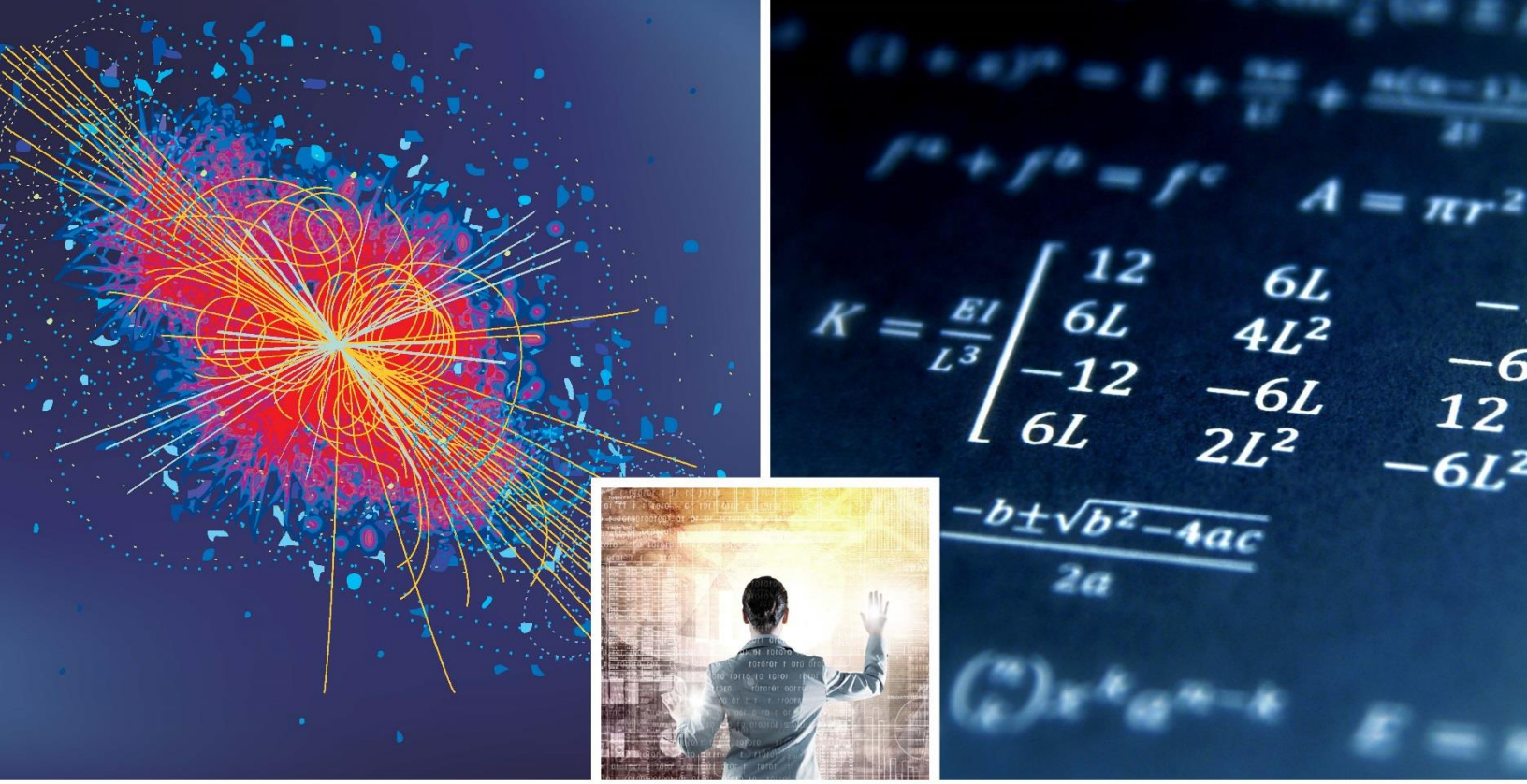


Mid-Term Evaluation of the Gordon and Betty Moore Foundation's Data-Driven Discovery (DDD) Initiative

Final Report

PREPARED FOR:
Julia M. Klebanov
Measurement, Evaluation & Learning Officer
The Gordon and Betty Moore Foundation
1661 Page Mill Road
Palo Alto, CA 94304

SUBMITTED BY:
Carter Epstein
Hannah Thomas
Samantha Burke
Alanah Hall
Marissa Hashizume
Luba Katz

With the assistance of:
Beth C. Gamse

Abt Associates
55 Wheeler Street
Cambridge, MA 02138

JULY 17, 2017

Acknowledgements

We have many individuals to thank for their contributions to this evaluation and report. We would first like to thank Julia Klebanov, Mari Wright, and Debra Joy Perez from the Gordon and Betty Moore Foundation's Measurement, Evaluation & Learning team for wise guidance and insightful feedback throughout the evaluation. In addition, we benefitted from thoughtful comments and suggestions from several others, including members of the evaluation steering committee Dusan Pejakovic, Richard Murray, and Ed Penhoet; Josh Greenberg of the Alfred P. Sloan Foundation; and our Technical Advisory Group, Daniel Katz, Julia Melkers, and Denis O'Gray. We thank you for your time and commitment to the quality of the evaluation.

To Chris Mentzel, Carly Strasser, and Natalie Caulk of the DDD team, and Robert Kirshner, Chief Program Officer for the Moore Foundation's Science Program, we thank you for entrusting us with this responsibility and for the tremendous opportunity to learn so much from you and your colleagues at the Moore Foundation. We have enormous respect for your commitment to science and your passion for the Data-Driven Discovery initiative. We learned firsthand from many of the DDD grantees about your sincere dedication to their success.

Among our many wonderful colleagues at Abt, we thank Sarah Sahni, Luba Katz, and Anna Jefferson for your contributions both to this evaluation and to its companion, the developmental evaluation of the Moore-Sloan Data Science Environments. For your help conducting site visits and interviews, as well as for your collaborative spirit, we thank Porsha Cropper, Djaniele Taylor, and Melissa Velez; for assistance with interview transcription, Audra Nakas; and for expert assistance implementing and fielding the online survey and supporting the survey analysis, Brian Freeman. For her steadfast support, mentorship, and technical review, we thank Beth C. Gamse; this report would not exist without your vital contributions. For editorial and production support, we thank Bry Pollack and Jan Nicholson.

Most importantly, we are especially grateful to all those who participated in the study by completing a survey, taking part in an interview, replying to a follow-up email, or warmly hosting us on campus.

Executive Summary

The Gordon and Betty Moore Foundation (Moore Foundation) contracted with Abt Associates in September 2016 to conduct a **mid-term evaluation** of its **Data-Driven Discovery (DDD) initiative**. The objectives of this evaluation were to assess the DDD initiative's progress to date toward achieving its key goals: to highlight the contributions of data-driven researchers in the natural sciences; to foster the dissemination of software, tools and other science-enabling resources; and to provide compelling exemplars of environments that nurture data-driven scientific inquiry at academic research institutions.

Overview of the DDD Initiative

The DDD initiative was motivated by the growing perception that increasingly data-rich scientific fields were "discovery poor" due to limitations in researchers' ability to exploit these data. Although scientists are generating vast quantities of data at unprecedented rates, to harness these data they must overcome three primary impediments:

- Researchers in the life and physical sciences too often lack expertise (or access to expertise) in computational, mathematical, or statistical methods or tools needed to manage and analyze large and complex data.
- Software, tools, and resources that enable the efficient manipulation and analysis of big data are not widely available, have technical limitations, or are insufficiently reliable.
- The traditional structures and normative practices in academic research institutions do not adequately nurture data-driven science or reward its practitioners.

In November 2012, the Moore Foundation approved the DDD initiative for an initial five-year phase, with a total investment of $60 million (a sixth year received approval in 2015). One of the largest privately funded programs of its type, the DDD initiative comprises three inter-related investment strategies—***People***, ***Practices***, and ***Institutions*** (Exhibit E1)—each of which aligns with one of three objectives. [1]

To highlight the value of data-driven scientists in academia, a key objective of the *People* strategy, the DDD team launched a **Data-Driven Discovery (DDD) Investigator award** competition open to doctoral-level early career or experienced researchers.[2] From an initial pool of more than 1,000 applicants, the Moore Foundation selected 14 awardees in October 2014, each of whom received a $1.5 million, five-year DDD Investigator award.

To drive the creation and dissemination of readily usable tools, methods, and techniques to enable data-driven discovery across the natural sciences, the DDD initiative has funded, to date,

---

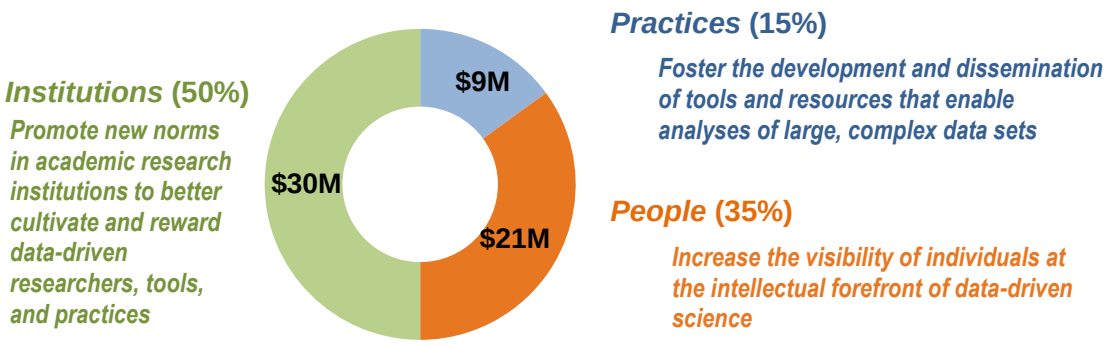
1 http://www.insidephilanthropy.com/science-research/2014/10/20/the-new-recruits-in-moores-huge-data-program.html; accessed July 15, 2016.

2 Early career applicants were those within six years of receiving their PhD; experienced career applicants were those with prior experience as a principal investigator (PI) or co-PI on a research award from a federal agency or private funder.

eight organizations under its *Practices* strategy; four were selected for inclusion in the mid-term evaluation:

- In July 2015, Project Jupyter received a three-year, $1.5 million award (with an additional $1.5 million from the Sloan Foundation and the Helmsley Charitable Trust) to improve the functionality and flexibility of **Jupyter Notebooks**—shareable, interactive electronic notebooks in which researchers store and document programming code data analysis output, visualizations, and text in an integrated platform—and to enhance the capabilities of **JupyterHub**, a multi-user instantiation of Jupyter Notebooks on a central server, cloud, or computing cluster.
- Julia Computing received a two-year, $600,000 award to develop **the Julia language**, an open source programming language that combines fast computational speed needed for processing large volumes of data with high-level, user-friendly scripting.
- Continuum Analytics received a two-year, $700,000 award to enhance **Dask** and **Numba**, two open source packages for use with Python, one of the most widely used scientific computing languages. Dask is a dynamic task scheduler for analysis of very large datasets; Numba integrates with the NumPy package and generates "just-in-time" machine code to optimize compilation speed.
- **Data Carpentry**, which offers training workshops, instructor trainings, and curricula on introductory computational skills geared to scientists' domain-specific data analysis tools, received a $750,000 two-year grant in September 2015 to expand its instructor pool and extend training into new domains (genomics, geosciences, neuroscience, and astronomy).

Exhibit E1: The DDD Initiative's Three Strategies: Key Goals and Allocation of $60 Million[3]



To foster academic environments that nurture data-driven research, researchers, and tools, the *Institutions* strategy targeted change in organizational structures and practices that affect the retention of data-driven scientists in academia, and sought to promote both cross-disciplinary collaborations between computational methodologists and domain-based scientists and training opportunities for researchers to acquire data-driven skills for scientific inquiry. In November 2013, the Moore

[3] Exhibit adapted from materials provided by the Moore Foundation's DDD team at the September 26, 2016, kickoff meeting

Foundation, in partnership with the Alfred P. Sloan Foundation, announced three five-year awards to the University of California at Berkeley (UCB), the University of Washington (UW), and New York University (NYU) for each institution to establish **Moore-Sloan Data Science Environments** (**MSDSEs**). Each institution received approximately $10 million from the Moore Foundation, with additional funding ($2.5 million each) from the Sloan Foundation, to build a community of domain scientists and methodologists engaged in research, to offer formal and informal data science training activities to enable data-driven science to flourish, and to establish new positions and career tracks for data-driven researchers. To foster successful implementation of these new environments, each MSDSE formed Working Groups to address challenges in six priority areas: (1) career paths and alternative metrics; (2) education and training; (3) software tools, environments, and supports; (4) reproducibility and open science; (5) working spaces and culture; and (6) data science studies. Each of these areas was seen as an important "bridge" between domain science/scientists and data science methods/methodologists.

The Evaluation Approach

Three research questions guided the mid-term evaluation:

(1) What results, thus far, has the DDD initiative achieved, and how effectively have the *People*, *Practices*, and *Institutions* strategies contributed to these results?

(2) What role has the DDD initiative played in changes in the data science landscape?

(3) How can positive outcomes of the DDD initiative be sustained? What insights and lessons learned have emerged? What are potential future opportunities for data-driven science?

To address these questions, the Abt team used primary and extant **data sources** including:

(1) **Interviews** conducted in February and March 2017 by telephone or during site visits with

- DDD Investigators (13 of the 14 invitees participated);
- Non-awardee finalists for the DDD Investigator Award (6 of 13 invitees participated);[4]
- Department chairs and an appropriate academic dean or other administrator at five DDD Investigators' institutions (10 of 11 invitees participated);
- Postdoctoral researchers, doctoral graduate students, or other research staff working in five DDD Investigators' research groups (all 11 invitees participated);
- Project leads at each of the four *Practices* grantee organizations included in the mid-term evaluation (all 4 invitees participated); and

[4] This group comprised individuals who had presented their proposed grant activities at the Moore Foundation's offices in the final round of the competition, but ultimately were not selected for the award. One additional non-awardee finalist who was unable to attend the in-person round of the DDD Investigator Award competition was not invited for an interview.

- Project users or key contributors nominated by the *Practices* project leads (and not employed by the grantee organization; 4 of 6 invitees participated).[5]

(2) **An online survey** fielded in March 2017 with DDD Investigators and non-awardee finalists and semifinalists in the DDD Investigator award competition (45 of 93 invitees, 13 DDD Investigators and 31 non-awardees, participated); and

(3) Existing data, including **annual reports** to the DDD team from each of the 14 DDD Investigators, three of the four *Practices* grantees included in the mid-term evaluation, and the three MSDSEs; **interviews with MSDSE lead personnel and administrators (e.g., deans or provosts) at each MSDSE host institution** in the Spring of 2017;[6] **online sources** including grantees' websites, scholarly publications and github postings; reports and commentaries from organizations pursuing goals similar to those of the DDD initiative; and information about federal, philanthropic and industry investments in "big data" or data-driven science initiatives.

