomenon (a supernova typically appears when it explodes, reaches peak brightness and then disappears within a matter of a few weeks, depending on the type) and then develop computational models to simulate the SN and compare it with the empirical data collected. Celestial events such as these can only be observed in very particular conditions (i.e., night time, away from urban light pollution) creating interrupted sequences of data. Sncosmo is a Python library for simulating, fitting and typing supernova light curves, built on NumPy, SciPy and Astropy. It has been downloaded 1,200 times and used in at least 10 projects.

TopoAngler is an interactive visualization tool to help biologists process image data. It was developed by a postdoctoral fellow and a senior data scientist at CDS in collaboration with a marine biologist at UW. One marine biologist at the UW received funding to create a cost-free database of scans of skeletons of all of the world's 30,000 species of fish. To do this, his team uses a Micro-CT device to scan the fish, but because each scan at a useful resolution requires 12 hours to complete, the team bundles up to a dozen fish to be scanned simultaneously. The resulting scan, however, must be "segmented" (i.e., the parts corresponding to the skeleton of each fish must be identified to isolate individual skeletons). The developers of TopoAngler used the visualization programming language Inviwo to create an interactive program that enables efficient processing of the images. The tool may ultimately be extendable to many biological research applications.

Viscm is a Python tool for visualizing and designing color maps to display data. Its BIDS-affiliated co-developers undertook the project as a way to begin a collaboration. The tool has been used to replace a much-maligned default color map in matplotlib and to optimize other maps. A critique of the default Matplotlib colormaps, particularly its previous default, jet, was that they distorted the data and were very difficult to read for individuals with color-vision deficiency. Other existing colormap packages (e.g., parula, in MATLAB) were not open source. Viscm is a package for analyzing and designing colormaps for data visualization and representation. The colormap created using this software is viridis, which has now replaced jet as the default in Matplotlib. Viridis does not suggest data features that are not there, uses perceptual ordering to facilitate interpretation, retains critical information when rendered in grayscale, and can be interpreted by color-impaired viewers. Viscm has been downloaded nearly 900 times and a presentation about the software viewed 86,000 times on YouTube.

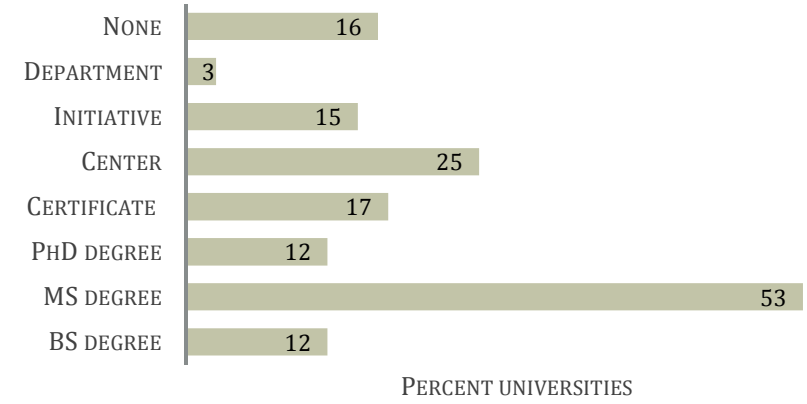
Discussions with the developers

We interviewed the developers of these tools about their experiences and the benefits of MSDSE to their careers. All respondents told us that the MSDSE funding offered opportunities for innovation and the freedom to explore new ideas. We also found that the developers strongly believed in open science and reproducibility. Some intentionally set out to explore novel approaches to share computational tools and acknowledge the developers' efforts. It was also clear from the interviews that software development projects benefited researchers at all career levels. Faculty members were able to advance their research program by learning from postdocs and data scientists. Perhaps more importantly, graduate students gained an opportunity to design and carry out their own projects.

Chapter 9: Academic Data Science Landscape

In the past 10 years, many universities in the United States launched academic degree programs, research centers, and even new departments in data science, data analytics, or related fields. A review of websites for 116 research universities in 2017 revealed that less than 20% did not list these types of entities. The remaining listed various centers and initiatives, as well as academic and professional programs (Exhibit 28).

EXHIBIT 28: DATA SCIENCE OFFERINGS AT RESEARCH UNIVERSITIES



Source: Moore Foundation, 2017. N=116.

We examined the mission, organization, and activities of data science entities at 20 universities. This information was collected to inform the evaluation of the Data Science Environments program at the Foundations,[30] which helped launch three such centers, but we hope that our findings will be helpful to the broader communities of data scientists, academic administrators, and funders.

The initial universities included in the study were suggested by the Foundations. The sample was expanded through recommendations from these early participants to 20 institutions (Exhibit 29). Because of this sampling strategy and some level of non-response, this report is neither representative nor inclusive of all university efforts in data science. Rather, we present a range of models being explored and some common challenges to establishing data science centers.

With the exception of the three MSDSEs for which extensive data were collected in the evaluation, the information presented in this report is based on three sources: (1) phone interviews with the leadership of the entities conducted between December 2016 and June 2018, (2) a review of the centers' websites and materials shared with us; and (3) a short survey of participants in the Data Science Leadership Summit held in October 2018.

In the next section, we summarize the organization, staffing, and programs at the 20 centers, which we attempted to illustrate with diverse examples, followed by more detailed profiles of

[30] http://msdse.org/.

each site.[31] We caution the reader that data science centers are rapidly evolving, and that consequently some of the information included in this report may be out-of-date.