Details of the data collection procedures, response rates, and analysis methods are included in the main report and its Appendix B.

Summary of Key Findings of the Mid-Term Evaluation

Evidence from across the evaluation's data sources support the following findings:

The DDD initiative has highlighted the value of data-driven scientists by:

- Enhancing the visibility and credibility of data-driven science and scientists;
- Expanding grantees' capacity by providing funding support for new personnel; and
- Enabling risk-taking and providing flexibility.

The DDD initiative has promoted the development and dissemination of science-enabling software, tools, and resources:

- All 14 DDD Investigators have disseminated or are developing software or other science-enabling tools for data-driven research, and some indicated that they had had difficulty advancing work on such tools before their DDD award.
- Project Jupyter, with an estimated 500,000 Jupyter Notebooks posted on github by January 2016, has released an alpha version of JupyterLab, an improved interactive development environment, and launched JupyterHub, a multi-user, browser-enabled version of Jupyter for

[5] We completed at least one such "user/contributor" interview for three of the four *Practices* grantees; despite repeated requests, one of the *Practices* grantees did not provide names of interview respondents. For Data Carpentry, we asked the project lead to nominate trained instructors (typically graduate students) whom we could invite to participate in an interview.

[6] To prevent duplication of effort and minimize burden on the MSDSEs, Abt and the Moore and Sloan Foundations agreed that the mid-term evaluation of the DDD initiative would use data from interviews collected as part of a separate evaluation of the MSDSEs that Abt is conducting under a current contract with the Sloan Foundation.

cloud or high performance computing environments, which more than a dozen universities have implemented.

- Data Carpentry has capitalized on its DDD grant to become a self-sustaining, transparently managed organization with over 800 volunteer instructors trained, and is helping to meet the high demand among scientists from a wide range of domains for training in data organization and analysis tools for research.
- The Julia Language, despite not yet having released version 1.0, has witnessed dramatic growth in the past few years, becoming one of the top ten programming languages in active development on github by May 2017 with a growing number of advocates for its applications to scientific domains such as astronomy, bioinformatics, geosciences, statistics, neuroscience, quantum physics, and data visualization.
- All three MSDSEs have research scientists, fellows, and other personnel who are actively producing and sharing a large number of tools for data-driven inquiry across multiple scientific domains, including both tools addressing particular domain-specific challenges and tools that have broad applicability across scientific domains.

The DDD initiative has begun to demonstrate the importance of new academic environments that nurture data-driven research and researchers, although some challenges remain. With respect to this goal, the DDD initiative has:

- Played a major role in catalyzing academic institutions' provision of training opportunities in data-driven skills for scientific inquiry;
- Fostered robust collaborations between computational methodologists and domain-based scientists; and
- Had a limited effect, to date, on promoting changes in academic research institutions' mechanisms for retaining data-driven scientists in academia.

Evidence for the DDD initiative's role in scientific discovery comes from DDD grantees' robust publication records (or, for *Practices* grantees, other evidence of their role in scientific findings). However, determining the DDD initiative's role in particular scientific findings presents challenges of causal attribution that arise frequently in evaluations of other research grant programs. Moreover, looking for links between the DDD initiative and discoveries is likely premature, given its short number of years relative to the typical time frames for peer review and publication. Although accelerating scientific discovery is an *ultimate* goal of the initiative, evidence reviewed above suggests that the initiative is meeting its more proximate goal: to facilitate the development of the "research infrastructure" (tools and methods) on which data-driven scientists rely. A more robust foundation of data analysis tools and methods (i.e., the short-term goal of the DDD initiative) is necessary for modern scientific inquiry and new discoveries to emerge (the long-term goal of the initiative).

Data reveal a network of links between the DDD Investigators, *Practices* grantees, and the MSDSEs:

- Eight DDD Investigators either have collaborated with researchers at an MSDSE or have an affiliation with an MSDSE and actively participate in its community at the institution.

- Data Carpentry and Project Jupyter are deeply integrated with the MSDSEs and the work of several DDD Investigators, and each MSDSE has one or more active users or contributors to the Julia Language.
- Dask was prototyped at the 2015 BIDS Data Structures for Data Scientists workshop, and data science fellows at the eScience Institute contribute to efforts to connect Dask with scikit-learn.

The DDD Initiative has played a role in three key changes in the landscape for data-driven science:

- The MSDSEs may have catalyzed the emergence of new data science initiatives at some academic research institutions, although some of these other universities' initiatives may not share the DDD initiative's focus on discovery in the natural sciences.
- Federal funding for data-driven scientific research has increased, but the DDD initiative was an early leader and continues to play a prominent role.
- There is increasing momentum toward open science and reproducibility, a movement that the DDD initiative has fully embraced and promoted.

As the DDD initiative moves toward the end of its initial phase:

- DDD Investigators and MSDSEs will likely require additional external funding after the DDD grant period concludes to continue supporting research software engineers, research scientists and data science fellows.
- All three MSDSE host institutions signaled enduring commitment to the data science environments, but respondents also raised concerns about continuity of funding.
- Formal career pathways for research software engineers may present a potential test case for the viability of alternative career paths for data-driven researchers in academia.
- Survey respondents representing 30 academic research institutions perceived multiple unmet needs for data-driven research at their institutions, including space to meet with colleagues from multiple domains; access to other data-driven faculty, data scientists and software engineers; and educational initiatives to build capacity of students to contribute to data-driven research.

Potential opportunities for advancing the goals of the DDD initiative include:

- Implementing an institutional-level "Challenges in Data-Driven Science" program to unite domain scientists and computational methodologists at non-MSDSE institutions around a shared problem that they propose. If feasible, such a program could present an opportunity to demonstrate the value of data-driven science at institutions without the impetus or resources to establish an MSDSE-like data science environment.
- Exploring ways to further engage academic research libraries and/or research computing in data-driven research.
- Supporting small-scale, cross-domain and cross-institutional community-building events for data-driven investigators or early career scientists to network and learn from each other.

The DDD Initiative Has Highlighted the Value of Data-Driven Scientists

Enhanced Data-Driven Scientists' Visibility and Credibility

The DDD initiative has raised the profile of DDD Investigators, which has translated into tangible new resources and opportunities. **Ten DDD Investigators interviewed indicated that the DDD Investigator award had validated their credibility as independent researchers and provided new opportunities and access to resources and colleagues. Several administrators at DDD Investigators' institutions concurred** that the awards represented validation, both for the individual researcher and for data-driven science writ more broadly. The award has also helped its recipients secure tangible resources and opportunities such as improved laboratory space, invited talks at their own and other institutions, and appointment to leadership roles in their institutions' data science initiatives or similar research centers.

MSDSE respondents also reported that the DDD initiative had "evangelized" the benefits of data-driven research and built cross-departmental bridges. At each MSDSE, joint appointments of faculty and postdoctoral fellows with academic departments fostered good will and cooperation: departments received a share of funding from the MSDSE for new personnel, who split their time between core MSDSE activities and responsibilities within their home department. One UW respondent indicated that this arrangement has given departments across campus an incentive to hire faculty who combine methodological and domain-specific expertise, and it has enabled the eScience Institute to find "friends for life" among department chairs. Not only do the departments gain an additional faculty member, argued one MSDSE respondent, but they also have a chance to witness the important contributions that these new data-driven scientists can make. At NYU, a university administrator saw the MSDSE's establishment of a common protocol for joint hires as a "clear, replicable path" for the university.

Expanded Grantees' Capacity By Supporting New Personnel

DDD funding gave grantees across the three strategies the ability to hire personnel who expanded these grantees' capacity for core research or development activities—either by contributing expertise in software development or computational methods, or by freeing grantee leaders from operational or administrative tasks.

About half (6 of 13) of responding DDD Investigators reported that their DDD award had given them flexibility to expand their research groups with new types of expertise; a few of these Investigators indicated that other grant funding would not have allowed them to hire some of these individuals. DDD Investigators specifically appreciated the ability to add to their research groups: postdoctoral fellows with specific interest in applying tools and methods from one scientific domain to another; software developers; and (for one DDD Investigator) a computational methodologist at a higher salary level than a postdoctoral researcher would typically receive.

For *Practices* grantees in particular, DDD funding has been transformative: it has allowed grantees to devote full-time staff to projects they had previously pursued in their spare time and provided support for operational, administrative, and outreach tasks, freeing up project staff to devote more time to substantive work on the projects. Other *Practices* grantees credited the DDD funding with freeing them from constraints imposed by commercial clients who are

generally more interested in solving specific problems than in investing in a free and open source tool developed for larger benefit.

At the MSDSEs, the data science fellows, research scientists, and program management staff supported with DDD funding (and with other funding that the Moore-Sloan funds allowed these institutions to leverage) have been critical for a broad range of research and educational activities. One MSDSE leader referred to the fellows as one of the data science environment's key "anchors," a view echoed in similar comments from respondents at the other MSDSEs. At all three MSDSEs these individuals have led Working Groups, contributed to "incubator projects" (MSDSE-sponsored project-based collaborations between methodologists and domain scientists, described below), offered trainings, and seeded collaborations with faculty across each of their institutions. At NYU, the MSDSE award enabled the hiring of talented researchers whose presence transformed its Center for Data Science (CDS) from primarily a master's-degree granting program into a robust research center. Data science fellows at the eScience Institute at UW have participated actively in collaborative, data-driven incubator and Data Science for Social Good projects that have won scholarly awards, attracted local press coverage, and received external grant funding. The data science and computational fellows at BIDS have likewise played key roles in a variety of collaborative activities, including a 2015 two-day Data Structures for Data Science workshop, and the formation of the Image Processing Across Domains (ImageXD) and Text Analysis Across Domains (TextXD) research collaborations.

Enabled Risk-Taking and Provided Flexibility

The DDD initiative has also enabled its grantees to take risks important for their professional advancement and provided flexibility to pursue emerging opportunities. Half (7 of 13) of the DDD Investigators credited the award with providing freedom to pursue potentially risky research agendas; three other DDD Investigators valued the DDD team's openness to unanticipated changes in the research, compared with the more rigid constraints of other funders. DDD funding also allowed DDD Investigators the freedom to focus more on the quality than the quantity of publications, and to work directly on developing methods, software, and tools other types of funding would not support. One *Practices* grantee lead and MSDSE leaders likewise valued the DDD team's flexibility, and encouragement to experiment with different approaches.

The DDD Initiative Has Promoted the Development and Dissemination of Science-Enabling Tools, Methods, and Resources

DDD Investigators, *Practices* grantees, and MSDSE personnel have made significant contributions to the development of a wide range of software, tools and resources for data-driven science.

All 14 DDD Investigators have disseminated or are developing tools for a variety of data-driven tasks

These tools include domain-specific tools as well as code used to produce analyses in publications or preprints. Others have developed or contributed to tools intended for a wider audience. These include workflow tools to support reproducibility, interactive data visualization tools, and data extraction tools. Some DDD Investigators indicated that moving these tools forward pre-DDD award was difficult, because other grant awards would not have funded these types of tools.

Among the four *Practices* grantees included in the evaluation, three have contributed demonstrably to a more robust infrastructure for data-driven science, as described below; whether the fourth grantee's tools achieve more widespread adoption is unclear to date.

Scientists across multiple domains have embraced Jupyter Notebooks

One of the key goals of Project Jupyter's July 2015 DDD grant was to improve the interface and user experience of Jupyter Notebook. By January 2016, Jupyter estimated there were approximately three million users and 500,000 Jupyter Notebooks on github. The success of Jupyter Notebooks is also indicated by notable examples of their use:[7]

- Data from the 2015 detection of gravitational waves are available in Jupyter Notebook form from LIGO's Open Science Center.[8]
- The entire contents of the bestselling *Python Data Science Handbook* are implemented in free Jupyter Notebooks in a github repository.[9]

Project Jupyter is also progressing well toward other goals of its DDD grant. In 2016, the Jupyter team announced an alpha version of JupyterLab, an interactive development environment designed to make Jupyter Notebooks more modular, powerful, and flexible; they expect to release version 1.0 in summer or fall 2017. JupyterHub, a multi-user, browser-enabled version of JupyterLab, appears poised for similar growth: More than dozen academic research universities have implemented JupyterHub, and the National Energy Research Scientific Computing Center (NERSC) has deployed JupyterHub on its Cori supercomputer.

Since receiving its DDD award in September 2015, Data Carpentry has become a self-sustaining organization

Data Carpentry, officially launched in 2014, has leveraged its DDD grant to hire an Executive Director, Deputy Director of Assessment, Community Development Lead,[10] and Program Coordinator, and to extend its offerings (training workshops and curricula) into new scientific domains. To date, Data Carpentry has developed material for ecologists, genomicists, biologists, and scientists working with geospatial data, with work underway focusing on image processing and neuroscience data. The organization has trained more than 800 volunteer instructors worldwide, helping to build much-needed capacity in the data-driven sciences, as affirmed by published survey data and interviews with DDD Investigators, their postdoctoral and graduate student colleagues, non-

---

7 Also see: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

8 LIGO stands for Laser Interferometer Gravitational-Wave Observatory. For event data in Jupyter Notebook form, see: https://losc.ligo.org/about/

9 VanderPlas, J.T. (2016). *Python data science handbook: Essential Tools for working with data*. Sebastopol, CA: O'Reilly Media. See: https://github.com/jakevdp/PythonDataScienceHandbook

10 Duckles, J., & Teal, T. (2017, June 7). Announcing Belinda Weaver as our community development lead [Blog post]. Retrieved from http://www.datacarpentry.org/blog/community-developement-lead/

awardees from the DDD Investigator competition, MSDSE respondents and Data Carpentry instructors.[11]

The Julia Language is among the top 10 languages in active development on github

As of May 2017, Julia was one of the top ten programming languages in active development on github with more than 8,500 stars and nearly 2,000 forks.[12] (Stars allow users to show appreciation and bookmark a github project for easy access; forks allow programmers to add features or make contributions to the project.) At the most recent user conference (JuliaCon 2017), five days of sessions included examples of applications of Julia to astronomy, biology, ecology, evolution, geosciences, statistics, systems biology, mathematics, machine learning, neuroscience, quantum physics, and visualization. Interview data suggest that Julia users value its combination of a high-level syntax with speed, but also recognize that because Julia has not yet reached version 1.0, "it's not ready for prime time" (DDD Investigator) and is "still growing in adoption" (*Practices* project lead).

The MSDSEs have each contributed domain-specific and more general-use tools for data-driven science

A key component of each MSDSE is a Working Group explicitly focused on building software tools for data-driven research. Some of the MSDSE-developed tools with applications to multiple scientific domains include:

- **rOpenSci** (BIDS) is a collection of R-based tools to support interactive access to, and analysis of, scientific data with support for efficient documentation and deposit of data in repositories.
- **Myria Big Data Management Service** (eScience Institute) is a cloud-based service intended to make initial data manipulation more efficient and automated.
- **ReproZip** (NYU's MSDSE) allows scientists to package the data files, libraries, variables, and other features associated with a project for others to explore on any machine.