EXHIBIT 29: CENTERS INCLUDED IN THE REPORT, ORDERED ALPHABETICALLY BY THE UNIVERSITY

	Institution	Name
1	Boston University (BU)	Data Science Initiative (DSI)
2	California Institute of Technology and Jet Propulsion Lab (Caltech and JPL)	Center for Data Driven Discovery (CD^3) and Center for Data Science and Technology (CDST)
3	Columbia University (Columbia)	Data Science Institute (DSI)
4	Duke University (Duke)	Information Initiative at Duke (iiD)
5	Harvard University (Harvard)	Harvard Data Science Initiative (HDSI)
6	Johns Hopkins University (JHU)	Institute for Data Intensive Engineering and Science (IDIES)
7	Massachusetts Institute of Technology (MIT)	Institute for Data, Systems, and Society (IDSS)
8	Michigan State University (MSU)	Department of Computational Mathematics, Science and Engineering (CMSE)
9	New York University (NYU)*	Center for Data Science (CDS)
10	Northwestern University (NW)	Data Science Initiative (DSI)
11	Ohio State University (OSU)	Translational Data Analytics Institute (TDAI)
12	Stanford University (Stanford)	Stanford Data Science Initiative (SDSI)
13	University of California Berkeley (UC Berkeley)*	Berkeley Institute for Data Science (BIDS)
14	University of Chicago (UChicago)	Computation Institute (CI) and Center for Data and Applied Computing (CDAC)
15	University of Massachusetts Amherst (UMass)	Center for Data Science (CDS)
16	University of Michigan Ann Arbor (UMichigan)	Michigan Institute for Data Science (MIDAS)
17	University of North Carolina Charlotte (UNCC)	Data Science Initiative (DSI)
18	University of Rochester (URochester)	Goergen Institute for Data Science (GIDS)
19	University of Virginia (UVA)	Data Science Institute (DSI)
20	University of Washington (UW)*	eScience Institute (eScience)

**Moore-Sloan Data Science Environments*

Mission, leadership, and organization

We found that 17 of the 20 centers were formed within the past five years and the remaining three (at BU, UChicago, and UW) stemmed from or extended pre-existing units. The creation of the centers was motivated by the growing interest in data science among the faculty and the perceived need to connect and/or build up existing programs. Several respondents recalled an elaborate planning phase, which lasted for several years and involved dozens of faculty and administrators; it was our impression that this highly participatory process was fairly typical. Interestingly, when asked to share any noteworthy observations about establishing a center, the most commonly mentioned "lesson learned" was the importance of listening to the university community about their needs and concerns, and frequently reporting back progress as the centers were being planned. Relatedly, several respondents said that understanding the

[31] All interview respondents were asked and most agreed to review and correct their profiles.

university's political landscape and persuading departments that they would benefit from the data science entity were their greatest challenges.

We heard opposing views about who should spearhead the creation of a data science center. Some respondents believed that it should be promoted by the university leadership, because the faculty have few incentives to "step out" of their area of expertise. Others argued that these types of initiatives should originate with the faculty in order to be accepted. One interviewee noted that when planning the center, it was important to not empower researchers in any single topical area to avoid disciplinary bias.

The mission statements of all centers articulated a commitment to collaborative and interdisciplinary research. Some of the centers also articulated the goal of education and/or workforce development and of contribution to society. Exhibit 30 is a word cloud generated using the mission statements, which highlighted the terms "education, research, science, interdisciplinary, methods, and data."

EXHIBIT 30: WORD CLOUD OF MISSION STATEMENTS

Note: the mission statements were edited to remove common terms such as "university" as well as titles that often contain the words "data science." N=20.

All of the centers were led by a faculty director (two co-directors at Harvard and UMichigan), and 9 of the 20 also included a non-faculty executive director (Exhibit 31). The directors were typically inaugural and had played key roles in designing and launching the centers. Most centers were overseen by Executive Committees composed of faculty.

In a testament to the interdisciplinary nature of data science, virtually all entities were administratively based outside of single departments and bridged multiple schools or colleges. The two exceptions were Umass and MSU. At Umass, CDS is located within one College of Information and Computer Science. This choice was made by the leadership of the center to retain control over the training of students by the college and simplify/speed-up the decision-making process. MSU launched a new department, the CMSE, administered jointly by the College of Natural Sciences and the College of Engineering. We were told that the planning committee concluded that data science centers at peer institutions were short-lived, and that a new department would be a more permanent solution. However, our respondent acknowledged that significant start-up and maintenance costs made widespread support difficult to secure. In

addition, compared to a center, a department could be perceived as more insulated, potentially discouraging faculty engagement.

EXHIBIT 31: CENTER ORGANIZATION AND PARTICIPANTS

	Year launched	Space	Single departm. or college	Non-faculty managing director	Faculty lines	Data scientists	Postdocs
BU DSI	2012/14	✓		✓	✓	✓	✓
Caltech CD^3 and JPL CDST	2015	✓				✓	
Columbia DSI	2012	✓			✓	✓	✓
Duke iiD	2013	✓				✓	✓
Harvard DSI	2017	✓					✓
JHU IDIES	2012	✓			✓	✓	
MIT IDSS	2015	✓		✓	✓	✓	✓
MSU CMSE	2015	✓	✓	✓	✓		
NYU CDS*	2013	✓			✓	✓	✓
NW DSI	2015			✓	✓		✓
OSU TDAI	2015	✓		✓	✓		
Stanford DSI	2014			✓			
UC Berkeley BIDS*	2013	✓		✓		✓	✓
UChicago CI and CDAC	2000/18	✓					
UMass CDS	2015		✓		✓	✓	✓
UMichigan MIDAS	2015	✓		✓			✓
UNCC DSI	2012	✓			✓	✓	
URochester GIDS	2013	✓			✓	✓	
UVA DSI	2013	✓			✓		
UW eScience*	2008	✓		✓	✓	✓	✓

**Moore-Sloan Data Science Environments.*

Space and funding

All but three centers had dedicated space (Exhibit 31), which ranged from a few meeting rooms and/or faculty offices to large portions of a building. The spaces were described as open, multi-purpose, configurable, vibrant, and collaborative. At UVA, faculty, staff, and students reside together in an open space, following the original "academical village" concept of Thomas Jefferson who founded the university. In several cases, the spaces were newly built or renovated with contributions from private donors or local governments. Several center directors were looking to move to accommodate their growing communities. The centers at Stanford, UMass, and JPL were "virtual," which was not viewed as an impediment to their function.