The DDD Initiative Has Begun To Foster Academic Environments That Nurture Data-Driven Research—Challenges Remain

The DDD Initiative has helped individuals advance professionally but has not effected formal changes in review criteria for data-driven scientists

Given industry demand for employees with data science skills, a major goal of the DDD initiative is to retain data-driven scientists in academia, but the DDD initiative has had a limited effect, to date, on promoting changes in institutional mechanisms for retaining data-driven scientists in academia.

Although the DDD Investigators are highly regarded and the majority (10 of 14) are associate or full professors with tenure (four others are on tenure-track), most administrators at DDD Investigators' institutions interviewed saw no effect of the DDD Investigator award on changing

---

[11] Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. bioRxiv 108555. Preprint. https://doi.org/10.1101/108555

[12] See Claster, A. (2017, May 25). Julia ranks among top 10 programming languages developed on github. [Blog]. Retrieved from https://juliacomputing.com/blog/

formal tenure or promotion criteria for data-driven researchers. Similarly, some MSDSE tenure-track faculty and research scientists at MSDSEs have received tenure or promotion, but the MSDSE institutions have not, to date, adopted changes in the formal review criteria for data-driven scientists. Several university administrators (e.g., deans or provosts) at DDD Investigators' and MSDSE institutions felt that their existing review criteria, coupled with the use of external letters from a candidate's field, adequately captured the contributions of data-driven researchers. Several lead faculty or staff at the MSDSEs disagreed.

Some DDD Investigators and MSDSE leaders expressed concern about the opportunities in academia for data-driven scientists seeking alternative career pathways, but other respondents, including MSDSE lead personnel, postdocs, and administrators at the three host institutions, thought that the unique opportunities available to MSDSE fellows and postdocs would make them attractive candidates when they entered the academic job market. Each MSDSE can also point to a few examples of former data science fellows or postdoctoral fellows who have transitioned successfully to faculty positions in academic research institutions. Nevertheless, to date, the MSDSEs have a relatively small number of "alumni" and there is little data to date to assess recent changes in academic opportunities for data-driven scientists. At the MSDSE host institutions themselves, MSDSE leaders and university administrators were actively exploring ways to make sustainable the new research scientist and data science fellowship positions that the Moore-Sloan support has enabled during the grant period. Respondents suggested potential partnerships with university research libraries or research computing departments; a model similar to medical schools' clinical professorships; and a salary buyback approach, in which a data science fellow who received salary support from an external grant could receive part of that salary back from the university in the form of added research funds.

MSDSE collaborations have nurtured data-driven scientists and demonstrated the value of data-driven science

All three MSDSEs reported several examples of successful collaborations with other university research centers and engagement of MSDSE personnel in a variety of research initiatives. Most notably, UCB and UW (along with UC-San Diego) are collaborating to operate **NSF's Western Big Data Regional Innovation Hub,** one of four regional data science hubs. Multiple faculty, data science and postdoctoral fellows at the eScience Institute work on the **Large Synoptic Survey Telescope** (**LSST**). At NYU, MSDSE personnel contribute to the **DIANA/HEP project**, a software development community for high-energy physics that supports Large Hadron Collider experiments.

To demonstrate further the unique contributions of an environment dedicated to data-driven scientists, the MSDSEs have each established "incubator" programs for one- to three-month project-based collaborations between scientists and data-driven methodologists. The incubator programs accept applications from scientists (or teams) who propose a specific, domain-based problem that would benefit from consultation with experts in data-driven methods or tools. Successful proposals receive a small amount of funding to conduct a short-term collaboration, typically one to three months, and work toward a solution. Multiple respondents from the eScience Institute spoke of the value of the incubator program for facilitating collaboration and building support for the MSDSE across the institution.

The MSDSEs have provided opportunities for training in data-driven skills and methods

The MSDSEs have proven particularly effective at providing training opportunities for data-driven skills, and these training opportunities have further raised the profile of the MSDSEs at their host institutions. Primary responsibility for providing such opportunities rests in each MSDSE's Education and Training Working Group. Through the efforts of these groups, all three MSDSEs have offered Data Carpentry and Software Carpentry workshops, Python boot camps, and topic-specific hackathons (e.g., AstroHackWeek, GeoHackWeek, NeuroHackWeek), as well as various workshops specific to each MSDSE (BIDS' 2015 Multiphysics Object-Oriented Simulation Environment [MOOSE] Framework Workshop; a 2015 NSF-funded graduate Data Science Workshop at the eScience Institute; the 2016 Atlantic Causal Inference Conference co-sponsored by NYU's MSDSE). Notable developments in formal educational offerings linked to each MSDSE include:

- BIDS personnel participated in the campus-wide, faculty-led effort to establish a Data Science education program (DSEP). DSEP's introductory-level Foundations of Data Science (Data8) course launched in the spring of 2016. A BIDS Senior Fellow assists faculty teaching DSEP courses, and the program uses Jupyter Notebooks and a JupyterHub to host course materials and manage student assignments.
- UCB approved the formation of a new Division of Data Sciences, for which it hired an interim dean in May 2017. One UCB administrator directly credited BIDS as the catalyst for this new division.
- Adding to the CDS's highly competitive master's degree program in data science, NYU's MSDSE helped win university approval of a new doctoral degree program in data science.
- NYU's MSDSE Education and Training Working Group is developing an undergraduate data science minor, and is drawing plans for its Introduction to Data Science course based on the precedent set by UCB's Data8 course.
- UW approved both a Master of Science in Data Science program, coordinated through the eScience Institute and, building off the success of the Big Data and Data Science IGERT program, an Advanced Data Science (ADS) Option for doctoral students, with nine participating departments (Applied Mathematics, Astronomy, Biology, Chemical Engineering, Computer Science and Engineering, Genome Sciences, Mathematics, Oceanography, and Statistics).

The DDD Initiative's Role in Scientific Discovery

Although accelerating scientific discovery is one of the *ultimate* goals of the DDD initiative, the more proximate goal of the DDD initiative is to facilitate the development of software, tools, practices, and other kinds of "research infrastructure" on which scientists—particularly those working with large or complex data—increasingly rely. The evaluation revealed clear evidence that the DDD initiative is meeting its immediate goals to support the people, practices, and institutions of data-driven science. Several DDD investigators and MSDSE leaders cautioned that it was relatively early to look for signs of their role in new discoveries, especially given the lengthy process of peer review, revision, and resubmission. Nevertheless, **the DDD Investigators and the MSDSEs provided robust publication**

records in their annual reports to the DDD team, and two *Practices* grantees track their projects' contributions to science on their websites. However, we briefly note just two of several difficulties in determining the unique role of the DDD initiative—or of any specific research grant—in these publications. Many DDD awardees likely had a meritorious and promising pre-award record of publications that likely predicts future publication success even in the absence of a particular grant such as the DDD award. For DDD grantees, moreover, it would be difficult to distinguish publications derived from "data-driven" scientific methods from those not (even experts in the relevant research domains might disagree over such classifications). Finally, given the lack of standards for citing software and similar research products of data-driven scientists, it is not feasible, at present, to trace these contributions to future scientific findings.

DDD Investigators, *Practices* Grantees, and MSDSEs have active interconnections

Data from interviews and annual reports reveal a network of links across the DDD initiative's three strategies and suggest that the strategies are mutually reinforcing. Eight DDD Investigators have collaborated with researchers at an MSDSE or have an affiliation with an MSDSE, which has further enhanced their visibility and expanded their opportunities for collaboration. Data Carpentry and Project Jupyter have active engagement with a range of projects and events at the MSDSEs and each collaborates with a DDD Investigator. MSDSE faculty and data scientists contribute to projects that use Julia or Dask.

A DDD Investigator and BIDS data science fellow helped co-found Data Carpentry, and continue to develop workshops with the organization. Both BIDS and the eScience Institute have formal partnerships with Data Carpentry, and the organization has participated in hackathons and other events at the MSDSEs. BIDS provides Project Jupyter a home, and the Jupyter team has several collaborations with BIDS fellows. A custom JupyterHub hosts courses for UCB's Data Science Education Program. Jupyter and DDD Investigator Matthew Turk are collaborating on a scientific workflow project.

The DDD Initiative's Role in a Changing Landscape for Data-Driven Science

The broader landscape for data-driven science has clearly changed over the past five years, based on evidence of progress in several key areas:

(1) The prevalence of data science initiatives at major research universities is increasing, and there is evidence that the MSDSEs may have catalyzed some of these initiatives; nevertheless, some of these programs focus on applied sciences in contrast to the MSDSEs' focus on basic discovery in the natural sciences:

- All three MSDSEs reported multiple inquiries from other academic research institutions about their data science environments.
- Eight of the 15 universities invited by Moore Foundation and the Sloan Foundations to compete for an MSDSE award have since launched new "big data" or data science initiatives.

(2) Although there is now more federal research support for data-intensive science research in academic institutions, the DDD initiative was early to identify and act on this funding need and continues to play a prominent role in supporting data-driven science:

- The National Institutes of Health (NIH) and the National Science Foundation (NSF) launched major "big data" initiatives in 2013.
- However, the DDD initiative remains one of relatively few sources of funding for researchers approaching basic (i.e., non-applied) scientific inquiry with a data-driven lens, and for organizations developing science-enabling tools.

(3) There is increasing traction in the progress toward open science and reproducibility, and the DDD initiative has fostered this progress by:

- Supporting researchers who have made important contributions to open science and reproducibility;
- Funding the development and dissemination of tools such as Jupyter Notebooks that enable reproducible research practices; and
- Including an explicit focus on reproducibility as one of six key themes of its MSDSEs.

Sustaining Key Successes, Remaining Challenges, and Potential Opportunities

Sustaining DDD Investigators' data-driven research labs and the MSDSEs' future will require external funding

Both individual investigators and data science environments will likely continue to require external funding after the end of their DDD grant period to sustain productive programs of data-driven research and opportunities for training and cross-disciplinary collaboration. About a third of the DDD Investigators hired software engineers and/or research scientists above the level of a postdoc; all three MSDSEs hired research software and computational methodologists; and at least three of the four *Practices* grantees included in the mid-term evaluation hired research software engineers.[13] DDD Investigators will need to fund positions that add data-intensive expertise to their research teams, and MSDSEs will likewise need to support a critical mass of researchers with such expertise, as well as administrative staff necessary to run the programs.

Despite concerns about funding, evidence points to continued institutional commitment to the MSDSEs

Interviewees at MSDSE institutions voiced concerns about the availability of post-DDD funding to sustain a critical mass of fellows, research scientists and other personnel and apprehension about possible changes to the structure of the MSDSEs. Despite these concerns, institutional commitment to the MSDSEs is evident at all three host institutions.

- At UCB, BIDS has made visible contributions to the Data Science Education Program, a faculty-led initiative that has generated widespread enthusiasm. Likewise, university administrators saw BIDS as a catalyst for the new Division of Data Sciences. Both developments auger well for BIDS, although its role in the new Division is not yet clear.

[13] Arguably, the DDD grant allowed the Data Carpentry to hire an individual with data-driven expertise in bioinformatics as its Executive Director.

- At UW, the university has demonstrated commitment to the eScience Institute by approving half-faculty lines and research scientist positions and by approving a new data science master's degree program and the Advanced Data Science option for doctoral students. The success of the incubator and Data Science for Social Good programs has also helped build support for eScience.
- NYU's MSDSE received half-faculty lines and two research scientist positions from the provost. MSDSE faculty and researchers participate actively in the CDS data science master's degree program (e.g., mentoring student capstone projects) and the university has just approved a new data science doctoral program.

To sustain momentum in the dissemination of data-driven tools and practices, the scientific community needs standards for citing software

More than 90 percent of scientists agree that software plays an important role in their research, but software is not cited consistently, and informal acknowledgments often lack crediting information.[14] Standardized citation of software will encourage scientists to acknowledge the contribution of software to their research. In turn, citing software in research reports will enable assessments of the role of software in scientific discoveries. Proposed guidelines for citing software have emerged in the past two to three years,[15,16] and increasing attention of research funders and publishers to data and software citation issues also suggests that the scientific community may soon converge on a set of such principles.[17,18]

Research Software Engineer positions may provide a viable alternative career path for data-driven scientists in academia

The Engineering and Physical Sciences Research Council (EPSRC) in the U.K. is experimenting with Research Software Engineering (RSE) Fellowships to support early career doctorates who want to provide software "that is used as a research tool in science and engineering" in academic institutions. An inaugural conference of research software engineers in 2016 drew more than 200 attendees, including funders, academic researchers, industry representatives and research software engineers from 14 nations (attendees included a DDD Investigator, non-awardee, BIDS data science fellow, and

---

14 Howison, J. & Bullard, J.A. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology, 67*, 2137-2155.

15 Smith, A. M., Katz, D. S., Niemeyer, K. E., & FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science 2*:e86. https://doi.org/10.7717/peerj-cs.86

16 Gent, I., Jones, C., & Matthews, B. (2015). *Guidelines for persistently identifying software using DataCite*. [Report.] Swindon, UK: Science & Technology Facilities Council. Retrieved from https://epubs.stfc.ac.uk/work/24058274

17 White, O., Dhar, A., Bonazzi, V., Couch, J., Wellington, C. (2014). *NIH Software Discovery Index Meeting Report*. [Report]. Bethesda, MD: NIH. Retrieved from http://www.softwarediscoveryindex.org/

18 Stodden, V., Guo, P., Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE, 8*, e67111. Retrieved from https://doi.org/10.1371/journal.pone.0067111

other DDD stakeholders). These developments indicate growing recognition that new career pathways are needed for academic research institutions to retain this type of expertise.

Unmet needs for data-driven research

More than 60 percent of survey respondents representing 30 different academic research institutions saw multiple unmet needs for data-driven research at their institution. These needs fell into three main categories: (1) space to meet with colleagues from multiple domains; (2) access to other data-driven faculty and to data scientists and software engineers; and (3) educational initiatives to build capacity of students to contribute to data-driven research. Respondents gave highest priority to the need for their institutions to:

- Hire more full-time, permanent data scientists or software engineers;
- Hire junior faculty with data-driven expertise;
- Incorporate additional training in data-driven methods or tools into existing degree programs; and
- Create interdisciplinary centers for data-driven research.

The fact that the majority of respondents perceived unmet needs across several elements of their working environments suggests that these elements function synergistically, requiring institutions to consider a portfolio of coordinated initiatives to effect change.

Potential opportunities

Interview data yielded three possible avenues to continue and enhance the early momentum of the DDD initiative:

- Sponsoring an institutional-level "challenges program" to unite data-driven domain scientists and computational methodologists from non-MSDSE institutions around a shared problem that they propose;
- Further exploring partnerships with academic research libraries and/or research computing to identify mechanisms of support for data-driven research; and
- Supporting regular symposia or similar small-scale events for data-driven investigators or early career scientists from different domains and different institutions to meet and build a community.

Conclusion

As the initial funding phase nears conclusion, it is already clear that the DDD initiative has had a strong imprimatur on data-driven science. The initiative has been at the forefront of interest and engagement in this area at academic institutions and among research funders, and has made common cause with associations of scientists advocating for more transparent and reproducible research practices. The initiative has also filled a gap by devoting resources for fundamental tool development to enable scientific inquiry.

Despite signs of increasing attention to the needs of data-driven science, the DDD initiative remains unique in its orientation and strategies. Advancing the initiative's goals further may require continued

attention to maintain an emphasis on two critical needs: (1) viable career pathways in academia for research scientists, particularly software developers and computational specialists; and (2) support for those organizations (or emerging organizations) that focus on providing a broad suite of tools and resources for data-driven science. While a five- to six-year investment strategy may seem long at conception and its outset, it is also a relatively short amount of time in which to achieve goals that only now may be gaining momentum.

Introduction

The Gordon and Betty Moore Foundation (Moore Foundation) contracted with Abt Associates in September 2016 to conduct a **mid-term evaluation** of its **Data-Driven Discovery (DDD) initiative**. The objectives of this evaluation were to assess the DDD initiative's progress to date toward achieving its key goals: to highlight the contributions of data-driven researchers in the natural sciences; to foster the dissemination of software, tools and other science-enabling resources; and to provide compelling exemplars of environments that nurture data-driven scientific inquiry at academic research institutions.

The DDD initiative is an ambitious endeavor to call attention to the value of data-driven science and to cultivate it with targeted investments in individual scientists, science-enabling software, tools, and other resources, and in academic research institutions. Though scientists now have access to vast quantities of complex data, they need new types of expertise and new computational tools and methods to harness that data's potential. To cultivate this expertise and foster the development of new tools, academic research institutions need to provide environments that reward researchers' investments in this science-enabling infrastructure and that promote meaningful collaborations among computational, statistical, and natural scientists.

The DDD initiative comprises a $60 million investment allocated across three strategies:

- A ***People*** **strategy**, which granted 14 DDD Investigator awards to researchers from a diverse set of scientific domains;
- A ***Practices*** **strategy**, which has provided support to organizations developing software, tools, and trainings for data-driven inquiry; and
- An ***Institutions*** **strategy**, which, in partnership with the Alfred P. Sloan Foundation, has funded three academic research institutions to establish Moore-Sloan Data Science Environments (MSDSEs).

The mid-term evaluation of the DDD initiative began in September 2016 with a kickoff meeting attended by the Abt evaluation team, the Moore Foundation's DDD initiative program staff (DDD team), Measurement, Evaluation and Learning (MEL) team, an evaluation steering committee, and a representative from the Alfred P. Sloan Foundation. Following this meeting, the Abt, MEL and DDD teams collaborated to finalize details of the evaluation, including a combination of qualitative and quantitative data sources and analyses, an evaluation timeline, and recruitment of an external Technical Advisory Group (TAG).

Data sources included interviews (conducted in person, by telephone, or by WebEx); an online survey; secondary data analysis of grantee annual report materials provided by the DDD team; and extensive reviews of literature and online sources (e.g., individual researchers,' institutions,' and other grantee organizations' websites). The evaluation was also informed by, but independent of, a three-year developmental evaluation of the MSDSEs co-funded by the Sloan Foundation, also

conducted by Abt, under separate contracts.[19] To avoid duplication of effort and overburdening MSDSE stakeholders, Abt and the Moore Foundation staff agreed, to the extent possible, that the mid-term evaluation of the DDD initiative would coordinate with the developmental evaluation of the MSDSEs to leverage the data available from that effort.

Organization of the Report

The report begins in **Chapter One** with a brief overview of key elements of the DDD initiative, including its motivation and goals, followed by a description of the evaluation approach. **Chapter Two** presents key results of the DDD initiative and the role of its strategies in these outcomes. **Chapter Three** considers the role of the DDD initiative across the landscape of data-driven science, including changes in enthusiasm for data-driven science and changes in the environments in which data-driven scientific inquiry occurs. **Chapter Four** examines the sustainability of the DDD initiative's gains, remaining challenges or unmet needs, and potential future opportunities to foster data-driven science. The report concludes with final thoughts and a brief summary of the strengths and limitations of the evaluation.

Appendices include:

- Brief profiles of each MSDSE (Appendix A);
- Details of data collection procedures and analysis methods (Appendix B);
- Interview protocols (Appendix C);
- The survey questionnaire (Appendix D); and
- Supplemental exhibits (Appendix E).

Approach to Protecting Privacy

To preserve the anonymity of interview respondents, we have omitted the names of respondents and their institutions, along with other characteristics (e.g., specific research domain) that could identify an individual person or institution, except in cases when we report publically available information (e.g., online blog posts, publications, github posts). In addition, because of the small total number of women in the respondent pool, we have altered gender pronouns for both male and female respondents at random. In some cases, we identify the individual MSDSE when the information comes from annual report materials submitted to the DDD team and the name of the MSDSE is a critical part of understanding the reported information.

Abt also took precautions to safeguard data security. Interview transcripts and survey data have been stored on a FedRamp-certified Amazon Web Services server. Prior to analysis, we replaced names and other identifying data with unique codes assigned to each case. By prior agreement with the MEL team, and as communicated to respondents in consent forms, Abt will not share interview notes, audio-recordings, or data files with the Moore Foundation.

[19] In Year 1, a contract with the Moore Foundation under direction of Chris Mentzel; in Year 2, a contract with the Alfred P. Sloan Foundation, under direction of Josh Greenberg.

1. The Data-Driven Discovery Initiative

1.1 Motivation

The DDD initiative was motivated by the growing perception that increasingly data-rich scientific fields were "discovery poor" due to limitations in researchers' ability to exploit these data. Although federal funding agencies and academic research institutions have made considerable investments in the physical infrastructure needed for storing, transmitting, and processing large amounts of data, the "intellectual infrastructure" for data-driven science remains relatively underdeveloped. The large quantities of scientific data may enable transformative scientific discoveries, on one hand, yet harnessing this potential means that data-driven scientists working in academic settings must overcome three impediments:

- Researchers in the life and physical sciences too often lack expertise (or access to expertise) in computational, mathematical, or statistical methods or tools needed to manage and analyze large and complex data.
- Software, tools, and resources that enable the efficient manipulation and analysis of big data are not widely available, have technical limitations, or are insufficiently reliable.
- The traditional structures and normative practices in academic research institutions do not adequately nurture data-driven science or reward its practitioners.

We briefly discuss each of these impediments to data-driven discovery below.[20]

1.1.1 Domain Scientists Need Expertise in Data-Driven Methods and Practices

Typically, scientists learn to manipulate and analyze their data on their own, by adopting the practices of more experienced researchers while they are graduate students or postdocs, or by learning from other mentors and colleagues early in their careers. The resulting ad hoc patchwork of less than optimal software programs, packages, and tools are prey to technical limitations, incomplete documentation, and difficulty of use for individuals not skilled in their creation.[21]

Until the more recent movements toward reproducibility and open software, scientists also lacked experience sharing their data with other researchers, let alone sharing the details of analyses that increasingly were embedded in software or data management tools and thus opaque to others. As the complexity and scale of data and corresponding analyses have grown, scientists have increasingly uncovered errors in analysis (leading to retractions) or found that the reproducibility of scientific findings—a key tenet of the scientific enterprise—is severely constrained, further impeding scientific progress.

---

[20] This discussion omits other barriers that also impede data-driven discovery—notably, the same historical lack of appreciation for its promise among most research funders—but these three points were the primary motivating forces that led to the DDD initiative.

[21] Hannay, J.E., MacLeod, C., Singer, J., Langtangen, H.P., Pfahl, D., & Wilson, G. (2009). How do scientists develop and use scientific software? *Software Engineering for Computational Science and Engineering*. Retrieved from http://ieeexplore.ieee.org/document/5069155/?part=1

1.1.2 Data-Driven Science Needs Better Tools and Resources

Limitations in the availability and design of programming languages, software, and other tools compound the problems caused by lack of expertise in data-driven methods. Scientific computing products developed commercially may be ill suited for the highly specialized problems that scientists tackle, and they cannot be readily adapted (e.g., because the source code is protected by copyright).

Although open source software, programming languages, and compilers provide alternatives to commercial software, those tools may be designed for experienced software developers, or they may lack capabilities (e.g., packages with detailed documentation) needed by an individual scientist or researchers within a domain. Scientists who use such tools to build data management solutions may simply not know about resources developed in other research labs, because there has been little incentive for scientists to disseminate these tools or to adopt shared standards or best practices in data management. Academic research institutions have had little awareness and few mechanisms for rewarding scientists for contributions to this scientific infrastructure. Moreover, scientists lack training in the principles of software development that can help avoid errors in reported results and even the retraction of papers.[22,23]

1.1.3 Academic Research Institutions Need to Better Cultivate Data-Driven Science

Because most basic science in the United States takes place within academic research universities, these institutions play a critical role in the future of data-driven science. The traditional organization of faculty into departments that conduct research within specific domains has further impeded the contributions of data-driven science to new discoveries. Academic departments largely determine which scientists advance professionally. [24] Departments are the primary decision-making bodies for hiring, promoting, and granting tenure to scientists, and these decisions reflect the values of their faculty members. If senior faculty do not recognize the value of data-driven science, data-driven scientists may face limited opportunities and difficulties in professional advancement.

A prevailing emphasis on peer-reviewed publications, coupled with little acknowledgement of the tools or resources that undergird such publications, means that data-driven scientists have little incentive to invest time in developing or sharing the "intermediate" research products—data, software, and similar tools—that play an increasingly critical role in scientific research.

Academic research institutions have also been slow to recognize the potential gains of providing rewarding and sustainable career paths to individuals who may be best poised to support data-driven research. Although there are highly trained, doctoral-level researchers interested in applying their computational or methodological expertise to domain-specific research, not all of them want to compete for the limited number of tenure-track faculty positions.

---

22 Groble, C. (2014). Better software, better research. *IEEE Internet Computing, 18*, 4-8. Retreived from http://ieeexplore.ieee.org/document/6886129/

23 Merali, Z. (2010). Computational science: ...Error...Why scientific programming does not compute. *Nature, 467*, 775-777.

24 National Research Council, National Academy of Engineering, & Institute of Medicine. (2014). *The arc of the academic research career: Issues and implications for U.S. science and engineering leadership: Summary of a workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18627

Yet, there are few other career options for these researchers within academia. Because these experts depend on others' grant funding for their livelihood, they may have limited job security and lower salaries than are increasingly available in private sector data science industry. As a result, research labs too often face cycles of hiring, training, and replacing cadres of term-limited staff whose accumulated knowledge is lost when they depart academia for more secure employment. Consequently, faculty members lose valuable time and resources in their scientific endeavors.

> *[Data-intensive science] is something qualitatively different. It's not the same stuff with more data. When you have vast amounts of data … the focus changes from ownership of data or access to data to ownership of expertise . . . . Somebody may not be the world's greatest astronomer, [or] the world's greatest computer scientist, but performs an extremely valuable role of bridging the two … and people like that tend not to have well-defined career paths in … academia. And we have to change that. If you want to have good people, you have to give them proper professional recognition and reward.*
>
> *—George Djorgovski, Director, Center for Data-Driven Discovery, Caltech*

1.2 About the DDD Initiative

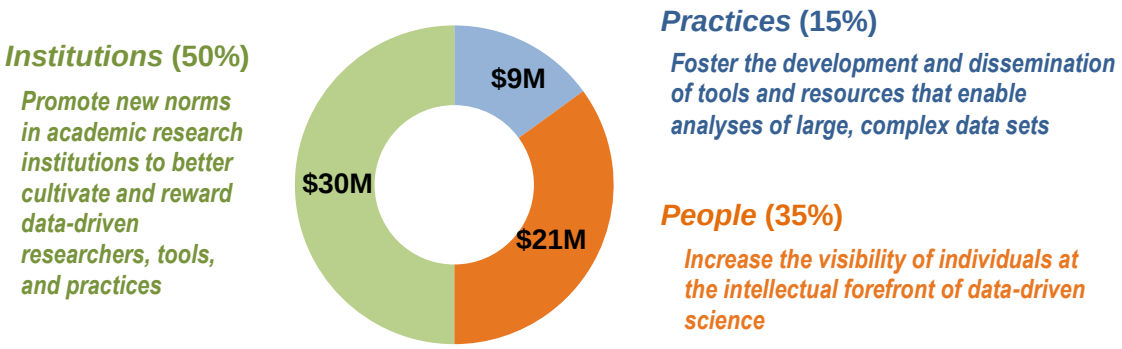
In November 2012, the Moore Foundation approved the DDD initiative for an initial five-year phase, with a total investment of $60 million (a sixth year received approval in 2015). One of the largest privately funded programs of its type,[25] the DDD initiative is intended to

- Increase the visibility of individuals at the intellectual forefront of data-driven science;
- Foster the development and dissemination of accessible, user-friendly tools and resources to handle the increasingly complex data across diverse fields including biology, neuroscience, astronomy, and geosciences; and
- Promote new norms in academic research institutions that will better cultivate and reward the expertise needed for data-driven research.

The DDD initiative distributed its funding across three inter-related investment strategies, each of which aligns with one of these three objectives (Exhibit 1.1). Each strategy was seen as a critical foundation of data-driven science: Data-driven natural scientists (*People*) rely on a suite of software, packages, tools, and other resources (*Practices*) to observe and analyze large, complex, and fast-accumulating scientific data; these investigators work in research universities (*Institutions*) that provide the physical space, intellectual community, and cultures in which research is conducted.

[25] http://www.insidephilanthropy.com/science-research/2014/10/20/the-new-recruits-in-moores-huge-data-program.html; accessed July 15, 2016.

Exhibit 1.1: The DDD Initiative's Three Strategies: Key Goals and Allocation of $60 Million



Source: Adapted from materials provided by the Moore Foundation's DDD team at the September 26, 2016, kickoff meeting

1.2.1 The DDD Initiative's *People* Strategy

The primary goal of the *People* strategy is to highlight the value of data scientists in academia. To achieve this goal, the DDD team launched the Data-Driven Discovery Investigator award (DDD Investigator award). Targeting doctoral-level researchers (not necessarily tenured or in tenure-track positions) in either U.S.-based doctorate-granting institutions or private research institutes, the Moore Foundation solicited applications from early-career (i.e., those within six years of receiving their PhD) or experienced researchers (i.e., those with prior experience as a principal investigator (PI) or co-PI on a research award from a federal agency or private funder). Eligible applicants were those working in the natural sciences, "science-enabling" methodologies, or a combination.