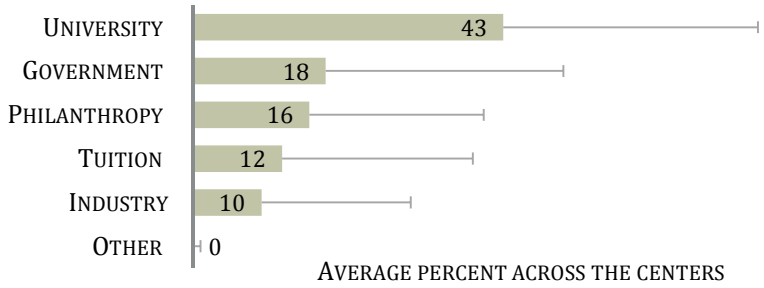
To collect more systematic data on the funding sources for the centers, we supplemented the interview data with a survey of data science leaders who attended a recent summit. Generally speaking, the centers pieced together funding from multiple sources to support their operations and programs (Exhibit 32). The university provided the bulk of the funding, at 43% of the total,

followed by the government and by philanthropy, at 18% and 16%, respectively. We note large standard deviations for each funding source, indicating a high degree of diversity.

Based on the interviews, the university investments in data science initiatives were in some cases quite large: $125 million at OSU, which included the hiring of approximately 70 faculty; $20.5 million at UMichigan; $50 million at URochester; and $25 million at JHU. In most cases, the university provided funding for the initial period (on the order of five years), after which the centers were expected to become self-sufficient. However, UMass and UW committed to support their data science centers for the foreseeable future. The leadership of several centers estimated that the annual budget to fund space, staff, equipment, programs, and events was in the range of $1–2 million.

EXHIBIT 32. FUNDING SOURCES TO SUPPORT THE CENTERS

Q: Please indicate the approximate percent of funding from the following sources that support the core activities of your program (e.g., staff salaries, space, community events, collaborative research projects). Do not include individual investigator research grants, matching time from faculty, or leveraged resources.



Note: Survey and interview respondents are partially overlapping groups. N=28.

Participants

Faculty

Thirteen of the 20 universities allocated faculty lines to the centers (Exhibit 31). As the centers cannot grant tenure, these faculty were appointed jointly with departments, and their duties were divided. Many centers also included numerous affiliated faculty without appointments.

Some universities developed new hiring and promotion policies for joint faculty. For example, at MIT's IDSS, both the department and the center participate in tenure review, but the department has a stronger vote because it can fully "absorb" faculty who wish to leave the center, while IDSS cannot. Similarly, CDS at NYU developed and successfully implemented a systematic process and criteria for joint faculty hiring. While administratively a department, CMSE aimed to replicate the research model used by national laboratories and recruited faculty with the skills to develop tools for solving difficult scientific challenges. All 27 faculty in the department have joint appointments (with 12 different departments), which are either a 70:30 or 30:70 split, to ensure that each faculty member has a primary "home."

UMass' CDS decided against joint appointments as it could be detrimental to faculty, especially assistant professors. To promote collaboration, the center uses cluster hiring. Finally, some centers established non-tenure tracked faculty positions, which do not require department involvement. For example, DSI at UVA hires "general" faculty, who teach students and conduct research, but are not eligible for tenure (the center also has joint tenure-track faculty).

Data scientists

Of the 20 centers, 12 created data scientist/engineer positions (Exhibit 31). In some cases, these staff spent all or most of their time providing consulting services (e.g., at UNCC), but more typically they were actively involved in collaborative research projects (e.g., at MIT, JHU, JPL, and UW). Several respondents mentioned that data scientist positions are difficult to create, even though many laboratories struggle to obtain the computational support they need. A respondent from BU told us that he was able to persuade the university to pilot several lines for software engineers, with the understanding that the investment was temporary as these staff would quickly transition to grant support. The pilot was an immediate success: within a year, nine software engineers were hired and the number continues to grow.

Through our team's evaluation of the MSDSE program, we became particularly familiar with the data scientist track created by eScience at UW. These staff are the engine of the center, leading most activities and programs, while simultaneously conducting their own research. eScience put in place several mechanisms to support these researchers. One is a "salary buyback program," whereby a portion of the grant funding obtained by data scientists is returned to them as a stipend. In addition, data scientists can obtain a PI status, which allows them to apply for independent funding. Finally, their salaries have been adjusted to be more competitive with industry.

Postdocs

Roughly half of the centers launched postdoctoral fellowship programs (Exhibit 31), which were at least in some cases very competitive. At Harvard, postdocs were based in the department of their primary mentor, but had access to the DSI's shared space and participated in monthly lunches with the center co-directors. DSI at NW established the Data Science Scholars program to diversify the domain-focused research portfolio of recent PhD graduates and to establish their reputation as leaders in data science. The participants held joint appointments with the Northwestern Institute on Complex Systems and at least one other research center on campus that matched their area of expertise. At MIT, IDSS postdocs "belonged" to the center rather than an individual faculty, and were expected to support its mission by working collaboratively with faculty across multiple schools. The fellows program at NYU's CDS was viewed as one of its key accomplishments. Recruited through a competitive national search and fully funded for one–three years, these researchers functioned at the level of assistant professors. Not encumbered by academic duties, the fellows flourished in the collaborative environment of the center and were highly successful on the academic job market.

Research activities

Data science centers included in the study established three types of research programs: small seed grants to provide short-term support for a student or postdoc, larger team projects, and student research experience (Exhibit 33).

Small seed grants and larger team projects

The majority of the centers offered internal funding opportunities, which typically aimed to bring together interdisciplinary faculty to work on projects that may lead to follow-up funding. Some programs paired domain scientists with methodologists, others required that faculty had not previously worked together or represented intellectually distinct disciplines, yet others targeted junior scholars. In most cases, these grants were in the range of $25,000–100,000 and could support a graduate student or postdoc for a short period of time (Exhibit 33). Typically,

all faculty at the university were eligible to participate, and the awards were made through a competitive but simple application process. Some calls for proposals incorporated industry co-sponsors, who reviewed the applications and followed up with the applicants directly if they were interested.

The MIDAS center at UMichigan used a different model. Under its Challenge Initiatives Program, MIDAS awarded over $10 million to support 9 projects in predetermined priority areas, which brought together multidisciplinary teams totaling 75 faculty and 79 students/postdocs. The projects were chosen for their potential scientific, educational, and societal impact through two rounds of internal review. However, MIDAS plans to switch to the small seed grant model ($75,000 per project) in the future.

EXHIBIT 33: INTERNAL FUNDING PROGRAMS

	Small seed grants	Larger team projects	Student research experience outside of degree programs
BU DSI	✓		✓
Caltech CD^3 and JPL CDST	✓		✓
Columbia DSI	✓		✓
Duke iiD			✓
Harvard DSI	✓		
JHU IDIES	✓		
MIT IDSS	✓		
MSU CMSE			
NYU CDS*	✓		✓
NW DSI	✓		
OSU TDAI	✓		✓
Stanford DSI			
UC Berkeley BIDS*	✓		✓
UChicago CI and CDAC			
UMass CDS			
UMichigan MIDAS	✓	✓	✓
UNCC DSI	✓		
URochester GIDS	✓		
UVA DSI			
UW eScience*			✓

**Moore-Sloan Data Science Environments. Note: We are uncertain about the completeness of these data.*

Student research experience

Half of the data science centers supported student research experiences not linked to degree programs. For example, *Data+* at Duke's iiD is a 10-week summer program that offers undergraduates the opportunity to explore data-intensive problems from nonprofit and corporate clients. Students form several small teams that work in a communal environment. In 2017 the program sponsored 25 projects involving 75 students, who were chosen from 300 applicants. MIDAS supports 4 data science student groups with a combined membership of

more than 400 students and 50 faculty members. These groups have completed 14 public service projects across southeast Michigan. eScience at UW runs winter and summer incubator programs, which bring together data scientists and domain scientists. The summer session (DSSG) supports projects with a potential for societal impact, while the winter session focuses on high-risk/high-reward projects with an engineering component.[32] Two of the centers had programs for high school students. OSU's TDAI runs a cost-free data science summer camp for girls attending Columbus high schools, where participants gain experience using software tools and presenting their work. A similar program is offered by MIDAS to economically disadvantaged high school students in southeast Michigan. Finally, many centers incorporated data science projects as components in courses or degree programs.

Community engagement

Almost all centers offer seminars, workshops, consulting services, and annual meetings (Exhibit 34). While we did not plan to explore these community-building activities in interviews due to limited time, some respondents described them as highlights of their centers. For example, BU hosts a Data Science Day – a symposium to connect data scientists and methodologists. Each year, the event is organized around several themes, which have most recently included such diverse topics as artificial intelligence, cybersecurity and law, and epigenetics. A Data Science Day at Columbia is a showcase for data science research and educational activities on campus that attracts hundreds of government, corporate, nonprofit, and academic leaders. UMass runs an annual career event, where students present posters to industry partners and visit company tables for further discussion, and which lead to numerous internship opportunities and offers of employment. The MSDSEs get together for an annual data summit where they discuss pertinent issues, such as reproducibility, careers tracks, and ethics of data, in addition to giving scientific presentations.

Industry partnerships

Eight of the 20 centers (at Columbia, MIT, NYU, Stanford, UChicago, UMass, UMichigan, and URochester) launched industry partnership programs, and several others receive some industry funding on a more ad hoc basis. Some centers cited substantial industry contributions (e.g., CDS at UMass received $15 million from MassMutual and significant additional funding of an unspecified amount from IBM, Pratt & Whitney, Google, Oracle, Microsoft, Amazon, and the Chan Zuckerberg Initiative). Stanford's DSI raised approximately $4 million per year through its industry program. This center chose to support its activities almost entirely through corporate contributions, citing spending flexibility and larger budgets relative to the government sources, as well as the practical nature of the problems of interest to industry that resonates with the Stanford community. It was our impression that most, if not all, centers were trying to make industry connections for access to research dollars and "real life" problems.

[32] As part of the MSDSE evaluation, we conducted a survey of participants in these programs and found that they produced lasting collaborations.