The DDD team and external peer reviewers screened pre-applications to identify 93 applicants who met eligibility criteria and provided a compelling record of past and anticipated future accomplishments. Of these, the DDD team invited 28 to submit a full application and deliver an in-person presentation to the Moore Foundation's staff, external reviewers, and fellow finalists.[26] In October 2014, the Moore Foundation announced the selection of 14 DDD Investigators, each of whom received a $1.5-million five-year grant.

By sponsoring a national competition and providing substantial funding to researchers engaged in data-driven scientific inquiry, the DDD initiative hypothesized that DDD Investigators would serve as exemplars within their own institutions, their research fields, and academia more broadly. By acknowledging these individuals as worthy of recognition, the DDD team hypothesized that others would begin to value more explicitly the potential contributions from a new kind of investigator.

1.2.2 The DDD Initiative's *Practices* Strategy

The primary purpose of the DDD initiative's *Practices* strategy is to drive the creation and dissemination of readily usable tools, methods, and techniques to enable data-driven discovery across the natural sciences. Under the *Practices* strategy, the DDD initiative has funded, to date, eight organizations, some for more than one project; the Moore Foundation identified five of these projects

[26] One of the 28 applicants invited could not attend the in-person event, and did not receive an award.

(four grantee organizations) for inclusion in this evaluation.[27] Three organizations were funded to build software and tools for data-driven scientists (Project Jupyter, Julia Computing, Continuum Analytics), and one organization was funded to train scientific researchers to use data management tools and adopt best practices to support reproducible results (Data Carpentry via NumFOCUS as the fiscal sponsor). Exhibit 1.2 summarizes the award type, amount, start date, and grant durations for each of these four grantees.

Exhibit 1.2: Award Type, Amount, Start Date, and Duration of Five DDD *Practices* Projects (Four Grantees) Included in the Mid-Term Evaluation

Projects (Grantee Organization)	Award Type	Award Amount	Start of DDD Grant Period	Duration (months)
1. Project Jupyter (UC Berkeley)	B	$1,500,000	07/2015	36
2. The Julia Language (Julia Computing)	B	$600,000	10/2015	24
3. Dask (Continuum Analytics) 4. Numba (Continuum Analytics)	B	$700,000	07/2016	24
5. Data Carpentry (NumFOCUS)	T	$750,000	09/2015	24

Key: B=builds software and tools; T=trains researchers

Project Jupyter

Project Jupyter is a set of open source software tools for interactive scientific computing. The project began as iPython, an interactive shell for scientific computing in Python; it subsequently evolved into a more generic tool for building "computational narratives" in any programming language.[28] The "narratives" are called **Jupyter Notebooks**. Jupyter Notebooks are shareable, interactive electronic notebooks in which researchers store and document their research workflow, including programming code used in data analysis, analysis output, data visualizations, and text in an integrated platform. Jupyter Notebooks enable collaboration and reproducibility.

The DDD initiative funded Project Jupyter to improve the functionality and applicability of Jupyter Notebooks, by including a better user interface and greater modularity to give users more flexibility; by adding the ability to move from a single notebook to smaller notebooks with external modules; and by improving the functionality of tools that allows users to export notebooks into other formats (e.g., Microsoft Word); and improved documentation. The grant support extended to enhancements to JupyterHub, a multi-user instantiation of Jupyter Notebooks on a central server, cloud, or computing cluster. JupyterHub has fewer installation requirements for use on local machines and expands the possible uses of Jupyter tools.

---

[27] The DDD initiative is also funding the R Consortium/NumFOCUS for community-driven development of R-language research tools, and the National Academy of Sciences to convene and disseminate the results of four Data Science Education roundtables per year for three years. Other grants included as part of the *Practices* strategy included $1.5 million in seed grants to three non-awardee runners-up in the MSDSE competitio; planning grants totalling $0.7 million to IDEO and the three MSDSE host institutions; and $1.4 million in SCPs.

[28] In 2010, iPython was renamed "Project Jupyter" to reflect the fact that core products (such as the iPython/Jupyter Notebook) support languages in addition to Python.

The Julia Language

Julia Computing received a grant to develop Julia, an open source programming language designed to combine efficiency of computation (often a limiting factor with big data) with high-level, user-friendly scripting. The development of the Julia language was motivated by frustration that no single programming language combined high-level syntax with the fast computational speed needed for processing large volumes of data. Most scripting languages (such as Python) enable fast, easy coding. However, such languages generally depend on a slower compiler that converts the code into machine language, and the process of compiling slows down the processing speed. Julia was designed to compile "on the fly," reducing the intermediate slowdown from typical compilers. Julia is also designed to enable parallel computing, often needed in high-performance computing environments (e.g., supercomputing environments).

Dask, Numba

Continuum Analytics received an award for two related projects, Dask and Numba, both open source Python packages. Although Python is one of the most widely used scientific computing languages, it is not ideal for algorithms that take advantage of parallel computing, nor for allowing these algorithms to process data in distributed environments. Dask is a dynamic task scheduler for analysis of very large datasets that would otherwise overwhelm the finite memory of a typical single CPU.[29] It also features a scheduler that coordinates parallel computations across distributed environments (multiple CPUs). Numba generates "just-in-time" machine code from Python (including integration with the widely used NumPy package) to optimize compilation speed.

The DDD initiative funded Continuum Analytics to bring Dask and Numba up to production-quality standards via the release of version 1.0; to engage in community outreach activities to build acceptance of these tools among core scientific communities; and to establish a community governance structure that enables participation by developers and users.

Data Carpentry

Data Carpentry offers training on introductory computational skills that scientists need for data management and analysis. Launched in 2014, its initial focus was on lessons in ecology (reflecting the backgrounds of its founders). The target audiences now are researchers in the life, physical, and social sciences. Data Carpentry is a sister organization to Software Carpentry, which focuses on best practices in software engineering. The two organizations share a commitment to open source licensing and collaborative lesson development. Data Carpentry's main training forum is a two-day workshop geared to attendees' domain-specific tools and needs. Instructors are volunteers who have successfully completed the joint Software Carpentry/Data Carpentry Instructor Training. In addition to workshops, Data Carpentry also posts curricular and training materials online under a Creative Commons-Attribution Only (CC-BY) license for download and use by anyone.[30]

The DDD initiative funded Data Carpentry to expand its instructor pool and create workshops for scientists in new domains (e.g., genomics, geosciences, neuroscience, and astronomy). Other goals

---

[29] Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In K. Huff & J. Bergstra (Eds.), *Proceedings of the 14th Python in Science Conference* (pp. 130–136). http://conference.scipy.org/proceedings/scipy2015/pdfs/matthew_rocklin.pdf

[30] https://creativecommons.org/licenses/by/4.0/

included incorporating continuous assessment and supporting instructors through an online community.

1.2.3 The DDD Initiative's *Institutions* Strategy

The *Institutions* strategy is designed to demonstrate innovative organizational structures in academia that foster data-driven scientists and practices. This strategy represents 50 percent of the total DDD initiative funds allocated, and it is the most ambitious of the three strategies because it targets changes at multiple levels of the institution simultaneously, as well as changes in long-standing cultural norms in reward systems and the career pathways for data-driven researchers. In partnership with the Alfred P. Sloan Foundation, the DDD team invited 15 academic research institutions to submit a letter of intent to be Moore-Sloan Data Science Environments, selecting six for site visits. Three institutions were invited to participate in a design process starting in May 2013 to develop new academic environments for data-driven discovery. This process led to the formation of several Working Groups, each with its own mandate, to address barriers to implementation, and it culminated in approval by the two foundations of three "linked proposals."

In November 2013, the DDD team announced the MSDSE partnership and selection of the three institutions: University of California at Berkeley (UCB), the University of Washington (UW), and New York University (NYU). The event, sponsored by the White House Office of Science and Technology Policy, highlighted public and private partnerships for big data analysis and research.[31] Each of the three MSDSEs received approximately $10 million from the Moore Foundation, with additional funding from the Sloan Foundation ($2.5 million). Each institution also invested funds into its MSDSE, and all three MSDSEs capitalized on additional funding from private, state, and federal sources.

Two of the three MSDSEs built on pre-existing centers. At UW, the MSDSE built on the 2008 establishment of an eScience Institute that had enabled the formation of research partnerships between methodologists and domain sciences in the biological, environmental, physical, and social sciences. The eScience Institute had received awards from the Moore Foundation, Microsoft Research, and the National Science Foundation (NSF), including an NSF-funded Data Science IGERT (Integrative Graduate Education and Research Traineeship) award in 2012.[32] At NYU, the MSDSE became the research arm of a new Center for Data Science (CDS) and its two-year master's degree program in data science, both developed earlier in 2013.[33] At UCB, the MSDSE began largely as a *de novo* organization, called the Berkeley Institute for Data Science (BIDS).

Each MSDSE brings together domain scientists with data science methodologists from within the institution to lead the implementation of the MSDSE. A key focus of the spring 2013 design process and the first year after award was to establish positions and career tracks for both hired and affiliated personnel. It also was to build a community of scholars engaged in research, formal and informal data

---

31 http://news.berkeley.edu/2013/11/13/new-data-science-institute-to-help-scholars-harness-big-data/

32 The MSDSE at UW also built on its Center for Statistics and Social Sciences, more than 10 years old as of 2013. See https://news.cs.washington.edu/2013/11/12/uw-berkeley-nyu-collaborate-on-37-8m-data-science-initiative/.

33 http://www.nyu.edu/about/news-publications/news/2013/november/nyu-part-of-five-year-initiative-to-harness-potential-of-data-scientists-big-data-with-support-from-moore-sloan-foundations.html

science training activities, lecture series and talks, and other events to enable data science to flourish. Members of this community included existing faculty (tenured or tenure-track) on campus, new faculty hires (both tenure-track and not), other positions not tenure-track such as research scientists, and "data scientists" or "data science fellows." In addition, each MSDSE worked to engage postdoctoral fellows and graduate students.

Each MSDSE also formed six Working Groups comprising MSDSE personnel to collaborate with their counterparts at the other two institutions in the following areas identified during the design process as needing explicit, cross-university attention:

- Career paths and alternative metrics;
- Education and training;
- Software tools, environments, and supports;
- Reproducibility and open science;
- Working spaces and culture; and
- Data science studies (originally called Ethnography and evaluation).

Each of these areas was seen as an important "bridge" between domain science/scientists and data science methods/methodologists: the charge of the Working Groups was to foster successful implementation of the MSDSEs. A profile of each MSDSE is included in Appendix A.

1.2.4 Other DDD Initiative Components

In addition to the awards made in each of the three strategic areas described above, the DDD team also supported several additional components intended to maximize collaboration and the dissemination of new knowledge emerging from DDD stakeholders. These included:

- An annual DDD Investigator symposium, for the 14 DDD Investigators and guests to network;
- An annual MSDSE Summit, where the three data science environments convene to exchange knowledge, provide progress updates, and work towards common partnership goals;
- A postdoctoral/early-career symposium for individuals working in the labs of the DDD Investigators, the data science environments, or associated with *Practices* grants; and
- A 2016 Data Science Summit hosted at NYU, providing an opportunity for the DDD communities to exchange knowledge and interact with representatives from industry, academia, and philanthropic and government funders.

These events were intended to augment the ability of stakeholders within and across each of the DDD initiative's strategies to meet their individual goals, to meet the initiative's goals, and to advance scientific discovery.

1.3 The Evaluation Approach

Working with the Moore Foundation, the Abt team identified three overarching research questions to frame the evaluation, and determined that both qualitative and quantitative data sources would best

address these questions. A brief description of the research questions and data sources follows. (Appendix B provides additional detail on data collection procedures and analysis methods.)

1.3.1 Research Questions

Three key research questions guided this evaluation:

RQ 1. What results, thus far, has the DDD initiative achieved, and how effectively have the People, Practices, and Institutions strategies contributed to these results? To address this question, the evaluation documents the DDD initiative's role in enhancing the visibility of data-driven researchers; in facilitating the development and dissemination of software, computational methods, and other resources to enable data-driven science; and in fostering environments at academic research institutions that better support and reward the contributions of data-driven researchers. We also examine how each of the DDD initiative's three strategies contributed to these results.

RQ 2. What role has the DDD initiative played in changes in the data science landscape? For the second research question, we examine contributions of the DDD initiative to changes in the landscape in which data-driven science is practiced. This landscape encompasses academic research institutions and research funding organizations, as well as trends in the practices for disseminating scientific findings, and tracking and rewarding the contributions of data-driven researchers to these findings.

RQ 3. How can positive outcomes of the DDD initiative be sustained? What insights and lessons learned have emerged? What are potential future opportunities for data-driven science, and what are their relative risks and potential gains? Finally, we explore the sustainability of the results observed to date, focusing on future plans of individual grantees, signs of lasting institutional commitment to data-driven science, and potential opportunities that have emerged from grantees' successes and challenges. The report concludes with a look at potential future opportunities for furthering the DDD initiative's goals.

1.3.2 Data Sources

Exhibit 1.3 summarizes the primary and extant data sources used to address the evaluation research questions. Semi-structured interviews comprised the principal data collection activity, complemented by an online survey and existing data sources.

Primary Data: Interviews

We invited 59 respondents to participate in an interview, and completed interviews with 48 of these in February and March 2017 (an overall response rate of 81 percent; see Exhibit 1.3). Interviews were conducted either by telephone or WebEx, or in-person at site visits to five of the DDD Investigators' institutions (See Appendix B for criteria used to select DDD Investigators for a site visit). Interview respondents included:

- DDD Investigators;
- Non-awardee finalists for the DDD Investigator Award;[34]

[34] This group comprised individuals who had presented their proposed grant activities at the Moore Foundation's offices in the final round of the competition, but ultimately were not selected for the award.

- Department chairs and an appropriate academic dean or other administrator at five DDD Investigators' institutions;
- Postdoctoral researchers, doctoral graduate students or other research staff working in five DDD Investigators' research groups;
- Project leads at each of the four *Practices* grantee organizations included in the mid-term evaluation; and
- Project users or key contributors nominated by the *Practices* project leads (and not employed by the grantee organization).[35]

Exhibit 1.3 summarizes response rates for each type of interview respondent (see Appendix B for more information on response rates, as well as key interview topics; Appendix C has a copy of each interview protocol).

Primary Data: Survey

In addition to interview data, we received 45 responses to an online survey fielded in March 2017 with DDD Investigators and non-awardee finalists and semifinalists in the DDD Investigator award competition (Exhibit 1.3). The survey included questions about respondents' rank, tenure status, and institutional affiliations; characteristics of members of their research groups and their institutions; usage of scientific computing software and other resources (including some tools funded by the DDD *Practices* strategy); perceived unmet needs for data-driven researchers; and for non-awardees only, the impact of the DDD initiative on their own research program and their research field (see Appendix D for a copy of the survey items). The overall survey response rate was 48 percent.