EXHIBIT 34: COMMUNITY-BUILDING PROGRAMS OFFERED BY THE CENTERS

	Annual meeting, summit, retreat	Workshops, boot camps	Data science consulting
BU DSI	✓	✓	✓
Caltech CD³ and JPL CDST		✓	✓
Columbia DSI	✓	✓	✓
Duke iiD		✓	✓
Harvard DSI			
JHU IDIES	✓	✓	✓
MIT IDSS	✓	✓	
MSU CMSE		✓	
NYU CDS*	✓	✓	✓
NW DSI		✓	✓
OSU TDAI		✓	
Stanford DSI	✓	✓	
UC Berkeley BIDS*	✓	✓	✓
UChicago CI and CDAC			
UMass CDS	✓	✓	
UMichigan MIDAS	✓	✓	✓
UNCC DSI			
URochester GIDS	✓	✓	✓
UVA DSI	✓	✓	
UW eScience*	✓	✓	✓

**Moore-Sloan Data Science Environments.*

Academic programs

As all universities surveyed have courses and programs related to data science, we asked interviewees to focus on what was managed by their centers. We found that master's programs were especially popular, offered by half of the centers (Exhibit 35). These programs typically combined courses in quantitative methods and domain sciences. For example, master's students at UNCC can mix and match courses to earn degrees in crime analytics, anthropology analytics, or health analytics; and a similar approach is planned for the undergraduate and PhD tracks. Another interesting example is the online micro-master's program at MIT. The graduates earn a certificate and can use it to earn a master's degree by taking additional courses elsewhere. Four of the entities launched a PhD program. Some centers offer online programs that attract very large audiences. For example, Caltech and JPL run a joint summer school, which incorporates data from space missions. Initially advertised to fewer than 100 people, the program attracted 30,000 registrants in the recent iteration.

Funding allocation

In the survey of data science leaders, we asked how the total center funding is currently allocated across various activities and how they would invest additional resources if they were available. Exhibit 36 shows that the centers spend on average 20% of their budget on management, 15% on technical staff (such as data scientists), 20% on graduate students and postdocs, and 18% on faculty. The remaining 20% is roughly divided between engagement activities and internal funding. If 50% more funding were available, the centers would allocate the largest fraction, 26% on average, to technical staff, which almost doubles the current spending level. The centers would invest 21% of the additional funding in fellowships. The

remaining half would be divided between management, community engagement, internal grants, and faculty in roughly equal allotments of 10–15%.

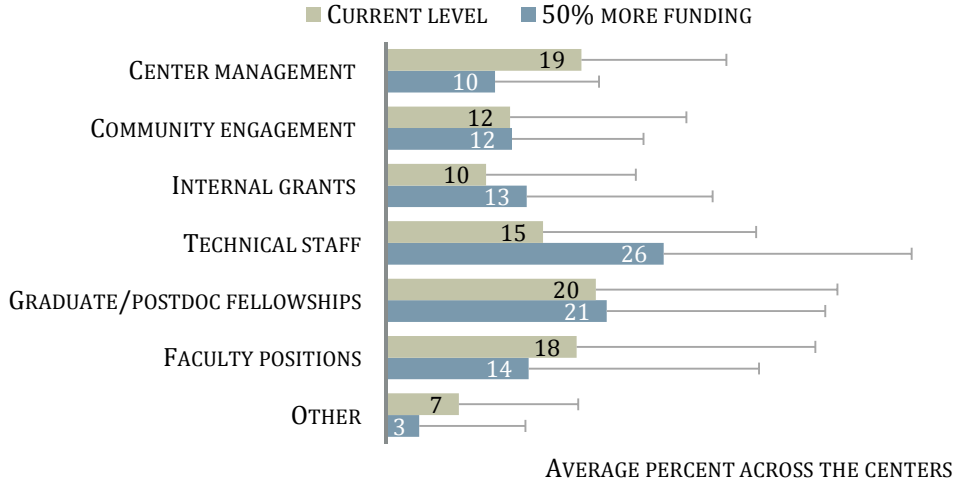
EXHIBIT 35: EDUCATIONAL PROGRAMS MANAGED BY THE CENTERS

	Certificate program	Undergraduate major/concent.	MS program	PhD program
BU DSI				
Caltech CD³ and JPL CDST				
Columbia DSI	✓		✓	
Duke iiD	✓	✓	✓	
Harvard DSI				
JHU IDIES				
MIT IDSS		✓	✓	✓
MSU CMSE	✓			✓
NYU CDS*			✓	✓
NW DSI		✓	✓	
OSU TDAI				
Stanford DSI				
UC Berkeley BIDS*				
UChicago CI and CDAC		✓	✓	
UMass CDS	✓		✓	
UMichigan MIDAS	✓			
UNCC DSI	✓		✓	
URochester GIDS		✓	✓	
UVA DSI			✓	✓
UW eScience*				

**Moore-Sloan Data Science Environments.*

EXHIBIT 36: FUNDING ALLOCATION BY THE CENTERS

Q: How is the funding for your center distributed among the various center activities? If you had 50% more unrestricted funding available, how would you distribute those additional resources?



Note: N=27 (current) and N=25 (50% more)

Conclusions

A survey of 20 data science entities in the United States revealed that they used a combination of similar building blocks to create unique entities that fit their specific context and goals. Several foundational elements were common to all or most centers – such as dedicated space, a strong emphasis on collaboration and interdisciplinarity, and a community-building mission. At the same time, clear differences between the centers also emerged. Some were primarily focused on research, while others combined research with training. Some centers were anchored by data scientists/engineers, while at others postdocs or faculty played more prominent roles. Several centers built strong connections to industry. The propensity of the centers to follow a different path was particularly notable for three MSDSEs, which received the same resources to implement a shared framework, and yet within a few years evolved along their own paths. Despite our conclusion that each data science center will chart its own course, we hope that our study offers some insights into the range of possibilities for supporting this emerging field.

Chapter 10: Summary

In this chapter, we draw on the data presented in the report to address the questions posed for the evaluation.

In what ways are the three data science environments different from and similar to one another? What are the relative strengths and weaknesses of the different approaches?

During the design phase, the grantees settled on the same framework of working groups around which the environments would be built, while exercising the flexibility to choose their own implementation strategies. Several years later, three unique data science environments emerged, shaped by their pre-MSDSE history, leadership style and vision, and institutional culture. In this section, we compare and contrast the centers.

Historical context

The MSDSEs began their journeys from different starting points. eScience predated the Moore-Sloan grant by approximately five years and at the time of award already had a mission, a leader, and a small team of data scientists. The MSDSE grant enabled eScience to increase its staff, launch new programs, and more firmly take root at the university, but it did not fundamentally change its character or direction. Data scientists remained "the center of gravity" at eScience, and this career track is viewed by the participants as one of their most important contributions to the MSDSE program and to the field more broadly. Similarly, CDS was also launched before the grant, but in this case the investment facilitated its transformation from an administrative home to the master's program to a flourishing research institute. The master's program continues to play an important role at CDS: its faculty teach courses in the program, some students participate in the research projects at the center, and the tuition revenue is being considered as a source of sustainable funding in the future. In contrast to the other two centers, BIDS was created with the MSDSE grant. While recognized as an exciting hub for data science, this center is still defining its direction and place at the university.

The environments also have several important features in common. First, all three are administratively and physically untethered from departments – a strategic choice meant to signal inclusivity and disregard for disciplinary borders, which is viewed as successful. The MSDSEs also share a collaborative working style, a focus on early-career researchers, and a strong commitment to the principles of open and reproducible science.

Staffing

Each center is governed by a faculty director/grant PI and a steering committee composed primarily of faculty. The core leadership team at eScience and BIDS also includes a non-faculty executive director(s), who oversee the day-to-day operations of the centers and help design and implement the programs. We concluded that the executive director position was important to the environments. Numerous participants interviewed highlighted the contributions of these staff to the internal functions of the MSDSEs and to building relationships on campus. The departure of the executive director at BIDS was universally widely seen as a serious setback for the center, further underscoring the significance of this position. CDS relied on a more junior manager and an outreach coordinator. While the work of these staff was universally appreciated, many participants (including the leadership) noted that it would have been helpful to also have hired an executive director. Based on the experience of BIDS and eScience, we

concluded that graduate-level training and/or a background in academic administration were important to this position.

Non-faculty staff formed the core of each MSDSE. While the titles and duties of these participants varied across the sites, they could be grouped into three broad categories: traditional postdocs, "super-postdocs," and data scientists. All MSDSEs had traditional postdocs, who were recruited either internally or externally, and were either fully funded by the sites or co-funded with departments. The primary objective of this track was to build relationships across the university. We found that traditional postdocs were the least engaged in the centers, possibly due to their dual affiliation, but the MSDSEs nevertheless viewed this track as worthwhile. CDS planned to keep these positions as partially funded and eScience as unfunded. BIDS was looking to expand this track as one approach to sustaining and growing the center. "Super-postdocs" – independent researchers described to us as assistant professors without teaching duties – was a career track used only by CDS, although the data scientists discussed below were in some ways similar. Selected through a competitive national search, super-postdocs are fully salaried through the center and have few non-research duties.

Finally, all three MSDSEs employed data scientists, also called research scientists or software engineers. At eScience, this position has two tracks: full-time data scientists recruited externally and part-time research scientists recruited internally, but they are similar in terms of staff skills and duties. Data/research scientists at eScience divide their time between collaborative research projects, consulting services, and running programs at the center. To make these positions more attractive, eScience offers relatively high salaries, flexibility to choose responsibilities, a PI status, and incentives to apply for the independent funding. Moreover, having enough people in these positions allowed eScience to launch successful programs that did not get off the ground at the other sites. In our view, eScience created a model career track for data scientists.

Based on our review of the three centers, we conclude that a critical mass of participants is necessary to run a data science center. Researchers at all career stages – data scientists, postdocs, graduate students, and faculty – can successfully staff a center, as long as they have the flexibility to define their duties according to interests and professional aspirations. It is probably worthwhile for a center to include several longer-term staff to anchor its programs and preserve the institutional history. Data scientists seem to be particularly well-suited to this role, as they have broad interests as well as valuable skills, making their positions relatively secure. In addition, data scientists are typically not interested in faculty positions and under less pressure to publish than postdocs. A staff of 4-5 data scientists heavily involved in running the center and a similar number of postdocs and/or graduate students contributing a portion of their time, along with a few faculty mentors (who are offered incentives to remain involved), could be a good model for a data science center that combines research with community programs.

Space

Physical space had both practical and symbolic value, clearly emerging as an important component of the environments at all three universities. We heard that having space where it is at a premium signaled confidence from the administration and elevated the status of the centers. The location of BIDS in the Doe Library, the physical and metaphorical heart of campus, was particularly fitting for becoming a campus hub. The MSDSEs used the open layout of their

spaces to convey its inclusive and neutral nature, although UW and UCB quickly discovered that more private/quiet space was needed. Being the last of the three centers to obtain space, CDS incorporated these considerations in its design; the center seemed to flourish after the move.

What challenges have the MSDSEs experienced and how have these been addressed? How have the environments evolved?