Extant Data

Finally, the evaluation also drew on extant data from DDD grantees (see Exhibit 1.3). These data included:

- Year 1 annual reports from each of the 14 DDD Investigators, including budget, expenditures, and written narratives and data on each grantee's publications, presentations, software and other research products; conferences attended; grants, prizes, honors and awards; and key collaborators.
- Year 1 annual reports from three of the four *Practices* grantees included in the mid-term evaluation, including budget, expenditures, and written narratives and data on each grantee's personnel; publications, presentations, software and other research products; conferences attended; grants, prizes, honors and awards; key collaborators; and usage data.[36]

One additional non-awardee finalist who was unable to attend the in-person round of the DDD Investigator Award competition was not invited for an interview.

35 We completed at least one such "user/contributor" interview for three of the four *Practices* grantees; despite repeated requests, one of the *Practices* grantees did not provide names of interview respondents. For Data Carpentry, we asked the project lead to nominate trained instructors (typically graduate students) whom we could invite to participate in an interview.

36 Because Continuum Analytics (projects include Dask and Numba) received its DDD grant in July 2016, no annual report data were available from this *Practices* grantee.

Exhibit 1.3: Data Sources for the Mid-Term Evaluation of the DDD Initiative

	Primary Data		Extant Data				
	Interviews (N=48)	Survey (N=45)	Grantee Annual Reports	Grantee Websites	Other Online Sources[a]	Year 1 Report, MSDSE Evaluation	Year 2 Interviews, MSDSE Evaluation
DDD investigators	13 of 14	13 of 14	Year 1	✓	✓		
Non-awardees from the DDD Investigator competition[b]	6 of 13	32 of 79					
Administrators from DDD Investigator institutions (department chairs, deans, vice provosts)	10 of 11						
Postdoctoral researchers, graduate students, research staff from DDD Investigator labs	11 of 11						
Practices project leaders	4 of 4		Year 1	✓	✓		
Practices project users/contributors	4 of 6						
MSDSE faculty, staff, Working Group leaders; administrators at host institutions			Years 1-2 (Shared & Individual)[c]	✓	✓	✓	Subsample: 18 of 65 completed[d]

Notes:

[a] The Abt team conducted online searches to identify relevant publications and github postings, to review project websites, to explore artifacts mentioned by interviewees or survey respondents, and when possible, to triangulate across data sources to confirm information reported by interviewees.

[b] The 79 non-awardees included 65 semi-finalists and 14 finalists who did not receive a DDD investigator award. 13 of the 14 finalists (i.e., those who delivered an in-person presentation during the award competition) were invited to participate in an interview; all 79 were invited to participate in the online survey.

[c] In Year 1, the three MSDSEs submitted both a shared report and individual, institution-specific reports. In Year 2, each MSDSE submitted an individual report.

[d] Systematic analyses of all interview data from site visits conducted as part of the MSDSE developmental evaluation was ongoing at the time of this report. As a result, the Abt team examined a small subsample of three to six interviews per MSDSE (i.e., 18 of 65 interviews total) identified by the MSDSE team as most relevant to the research questions of the mid-term evaluation. Respondents included faculty or staff MSDSE leaders, Working Group leaders, and administrators at the MSDSE host institutions

- Interviews with faculty and staff leaders at each of the three MSDSEs and administrators at each MSDSE host institution in the Spring of 2017.[37]
- Year 1 (2014) and Year 2 (2015) shared reports submitted by the three MSDSE host institutions together, separate Year 1 and Year 2 annual reports from each MSDSE host institution, and the MSDSEs' renewal proposals (July 2016) including data on each MSDSE's:
  - Data-science-related hiring;
  - Talks, seminars, and other events related to data-driven discovery at the institution;
  - Data science educational courses and other training events offered at the MSDSE host institution;
  - Participants in MSDSE-hosted events;
  - Grants related to data-driven science that the MSDSE host institution received;
  - Inquiries that the MSDSE or host institution received about the MSDSE or the DDD initiative as a whole;
  - Online resources at the host institution related to the MSDSE or similar initiatives; and
  - Data on individual MSDSE personnel or key affiliates such as publications, presentations, software and other research products; conferences attended; grants, prizes, honors and awards; and key relationships (e.g., collaborators, co-authors).

Other extant data were obtained from online sources, including grantees' websites, online searches for publications and github postings, reports and commentaries from organizations pursuing goals similar to those of the DDD initiative, and information about federal, philanthropic and industry investments in "big data" or data-driven science initiatives.

We now turn to the key findings of the evaluation.

[37] To prevent duplication of effort and minimize burden on the MSDSEs, Abt and the Moore Foundation agreed that the mid-term evaluation of the DDD initiative would use data from interviews collected as part of Abt's separate evaluation of the MSDSEs.

2. Results of the DDD Initiative

This chapter documents the key results, thus far, of the DDD initiative, in terms of its progress toward three key goals: highlighting the value of data-driven scientists; promoting the development and dissemination of science-enabling tools, methods, and resources; and fostering academic environments that nurture data-driven research and researchers. The chapter also examines the contributions of the *People*, *Practices*, and *Institutions* strategies to these results.

2.1 Key Findings

The DDD initiative has highlighted the value of data-driven scientists:

- Ten DDD Investigators and several administrators at these Investigators' institutions noted that the DDD award enhanced the investigator's credibility, leading to new opportunities and resources; for some institutions the award also bolstered the reputation of the investigator's department both with the institution's administration and nationally.
- Joint appointments of faculty, fellows, and postdoctoral researchers at all three MSDSEs helped to demonstrate the important contributions of data-driven scientists and build bridges between academic departments.
- DDD funding allowed DDD Investigators and *Practices* grantees to add staff who expanded their teams' expertise in software development or computational methods, or who freed up researchers' time by attending to operational or administrative tasks; *Practices* grantees in particular indicated that this extra capacity transformed their efforts from activities conducted in spare time to bona fide, funded projects.
- Half of the DDD Investigators reported that the DDD initiative enabled them to take risks in a way that other funding sources did not; both DDD Investigators and other grantees credited the DDD program officers with encouraging them to pursue emerging opportunities, again citing the program as unique among funders.

The DDD initiative has promoted the development and dissemination of science-enabling software, tools, and resources by DDD Investigators, *Practices* grantees, and the MSDSEs:

- All 14 DDD Investigators have disseminated or are developing software or other science-enabling tools data-driven research, and some indicated that they had had difficulty advancing work on such tools before their DDD award.
- Jupyter users value its usefulness in teaching, enabling research collaborations, and allowing the posting of data, analyses and findings in a reproducible format. By January 2016, Project Jupyter estimated that approximately 500,000 Jupyter Notebooks had been shared on github; the team released an alpha version of JupyterLab, an improved interactive development environment in the summer of 2016; and more than a dozen universities have implemented JupyterHub, a multi-user, browser-enabled version of Jupyter for implementation in a cloud or high performance computing environment.

- Data Carpentry has capitalized on its DDD grant to become a self-sustaining, transparently managed organization that is helping to meet the high demand among scientists from a wide range of domains for training in data organization and analysis tools for research.
- The Julia Language, despite not yet having released version 1.0, has witnessed dramatic growth in the past few years, becoming one of the top ten programming languages in active development on github by May 2017 with more than 8,500 stars and nearly 2,000 forks with a growing number of advocates for its applications to domain-specific research problems in astronomy, bioinformatics, geosciences, statistics, numerical computing and data visualization.
- All three MSDSEs have research scientists, fellows and other personnel who are actively producing and sharing a large number of tools for data-driven inquiry across multiple scientific domains, including both tools addressing particular domain-specific challenges and others that have broad applicability across scientific domains.

The DDD initiative has begun to demonstrate the importance of new academic environments that nurture data-driven research and researchers, although some challenges remain:

- The DDD initiative has played a major role in catalyzing academic institutions' provision of training opportunities to acquire data-driven skills for scientific inquiry at several scales.
- DDD Investigators have benefitted from the opportunities provided by the DDD initiative to network and build community.
- MSDSEs have demonstrated their value to host institutions via synergistic engagement with other research centers and initiatives at their universities, and via structured, project-based collaborations between domain scientists and computational methodologists focused on solving concrete problems.
- Although individual DDD Investigators, as well as MSDSE data science fellows, research scientists, postdoctoral fellows and graduate students have advanced professionally within academic research settings, the "host" institutions have not implemented formal changes in the criteria for evaluating data-driven scientists' contributions.
- MSDSE respondents believe their fellows and graduate students have had opportunities that will make them competitive on the academic job market.
- MSDSE institutions have begun experimenting with alternative career pathways for data-driven researchers, but limited data exist, to date, to assess the outcomes of these experiments, and respondents cited concern about the sustainability of these positions post-DDD.

Although scientific discovery is a long-term goal of the DDD initiative, there is clear evidence that the initiative is meeting the short-term goal—a more robust foundation of data analysis tools and methods for scientific inquiry—that is a necessary precursor to scientific discovery:

- DDD Investigators and the MSDSEs have robust publication records.

- *Practices* grantees have documented evidence of their role in scientific findings.
- Determining the role of any grant program, including the DDD initiative's role, in particular scientific findings presents challenges of causal attribution.
- Looking for links between the DDD initiative and discoveries is likely premature, given its short number of years relative to the typical time frames for peer review and publication.

Data reveal a network of links between the DDD Investigators, *Practices* grantees, and the MSDSEs:

- Eight DDD Investigators either have collaborated with researchers at an MSDSE or have an affiliation with an MSDSE and actively participate in its community at the institution.
- Data Carpentry and Project Jupyter are deeply integrated with the MSDSEs and the work of several DDD Investigators, and each MSDSE has one or more active users or contributors to the Julia Language.
- Dask was prototyped at the 2015 BIDS Data Structures for Data Scientists workshop, and data science fellows at the eScience Institute contribute to efforts to connect Dask with scikit-learn.

2.2 The DDD Initiative's Role in Highlighting the Value of Data-Driven Scientists

At mid-term, there is compelling evidence from interview data that the DDD initiative has demonstrated the value of data-driven researchers both to their own institutional colleagues and to researchers in many scientific domains. First, the DDD initiative awards **enhanced the visibility and credibility of individual researchers and data-driven research in general**; and, in a virtuous cycle, the increased visibility translated into tangible new resources and opportunities, further enhancing the profile of data-driven scientists. Second, by providing funding for grantees to hire personnel, the DDD initiative **expanded grantees' capacity to carry out their core research and development agenda**. Third, the DDD initiative **enabled risk-taking** that would not have occurred otherwise and that further advanced grantees' development **and gave grantees considerable flexibility to explore emerging opportunities**.

Terminology

When characterizing findings from the 13 DDD Investigators who participated in an interview or the survey, we use the following conventions throughout the report:

- *"Most" = 11 to 12 DDD Investigators*
- *"A majority" = 8 to 10 DDD Investigators*
- *"About half" = 6 to 7 DDD Investigators*
- *"About a third" = 4 to 5 DDD Investigators*
- *"A few" = 2 to 3 DDD Investigators*

2.2.1 Enhanced Data-Driven Scientists' Visibility and Credibility

Ten DDD Investigators indicated that the DDD Investigator award had validated their credibility as independent researchers and provided new opportunities and access to resources and colleagues:

> *It was completely because of the award that I gained this visibility within circles on campus and at [the MSDSE at my institution]. On Wednesday, I am giving a seminar in [another department] which is a direct consequence of me getting the DDD*

> *Investigator award. Had I not received this award I would have continued to try to get this work funded. … I would [have taken] the initiative to become a part of those communities but it probably wouldn't have happened as organically if I hadn't received this high-profile award. It kind of put me on the map as a data-driven researcher and has helped me establish credibility in this field, which is somewhat a new direction for my research. (DDD Investigator)*

Several administrators at DDD Investigators' institutions concurred that the awards represented validation, both for individual researcher and for data-driven science writ more broadly:

> *It put [the investigator] on the map with the university leadership … the award … from DDD represents an external vetting of [the investigator's] research project and viability of [that] research on a more national level. Because this new area is less familiar to most people, and people do not have the tools to evaluate this, so they look to see how others vet and evaluate these kinds of people. (Administrator at a DDD Investigator's institution)*

> *[The DDD award] gave [the investigator] enormous visibility and … not just at my school but at the entire university. It drew attention … and therefore by that alone it advanced and made … a little bit more coherent the data science agenda at [the university]. … It put before everyone's mind that there were opportunities like this and [that] we should start to integrate more. (Administrator at a DDD Investigator's institution)*

Particularly for newly formed departments or research centers, a DDD Investigator award to one of their faculty served as an important external endorsement. As one department chair said, "These awards put a young department on the map [and] are a validation of our strategy." This recognition, in turn, helps with faculty recruitment, as another department chair explained:

> *We have a very small number of well-known people and [the DDD Investigator] has become part of that next wave of those people. [The DDD Investigator] plays an important role in faculty recruiting ... and [because] the reputation of the department is enhanced ... that increases our ability to recruit faculty.*

Four DDD Investigators gave other examples of increased departmental or institutional support, including improved laboratory space, leverage with their administration to negotiate for the hiring of software engineers, invitations to speak both at their own and at other universities, and leadership roles or appointments to strategic committees for campus data science initiatives convened by the university. One administrator indicated that the university had renovated an entire building with upgraded IT support for an incoming DDD Investigator. An Investigator (at a different university) described the importance of the institution providing good laboratory space for fostering community:

> *Having the Moore award was essential in convincing the department to renovate and provide a better space for my students and my group because honestly money talks. … We have gone [from] having what I would consider a substandard space for my group to an absolutely fantastic space. … We embedded huge amounts of whiteboard space and a seating area for collaboration and video conferencing and space for visitors. And so all of those things we did not have before [the DDD Investigator*

> *award], and that has been really beneficial. I have noticed a huge improvement in how happy my group members are to be spending time here. (DDD Investigator)*

At the MSDSE institutions, too, the DDD initiative has shifted both the attitudes and behaviors of faculty. For example, a graduate student's successful experience collaborating with her university's MSDSE transformed her faculty advisor from a skeptical outsider to a strong proponent:

> *This student was struggling with a dataset and had a question that was … amenable to data science tools and techniques. ... At the end of it, the tool that [this student] built was taken up by everybody else in the lab. The [faculty advisor] … has been a major advocate … ever since. That lab's processes and how they handled their data was completely transformed. (MSDSE leader)*

Respondents at all three MSDSEs also perceived the MSDSEs' joint appointments of faculty and postdoctoral fellows as an important mechanism for "evangelizing" the benefits of data-driven research and building cross-departmental bridges. Under these joint appointments, faculty and postdoctoral researchers receive partial support from the MSDSE and partial support from an academic department or other unit on campus,[38] and split their time between participation in core MSDSE activities and fulfilling expectations of their department (such pursuing an independent research program).

Joint faculty hires have fostered good will and cooperation from departments that otherwise might not get a new faculty line (or half-line). The result, according to one MSDSE respondent, is that "you can't get rid of that focus on interdisciplinarity … the Moore Sloan program baked that in with the joint hires." At UW, for example, using funding from the provost, the eScience Institute has acquired seven half-faculty lines. One respondent from UW indicated that this arrangement has given departments across campus an incentive to hire faculty who combine methodological and domain-specific expertise, and it has enabled it to find "friends for life" among department chairs. Not only do the departments gain an additional faculty member, argued one MSDSE respondent, but they also have a chance to witness the important contributions that these new data-driven scientists can make.

Likewise, the MSDSE at NYU successfully established a new common protocol for joint hires between the Center for Data Science (CDS) and other academic units. This protocol specifies that search committees have equal CDS and departmental representation; candidates deliver their job talks to CDS and departmental representatives; CDS and the department agree in advance on funding arrangements for a new hire's start-up package; and CDS and the department have equal voting and veto power. Several joint appointments of faculty with computer science and engineering, mathematics, and politics have resulted (including one joint appointment hired with tenure). One NYU administrator saw this model as an impressive and "clear, replicable path" for the university.

All three MSDSEs have similar joint hiring arrangements for postdoctoral fellows. At UW, the eScience Institute normalized postdoctoral salaries and benefits across several departments, and established agreements that postdocs would receive dual mentoring from departmental faculty and methodologists at the Institute. According to its 2014 annual report, this arrangement constitutes "an

[38] Funding for participation in the MSDSE may come directly from the Moore-Sloan award or from other grants that also support the MSDSE (e.g., UW's Washington Research Foundation grant).

existence proof that cross-departmental postdoc programs in data science can be successfully established." At NYU, one MSDSE respondent noted that the joint hiring of postdocs had played a key role in attracting faculty engagement with the data science environment; NYU is also experimenting with joint appointments of data science fellows and hired their first joint research fellow in 2016 in partnership with NYU's GovLab.

One MSDSE faculty member noted that the MSDSE's successful experiences with joint hires increased his faculty colleagues' optimism about the potential contributions of a data-driven scientist.

> *It didn't seem that plausible that we were going to get that [joint] hire … except that by way of this joint hiring thing, we were able to argue that there is this very exciting campus initiative … and that enabled us [to discuss] what their contribution might be to the department. I hadn't heard that discussion be taken seriously by my general colleagues [before the MSDSE]. Just the idea that we would … have a discussion and [they would] say, "Well, this person seems really exciting because not only are they doing really interesting theoretical work, but they also have been developing some really important tools that are going to enable faster discovery and more computational, reproducible work."*

2.2.2 Expanded Grantees' Capacity by Supporting New Personnel

DDD funding gave grantees across the three strategies the ability to hire personnel who expanded these grantees' capacity for core research or development activities—either by contributing expertise in software development or computational methods, or by freeing grantee leaders from operational or administrative tasks.

The DDD Investigator award allowed its recipients build or expand their research groups. About half (6 of 13) of responding DDD Investigators described how their DDD award had given them flexibility to hire particular types of personnel. One of these successfully recruited her top two picks for postdoctoral fellows, both of whom had non-traditional interests in applying research tools and insights from one domain to others. Four DDD Investigators hired a software developer or engineer; one of these four Investigators hired four software developers. Two of these investigators indicated that their ability to hire a software developer was a unique benefit of the DDD award and something that would not have been possible with another funder.

> *We have much more software engineer expertise now. [A federal agency] is very willing to fund research to an extent; but to fund delivery and development of methods, many of the grant panels are less enthusiastic about it. (DDD Investigator)*

Other investigators indicated that by allowing them to hire individuals who could function more independently than a graduate student, the grant funding freed up more of their own time for research:

> *One of the things I've done is hire a computation scientist who can help my students gain these skills. I have good access to students who have all but one of those base skills, or ... two of those three skills.... We try to identify smart people who are willing to learn, and teach them those skills. Funding has allowed me to employ someone to help me in that endeavor, specifically someone who knows more computation than I do and has more time than I do to help students with their computational issues.*

One DDD Investigator not only saved significant time by hiring a full-time software engineer to tackle work that he had previously done, but also discovered that someone with more expertise could enhance his research group's productivity by improving their software:

> *Really good [software] development means … general improvement and testing of code to make sure you are doing what you want. [It takes a] huge amount of time and effort. If you look at software developed by the average science lab, [it] doesn't come with those things, since it's just one grad student or postdoc writing those things. [Our software engineer] has spent a huge amount of time improving underlying infrastructure … adding new functionality and developing new software.*

Similarly, the DDD award provided a few investigators the flexibility to hire staff that did not fit into typical salary/rank categories. One DDD investigator, for example, successfully hired a methodologist who had a competing offer from industry by persuading her institution to give the candidate status as a research scientist, with a corresponding salary that was more competitive than that of a postdoctoral fellow.

For *Practices* grantees in particular, DDD funding has been transformative; people who had been working largely in the evenings or on weekends (on top of existing responsibilities) now have direct support to devote their time to this work.

> *This is a project ... that cannot be done purely as a volunteer. ... Not only is it not viable, it's also extraordinarily unfair … to people who [cannot] afford, because of where they are in life or the privilege they have financially, to contribute to github on nights and weekends. … From those perspectives, these funds [from the Moore Foundation] are critical. (*Practices *grantee)*

Another *Practices* project lead extolled the fact that the DDD program allowed funding of full-time operational support that freed up time for working on the substantive aspects of the project:

> *What's really unique about the Moore grant is that it really was seed money. We write a lot of grants to [federal funding agencies], but every one of those grants is for specific projects. Every one of those grants is awesome, but it gives us more work to do. None of it really funds ... an executive director, a communications person. ... So what's different about the Moore funding is the ... support for the organization rather than a specific programmatic outcome. [The Moore grant] is very unique in this space. (*Practices *grantee)*

The DDD funding similarly freed the team developing another *Practices*-funded project from constraints imposed by commercial clients generally more interested in solving specific problems than in investing in a free and open source tool developed for larger benefit. What this grantee cited as the "single most important thing" for his user community was the DDD team's emphasis on the importance of developing and releasing version 1.0: a robust and stable tool on which users could rely.

At the MSDSEs, the data science fellows, research scientists, and program management staff supported with DDD funding (and with other funding that the Moore-Sloan funds allowed these institutions to leverage) have been critical for a broad range of research and educational activities.

One MSDSE leader referred to the fellows as one of the data science environment's key "anchors," a view echoed in similar comments from respondents at the other MSDSEs. At all three MSDSEs these individuals have led Working Groups, contributed to "incubator projects" (MSDSE-sponsored project-based collaborations between methodologists and domain scientists, described below), offered trainings, and seeded collaborations with faculty across each of their institutions. At NYU, the MSDSE award enabled the hiring of talented researchers whose presence transformed its Center for Data Science (CDS) from primarily a master's-degree granting program into a robust research center; the eScience Institute at UW has successfully recruited research staff and software engineers from industry, and its data science fellows have been active participants in the Data Science Incubator and Data Science for Social Good programs, yielding data-driven projects that have won scholarly awards, attracted local press coverage, and received external grant funding. The data science and computational fellows at BIDS have likewise played key roles in a variety of collaborative activities, including a 2015 two-day Data Structures for Data Science workshop, the compilation and publication of a volume of case studies in reproducible science, and the formation of the Image Processing Across Domains (ImageXD) and Text Analysis Across Domains (TextXD) research collaborations.

2.2.3 Enabled Risk-Taking and Provided Flexibility

Importantly, half (7 of 13) of the DDD Investigators credited the award with providing freedom to pursue potentially risky research agendas. The following quotes illustrate investigators'—and administrators'—explicit awareness of the flexibility they have to take risks, to pursue research that pushes boundaries, and to consider their research agendas more broadly than might be expected from other grants/funders.

> *As a new investigator, having a really generous-sized grant [from the DDD initiative] has freed me up to do what I think are the more exciting, risky items in my portfolio early. ... The Moore grant freed me up [and] I'm not under as much pressure to do those [more traditional] things [such as apply for a federal agency grant] immediately, and we can start to take on bigger projects. ... I'm working on a paper now ... and it's a statement paper about my lab and our approach. It's a proof of concept and gives an exciting new view of [biological] systems, and I have the time to collect and analyze that data and put it into a paper, instead of smaller papers. It's more like a manifesto. (DDD Investigator)*

Although the risks associated with the choice of research questions to pursue are particularly high for junior faculty, even for more senior faculty, pursuing challenging questions or innovative methods, even when these activities might produce novel discoveries, can be risky.

> *We have focused on this unusual research that we normally couldn't get funding for, and it's very discovery oriented but it's harder to publish. I have a hilarious quote from [a journal where] we tried to publish: "Sorry for the prolonged review process but this is an interdisciplinary manuscript and that presents challenges." So [publishing] has been a much more drawn out process than I'm used to. (DDD Investigator)*

With the DDD award, Investigators also noticed an increased tolerance for risk among members of their research group:

> *I would say that people within the lab have taken their cue from the grant and now that we have secure support, we can try crazy things. I had two grad students…who are much less concerned about traditional boundaries than they were before I received award. (DDD Investigator)*

Three DDD Investigators valued the DDD team's openness to unanticipated changes in the research, compared with the more rigid constraints of other funders.

> *Other awards were tied more closely to the original proposal … [but] the Moore Foundation Investigator awards … are funding our research program, not a specific project, and that has several implications … in terms of freedom to shift the focus of your research. (DDD Investigator)*

DDD funding also allowed DDD Investigators the freedom to focus more on the quality than the quantity of publications, and to work directly on developing methods, software, and tools other types of funding would not support:

> *I don't feel like I have to publish crappy work to succeed ... which leads to more interesting things to not have to go with first publishable result with small p value. My work is less dependent on number of publications. If you are a data scientist and need to maximize [productivity], there are a lot of poor scientific ways to do it. This [DDD] program emphasized the process of science, more than the need to publish. That has really made a difference to us as grantees. (DDD Investigator)*

> *Having the [DDD] award opens up a little more flexibility in what we do. We wrote a paper and came up with an approach for a new method ... and we ended up being able to do work in a robust and rigorous way. I am working on these projects that [a federal agency] will never fund and that's okay, I'm fine with that, because I have DDD funding. (DDD Investigator)*

One *Practices* grantee leader, who saw overlap between work on the project that was explicitly funded by the DDD grant and work funded by other sources, also cited the value in the DDD team's flexibility in allowing them to collaborate across these grant funding boundaries:

> *Some of the new stuff that has happened [on this project]—we are partially supported [but it is] not an official milestone [of the DDD grant]. I hope ... the evaluation [of the DDD initiative] recognizes the value of fluid usage of resources that crosses these boundaries. Because the fact that we're allowed to ...collaborate openly [with others who were not initially proposed as collaborators] without worrying [is] what matters; that they can work together, [that] they have the freedom to solve problems. That is [what is] enabling these things to happen.*

MSDSE leaders echoed these types of comments, and valued the somewhat open-ended nature of their awards for allowing them to experiment with different approaches. For example:

> *The Moore-Sloan funding gave the university prestige and the resources needed to make the experiment worthwhile … and to let them take some risks. (MSDSE leader)*

2.3 The DDD Initiative's Role in Promoting the Development and Dissemination of Science-Enabling Tools, Methods, and Resources

Initial indicators at mid-term show that the DDD initiative has played a key role in the development of a wide range of tools and practices applicable both within specific scientific domains and across domains. **DDD Investigators, *Practices* strategy grantees, and MSDSE personnel have made significant contributions to this infrastructure**.

2.3.1 DDD Investigators' Development of Science-Enabling Software, Tools, and Resources

All 14 DDD Investigators have disseminated or reported that they are developing tools for a variety of tasks. These tools include domain-specific tools (e.g., PySCA, Toboggan, Salmon, khmer, ASPIRE, CompCellScope), as well as code to produce analyses in publications or preprints. Others have developed or contributed to tools intended for a wider audience; these include workflow tools to support reproducibility (continuous analysis; Jupyter); interactive data visualization tools (e.g., Vega, Vega-Lite, Voyager), and data extraction tools (e.g., DeepDive, for extracting information from text; rData Retriever, to find, download, and merge publicly available data). Several of these DDD Investigators also contribute to training initiatives for data-intensive sciences, including Data Carpentry, Software Carpentry, Python workshops, and corollary projects such as Lab Carpentry (tools for starting and managing a scientific research laboratory).

Part of the work that investigators talked about was making sure that the tools they develop are accessible to different scientists in their respective fields, some of whom may not have programming experience.

> *One thing that has been spurred by [my] DDD award is taking that Python package and figuring out ways to make it accessible to real experimental[ists] who don't have Python knowledge. The [new] package is a first step. It's free and it's downloadable and has instructions but there's a learning curve if you invest time in it. We're trying to make a web app for the thing so if you don't do any programming, you can still go and make predictions with this tool and experimentally test them in the lab so the group of people who could use it will be larger. The vast majority of (scientists) don't have that programming background. (DDD Investigator)*

Some DDD Investigators indicated that moving these tools forward pre-DDD award was difficult.

> *Other grants I have received would not have funded this kind of tools development...they're all narrower in the actual area of what they would be willing to fund...whereas [with] the DDD grant...it just has to be useful in the big picture. (DDD Investigator)*

One investigator had wanted to develop a tool that could help to solve the problem of merging and cleaning datasets yet lacked enough time to develop the tool on his own. Before the DDD award, he was unable to do more than just fix bugs and errors in the software; the DDD award provided the resources to hire a software developer who is working to make the tool compatible with a range of different languages and ensure that the underlying infrastructure is robust.

2.3.2 *Practices* Grantees' Development of Science-Enabling Software, Tools, and Resources

The four *Practices* grantees included in this evaluation have demonstrably contributed to a more robust infrastructure for data-driven science. In particular, **Project Jupyter** and **Data Carpentry,** which received DDD funding in July and September 2015, respectively, have achieved strong adoption, name recognition, and endorsements; Data Carpentry's Executive Director now sees the organization as largely self-sustaining. Julia Computing, funded by the DDD initiative in October 2015, has made significant progress in its development and dissemination of **the Julia language**, although data suggest reluctance among data-driven scientists to adopt it before its anticipated release of version 1.0. **Dask** and **Numba**, two Python packages under development by Continuum Analytics and funded by DDD beginning in July 2016, appear to have achieved less widespread adoption to date, although we had limited data with which to assess their mid-term progress.