Many fellows and postdocs at BIDS and some at eScience said that the involvement of faculty in the centers was limited. This view was confirmed by faculty members themselves, who cited competing duties as the reason for not playing a larger role. CDS is the only center where faculty have offices in the space and can be found there on most days. Our respondents at CDS told us that based on their experience it was unrealistic to expect that faculty who are not officially affiliated with the center would be more than superficially involved in it. Based on these findings, we believe that some incentives are probably necessary to recruit and retain faculty. These could include offices in the space, salary supplements, teaching buy-outs, and/or funding for students.

We found that none of the centers have yet found an optimal approach to mentoring of non-faculty staff. While satisfaction with mentoring so come extent correlated with the nature of the position, most staff would have preferred more support than what was being offered with grant writing, publishing, and job search. The MSDSE leadership is clearly aware of this challenge and has been experimenting with various models. For example, eScience and BIDS created a scientific research officer position, whose responsibilities include oversight of postdocs and data scientists. CDS considers which fellows are the best match to existing faculty during the application process to make it easier to find mentors when these researchers arrive at NYU.

The joint postdoc career track appeared to be a mixed success. At BIDS, these researchers initially struggled with dual demands on their time; at the time of the last visit two years later, they were uncertain what was expected of them. At eScience, postdocs participated in the center events, but their primary focus was on research projects. Finally, non-faculty staff at NYU and UCB would have liked to be more involved in the governance of the centers. The leadership at both universities is taking steps to be more inclusive and transparent by organizing community breakfasts (NYU) and inviting fellows to Executive Committee meetings (UCB). We are uncertain whether these efforts have addressed the challenge.

In addition to career tracks, the MSDSEs have experimented both with the working group framework and with the activities the groups launched. Initially, the three sites settled on six shared working groups and one unique to NYU. These included Education, Careers, Tools and Software, Reproducibility and Open Science, Space and Physical Organization, Evaluation and Ethnography, and Methods (NYU only). As the environments matured, some working groups dissolved, either because their mission had been accomplished or because the participants' were no longer interested or had time. For example, the Space Working Group became defunct after the centers moved. After the initial burst of activity that produced a book, interest in the Reproducibility Working Group seems to have declined at UW and UCB (NYU has a staff member committed to this area, who keeps it active). At UW, the Reproducibility and Software working groups merged into one Special Interest Group in Reproducible Science and Open Source Software because of the overlap in interests and staff involved.

At the same time, a new type of working groups began to emerge, pioneered by UCB. These groups are organized around data types rather than topics and include ImageXD, TextXD, and GraphXD, for image and text processing and for analyzing graphs (or networks), as the titles suggest. It is our impression that these groups are somewhat similar to the Methods Working Group unique to NYU.

When asked to comment on the utility of working groups, the MSDSE participants said that they were helpful for organizing people and programs in the beginning, but should be allowed to disband, and that new groups created based on the needs and interest of the community.

To what extent are the MSDSEs accomplishing their stated goals?

As discussed in the Introduction, the MSDSE program has three goals: to develop and maintain collaborations between domain scientists and methodologists, to establish rewarding and sustainable career paths, and to contribute to the ecosystem of analytical tools and research practices. We found that the centers made significant strides in each area.

Creating collaborative environments

The MSDSEs used their administrative and physical locations to project the image of a community-building hub open to researchers of all backgrounds and levels of expertise. The MSDSE leadership also made a concerted effort to staff the centers with junior scholars who had a record of collaboration in addition to academic accomplishments. Finally, each MSDSE used multiple venues to bring people together: from seminar series and workshops, often organized around methodologies rather than disciplines, to community-building office hours and training sessions, to internal grants, to interdisciplinary teams addressing a shared problem. In the course of the evaluation, we collected extensive evidence that these efforts led to learning, acquisition of skills, and collaboration. Numerous faculty, fellows, and postdocs interviewed told us that being part of the MSDSE had broadened their horizons, led to the acquisition of skills and partnerships, and in some cases resulted in changes to their research direction. Similar findings emerged from the survey of participants in the Seed Grant and Incubator programs. Finally, each MSDSE reported dozens of grants, publications, and software products that involved multiple staff.

Establishing career paths

The MSDSEs used a combination of traditional positions (such as full-time postdocs) and more innovative positions (part-time postdocs, data scientists/engineers, scientific research officers) to staff the centers. We interviewed many of these researchers, including several alumni, about their experiences, career aspirations, and benefits of participation. All respondents enjoyed the vibrant environment of the MSDSEs and appreciated the flexibility to spend their time developing software, which many saw as being outside of the normal academic experience. Most postdocs and fellows who entered the job market were able to secure a tenure-track faculty position. The perceived benefit of participation was mixed, however. In general and perhaps not surprisingly, those who sought academic positions that were interdisciplinary and/or focused on data science, found that a MSDSE postdoc was an asset. On the other hand, some (although not all) researchers interviewing at more traditional departments had to justify or even downplay their interest in data science and the MSDSE position specifically. In our view, these challenges highlighted the importance of faculty mentors to guiding postdocs through the job search process. While almost all of the postdocs interviewed preferred to stay in academia,

many were willing to move to industry if this was necessary to continue working in data science.

Most data scientists/research engineers were not interested in a faculty track, but preferred to stay at a university for its intellectual freedom, mentoring opportunities, and other advantages over industry. These researchers saw their MSDSE positions as ideal for the time being, as it allowed them to do the work that they enjoyed. At the same time, some expressed concerns about the long-term security of the position and lack of advancement within the career track. These limitations notwithstanding, we believe that the MSDSE grant played an important role in demonstrating the value of data scientists to universities, setting the stage for institutionalizing these positions in the future.

Contributing to the ecosystem of tools and practices

While the MSDSEs created over 200 software products between 2015 and 2017, we were unable to measure the dissemination, use, and impact of these tools beyond a few that were described in peer-reviewed journals. However, we found that nearly all of the tools were posted on GitHub, which was consistent with the participants' commitment to open and reproducible science. Importantly, not only did these researchers personally espouse these practices, they disseminated them to the broader scientific community by publishing a book of case studies drawn from their own projects and through workshops and consultations to promote the relevant skills and tools (e.g., docathons, use of reprozip).

What institutional and cultural changes have occurred, and can any of these changes be attributed to the MSDSE funding?

While three or four years is a short time for cultural changes to occur, particularly at universities that are steeped in tradition, we were able to document examples of institutional changes that were linked to the grant. At BIDS, members of the Education Working Group participated in the development of a new course known as "Data 8," which will change the education of thousands of Berkeley students. Through the course, and by creating the community and the enthusiasm around the center, BIDS contributed to the establishment of a new Division of Data Science at Berkeley - one of the largest reorganizations of the university in more than a decade. Finally, Berkeley created a faculty line for one of the BIDS research scientists, formally recognizing the value of his contributions and setting a precedent for this career path. Examples of institutional changes at UW include data scientist positions, new degree options at the graduate and undergraduate levels across a very large number of units (and growing), new courses including a complete introductory set of three parallel courses, and the inclusion of data science as an area of interest by several departments. The NYU MSDSE developed and successfully implemented a new process for hiring joint faculty.

Are the successful components of the environments sustainable?

As of May 2018, only eScience has secured long-term funding from the university at the level of $1.75 million per year. While this amount can sustain the center's space, programs, and staff at the current level, the leadership is strongly interested in identifying additional sources of support to pilot new programs and career tracks. The NYU MSDSE also seemed to be on a sustainable path through a combination of tuition revenue, university support, and grant funding, but we are uncertain about its direction under the new CDS leadership. Similarly, it is unclear whether the Division of Data Science will provide any financial support to BIDS and at

the time of our last site visit in May 2018, the leadership of BIDS was very focused on strategic planning and fundraising.

What are the characteristics of the data science programs established by other institutions and how do they differ from MSDSEs?

As part of the evaluation, we reviewed data science entities at 17 universities and found that they shared several key elements with the MSDSEs. Virtually all were based outside of departments, and many in open-plan spaces, to signal their interdisciplinary and inclusive nature. Similarly to the MSDSEs, most centers were committed to data science education. A professional master's degree was a particularly popular choice, but some centers were also establishing undergraduate and doctorate degrees and concentrations. Like MSDSEs, the centers offered internal grant programs, postdoctoral fellowships, and community-building activities (e.g., seminar series, workshops, summits). These similarities were perhaps not surprising, as the design of the MSDSE program was informed by the known, and presumably shared needs of the data science community. Furthermore, some universities had consulted with the MSDSEs and may have tried to replicate their efforts. Based on the landscape study, we concluded that the universities used similar building blocks in different combinations to create the entities that best fit their organizational contexts, and that at present there is no consensus for how to organize a data science center.

We also found that the leaders of the centers shared a strong commitment to serving the entire university community over the narrow departmental interests. Several faculty directors mentioned both the importance and the challenges of engaging a broad range of stakeholders when launching a center, suggesting that successful leaders need diplomatic skills, patience, and the ability to articulate the broader benefits of data science.

Roughly half of the non-MSDSE universities employed data scientists or engineers, and several others were interested in hiring these researchers. These positions were lower salaried than in industry and contingent on grant funding, but offered more intellectual environment and opportunities to mentor and teach, which appealed to data scientists. Several center directors commented that data scientists were very important to their community, but that these positions were challenging to create at universities. These views echoed what we heard at MSDSEs.

Did the Foundations select the right strategies to achieve their stated goals?

We found that the MSDSE grantees were chosen based on the perceived enthusiasm for data science, support from the administration, and a collaborative culture at applicant universities rather than particular pre-existing expertise or programs. In our view, this strategy played a positive role in creating the environments as the energy and the commitment to move forward were already in place. The Foundations selected the environments which were at a different stage in its development at the time of award. As all MSDSEs made significant progress to accomplishing program goals, the funding strategy seems to be robust. In the future, the Foundations can intentionally choose the grantees to affect transitions through the development continuum.

We found that the Foundations were able to create a highly exploratory program model by engaging grantees in the design of the environments, encouraging experimentation, and

articulating expectations for false starts. Simultaneously, the unusually close partnership between the funders and the grantees allowed all parties to learn from the program and adjust in real time. Relatedly, funding a cohort of three centers played a positive role in their evolution. We observed a strong interest among MSDSEs to learn from each other and a willingness to course-correct if necessary. In addition, having peers who received the same resources to achieve similar goals probably led to healthy competition between the centers and promoted growth. Finally, the program created a larger community of like-minded scientists and enabled collaboration and knowledge-sharing.

A working group-based framework for the centers was a good approach. While some MSDSE participants felt that the number of meetings was excessive early on, they also acknowledged that the groups were helpful for launching programs and attracting staff. We also note that without this organizational glue, the MSDSE participants would probably largely focus on their own research programs and would not play the same community-building role. We are uncertain whether the future centers should begin with the groups that proved popular at MSDSEs or design their own framework.

The landscape review revealed that the MSDSE program was one of the first efforts to promote data science at the institutional level. It was our impression based on interviews with center leaders, that some of the universities that were not selected by the Foundations were spurred to evaluate their capabilities in data science and to make investments to become more competitive in the future. However, we do not have more than anecdotal evidence to support this claim. We are more confident that the program demonstrated the value of data science at the funded institutions and advanced this field.