Project Jupyter

Project Jupyter emerged from Fernando Perez's early work on iPython beginning in 2001, with subsequent development in partnership with Brian Granger.[39] Both direct funding and in-kind support (server space) for this early work came from Rackspace, Microsoft, Google, and the Department of Defense, among other sponsors.[40] Project Jupyter officially launched in July 2014 with the recognition that iPython-developed tools supported several languages other than Python.[41] In 2015, the DDD initiative made a three-year, $1.5 million investment in Project Jupyter; the project leads leveraged the DDD grant for an additional $1.5 million from the Sloan Foundation and the Helmsley Charitable Trust; the project also received technology industry funding and other support.[42]

Scientists across multiple domains have embraced Jupyter Notebooks. One of the key goals of Project Jupyter's July 2015 DDD grant was to improve the interface and user experience of Jupyter Notebook. **By January 2016, Jupyter estimated there were approximately three million users and 500,000 Jupyter Notebooks on github**. The success of Jupyter Notebooks is also indicated by notable examples of their use:[43]

---

39 Pérez, F. & Granger, B.E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering, 9*, 21-29.

40 Other funders included Enthought, Fastly, the Ohio Supercomputer Center and the Department of Defense's High Performance Computing Modernization Program; for a complete list, see: http://ipython.org/ipython-doc/rel-0.12.1/about/credits.html

41 For more on the history of Jupyter, see: https://blog.jupyter.org/2015/04/15/the-big-split/ and the SciPy 2014 announcement: https://www.youtube.com/watch?t=258&v=JDrhn0-r9Eg as well as http://blog.fperez.org/2012/01/ipython-notebook-historical.html

42 Jupyter had received a smaller grant from the Sloan Foundation prior to the DDD grant. Other funding for Project Jupyter comes from Rackspace, Microsoft, Google, Fastly, and OpenDreamKit, a Horizon 2020 grant from the European Research Infrastructures Work Program. Its Steering Council includes employees from Bloomberg, Continuum Analytics , Netflix, QuantStack Scientific Computing, and the both UC-Berkeley and California Polytechnic State University in San Luis Obispo.

43 Also see: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

- Data from the 2016 detection of gravitational waves are available in Jupyter Notebook form from LIGO's Open Science Center.[44]
- The entire contents of the bestselling *Python Data Science Handbook* are implemented in free Jupyter Notebooks in a github repository.[45]

Based on a January 2016 survey, 25 percent of Jupyter users self-identified as data scientists, and 18 percent each as scientists and researchers.[46] The majority were daily or weekly Jupyter users who created Jupyter Notebooks for tens or hundreds of users; some shared their notebooks with thousands or more. Survey respondents noted that the then-current version (again, in January 2016) lacked features such as version control, text/code editing, debugging tools, and easier options for exporting to slides or reports.

Having identified several of these shortcomings themselves, **the Jupyter team was actively engaged in developing JupyterLab, and expects to release version 1.0 (a key deliverable of its DDD grant) in summer or fall 2017,** after which it will intensify focus on enabling real-time collaboration.[47,48] The team announced an alpha version of JupyterLab in summer 2016. JupyterLab is an interactive development environment, designed to give Jupyter Notebook users a more modular set of "building blocks" (such as a file browser, text editor, widgets, and output) that they can mix and match within a more powerful and flexible workspace. In addition to project staff, Jupyter had several industry partnerships (with IBM, Bloomberg, Microsoft, and Continuum Analytics) and 35 contributors on github.

More than a dozen academic research universities have deployed JupyterHub, another important contribution to the suite of tools for data-driven science (e.g., University of Colorado-Boulder, University of Illinois, Massachusetts Institute of Technology, Michigan State, University of Rochester, UC-San Diego, Texas A&M).[49] JupyterHub is a multi-user, browser-enabled version of Jupyter designed for implementation in a cloud computing environment or with high-performance computing resources. Other notable examples of its use include these:

- UC Berkeley's Data Science Initiative (for use across data science courses);

---

44 LIGO stands for Laser Interferometer Gravitational-Wave Observatory. For event data in Jupyter Notebook form, see: https://losc.ligo.org/about/

45 VanderPlas, J.T. (2016). *Python data science handbook: Essential Tools for working with data*. Sebastopol, CA: O'Reilly Media. The repository is available at: https://github.com/jakevdp/PythonDataScienceHandbook

46 Survey data are available at http://github.com/jupyter/design/tree/master/surveys/2015-notebook-ux. More than 1,700 responses from a convenience sample (i.e., participation was invited via posts on Jupyter's blog, Twitter accounts, and Google group) were submitted.

47 Granger, B. E. (2016, July 14). JupyterLab: Building blocks for interactive computing [Slides]. Retrieved from archive.ipython.org/media/SciPy2016JupyterLab.pdf

48 Tache, N. (2017, May 25). *JupyterLab: The evolution of the Jupyter web interface* [Blog post]. Retrieved from https://www.oreilly.com/ideas/jupyterlab-the-evolution-of-the-jupyter-web-interface

49 http://jupyterhub.readthedocs.io/en/latest/gallery-jhub-deployments.html

- Implementation in the high performance computing platform via a partnership between the Berkeley Research Computing program, Pacific Research Platform project, and BIDS for experimentation by approved researchers;[50] and
- Implementation at the National Energy Research Scientific Computing Center's (NERSC) for data-intensive computing on a Cray XC40 Cori supercomputer.[51]

The NERSC's implementation of JupyterHub is facilitating scientific advances. Researchers in the Large Synoptic Survey Telescope's (LSST) Dark Energy Science Collaboration are using NERSC's JupyterHub to develop "Twinkles," an open source research project to test and validate software that will be used to make precise measurements of supernovae and strong gravitational lens time delays.[52]

Interview respondents cited several benefits of Jupyter, including its usefulness in teaching, its role in enabling of research collaborations, its interactivity for exploring data, and its potential for allowing researchers to disseminate reproducible findings. First, Jupyter appears well suited for teaching data science concepts to undergraduates, even for relatively novice-level programmers.[53] One DDD Investigator found the Jupyter Notebook environment ideal for student assignments: "I'm teaching a data science capstone class [for undergraduates] … and I have been having them do all of their assignments in Jupyter." A graduate teaching assistant in the geosciences reported that Jupyter also allows students to experiment interactively with complex numerical simulations, providing an instructional tool that she and her professor would not have been able to develop on their own.

Second, Jupyter enables collaboration among members of a research team. Two DDD Investigators, one postdoctoral researcher, and three graduate students all cited the ease of sharing Jupyter Notebooks as one of its key benefits. From a DDD Investigator's perspective:

> *We use them a lot for lab notebooks for experiments and data analysis and [lab members] can put in comments, parameters, goals, all of [the] data analysis, interpretation of the results. My students can give me a summary of what they've been working on and it's contained in a single document and I can easily look at their calculations. I can see every step they did and change things to see how it changes the output. So the ability to interact with and communicate the data is really powerful with all the interpretation and notes. And it's free, which is huge.*

Finally, Jupyter is an important tool for supporting reproducible research. As early as 2012, Titus Brown (now a DDD Investigator) experimented with posting his iPython Notebook (since renamed

---

50 http://research-it.berkeley.edu/blog/17/01/24/free-fully-loaded-jupyterhub-server-supports-campus-research-computation

51 http://www.nersc.gov/news-publications/nersc-news/nersc-center-news/2016/jupyter-notebooks-will-open-up-new-possibilities-on-nerscs-cori-supercomputer/

52 https://community.lsst.org/t/analyzing-lsst-desc-twinkles-data-at-nersc-via-remote-jupyter-notebooks/1533

53 High school teachers have also used Jupyter Notebooks for introductory computer programming and scientific methods courses. See https://peak5390.wordpress.com/2013/09/22/how-ipython-notebook-and-github-have-changed-the-way-i-teach-python/ and http://srm12.weebly.com/current-handouts.html accessed May 17, 2017

Jupyter Notebook) alongside research findings.[54] He has recently advocated for "mybinder.org," a site that runs Jupyter Notebooks on github for free as "a solution to publishing reproducible computational work."[55]

Of the 13 DDD Investigators interviewed for the evaluation, nine reported that their research group used Jupyter routinely, and nearly three-quarters of the postdocs, graduate students, and research scientists in five of the DDD Investigators' research groups reported using Jupyter either routinely or occasionally. Several interview respondents (including a DDD Investigator, a software developer, a graduate student, and a postdoctoral researcher) described Jupyter as a valuable, even "indispensable" tool, for which there were no good alternatives.

Data Carpentry

Officially launched in July 2014, Data Carpentry emerged from a 2013 collaborative workshop between IT members of several National Science Foundation (NSF) supported centers for research in biology,[56] and NSF-supported investigators developed much of the initial curricula.[57,58] This vision culminated in a prototype workshop in May 2014 at the NSF-funded National Evolutionary Synthesis Center (NESCent).[59] Data Carpentry's co-founders include a DDD Investigator (White) and a MSDSE member (Karthik Ram at BIDS); other MSDSE members (e.g., Ben Marwick at UW's eScience Institute) have been key early contributors.[60]

Since receiving its DDD award in September 2015, Data Carpentry has become a self-sustaining organization with an Executive Director, Deputy Director of Assessment, Community Development Lead,[61] and Program Coordinator. A key mechanism for making Data Carpentry self-sustainable has been the tiered partnership model for institutions. These partnerships (offered in conjunction with Software Carpentry) not only generate revenue for Data Carpentry but also help

---

54 Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. Retrieved from https://arxiv.org/abs/1203.4802

55 http://ivory.idyll.org/blog/tag/reproducibility.html, accessed May 17, 2017

56 Centers included: the National Evolutionary Synthesis Center (NESCent); the BEACON Center for the Study of Evolution in Action; Integrated, Digitized Biocollections (iDigBio); the National Ecological Observatory Network (NEON); the iPlant Collaborative (iPlant); the National Socio-Environmental Synthesis Center (SESYNC); Data Observation Network for Earth (DataONE) and the National Institute for Mathematical and Biological Synthesis (NIMBios). See: https://www.idigbio.org/content/data-carpentry-please-can-we-have-some-more

57 NSF support included a 2010 CAREER award to (subsequent) DDD Investigator Ethan White; SESYNC support for Mike Smorul; NESCent support for Hilmar Lapp, Karen Cranston; iDigBio support for Deborah Paul, Matt Collins, Kevin Love, and Francois Michonneau, and both BEACON and NSF 12-101 support to Tracy Teal.

58 Teal, T. (2014). Moore Foundation contribution to Data Carpentry. [Blog]. Available at http://www.datacarpentry.org/blog/announce/

59 https://software-carpentry.org/blog/2017/02/uf-program.html

60 See Tracy Teal's July 9, 2014 SciPy talk at https://www.youtube.com/watch?v=SMyto7WhiNs

61 Duckles, J., & Teal, T. (2017, June 7). Announcing Belinda Weaver as our community development lead [Blog post]. Retrieved from http://www.datacarpentry.org/blog/community-developement-lead/

build local capacity for training in data management skills for research. Depending on their membership level, partners receive two to six workshops coordinated by Data Carpentry (discounted rates for additional workshops), an online or in-person instructor training for 6 to 15 people, and the option to host their own self-managed workshops locally.[62] Partners include the MSDSEs BIDS and eScience Institute, Michigan State's Institute for Cyber-Enabled Research, the University of Florida, University of Michigan, UC-Davis, and UC-San Diego, and University of Miami, as well as several European universities and research institutes.

Data Carpentry has trained more than 800 volunteer instructors worldwide, with a long waiting list for future instructor trainings, helping to build much-needed capacity in the data-driven sciences. Data Carpentry's instructor training program is a two-day course offered in partnership with sister organization Software Carpentry. The programs provide instructors with training in educational psychology, instructional pedagogy, and practical issues of leading a Data Carpentry workshop. New instructors who have completed the training also receive support through a mentorship program, in which they can meet online with an experienced instructor on a regular or ad hoc basis.

The need for training opportunities like those provided by Data Carpentry is especially great for the next generation of data-driven scientists, and even for experienced researchers. In a 2016 survey of principal investigators funded through NSF's Biological Sciences directorate, more than 70 percent of respondents identified training in data integration, management, and scaling for cloud/high-performance computing as their greatest unmet need.[63] As early as 2015, NSF listed Data Carpentry as a resource, after the agency had required all full proposals to include a data management plan.[64,65]

DDD Investigators, their postdoctoral and graduate student colleagues, non-awardees from the DDD Investigator competition, MSDSE respondents, and experienced Data Carpentry instructors **all affirmed both the need for the kinds of hands-on training opportunities provided by Data Carpentry and its impact**.

> *A lot of researchers don't have training in [data management]. The workshop format of being hands-on and working together is really good. It gives [people] another chance to connect with each other outside of research bubbles. … I definitely recommend it to researchers we work with, people who want to get started … and don't quite know where to begin. (Data Carpentry instructor)*

---

62 Self-managed workshops can be branded as a "Data Carpentry" workshop provided that the host institution registers the training with Data Carpentry, the content is Data Carpentry, and at least one instructor is Data Carpentry certified.

63 Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *bioRxiv 108555*. Preprint. https://doi.org/10.1101/108555

64 National Science Foundation. (n.d.). *Advisory Committee for Geosciences. October 21-22, 2015. Meeting minutes* [Word document]. Retrieved from http://www.nsf.gov/geo/adgeo/advcomm/acgeo-oct2015-minutes.docx

65 National Science Foundation. (n.d.). *Directorate for Biological Sciences. Updated information about the data management plan required for full proposals. October 1, 2015.* Retrieved from http://www.nsf.gov/bio/pubs/BIODMP_Guidance.pdf

Interview data suggest that graduate students constitute a large proportion of Data Carpentry workshop attendees. Nine interviewees (DDD Investigators, postdocs and graduate students, and non-awardees from the Investigator award competition) noted the importance of Data Carpentry trainings for graduate students. One DDD Investigator pointed graduate students to Data Carpentry's online materials because "the examples are more attuned to people doing research and academic science than a generic Python tutorial." Several postdocs and graduate students working with DDD Investigators had completed its instructor training program and volunteered to lead Data Carpentry workshops.

Data Carpentry's own October 2016 assessment report indicates that more than 36 percent of their workshop attendees are graduate students, the largest single constituent group.[66] These assessment data, based on pre- and post-workshop surveys, showed that:

- Some 74 percent of learners agreed or strongly agreed that they could immediately apply what they learned at a workshop (n=421).
- Of learners who rated their pre-workshop skills as "very low," "low," or "neither low nor high," 95 to 97 percent reported "somewhat higher" or "higher" levels of skill post-workshop.
- Of those who rated their pre-workshop skills as "high" or "very high," 78 to 87 percent reported "somewhat higher" or "higher" levels of skill post-workshop.

Post-workshop, learners also endorsed statements about the importance of data organization for reproducible research; the value of scripting languages (R, Python) for making analysis more efficient and reproducible; and the value of these languages for preventing accidental changes to data.[67]

Data Carpentry is poised for continued growth. With its DDD funding, Data Carpentry has invested considerable effort into building a transparent, accountable organization responsive to its community of trainees and instructors. A full-time Community Development Lead coordinates multiple communication channels and outreach initiatives, including social media and online discussion boards a monthly newsletter; and virtual events such as a "bug barbecue" to fix typos and identify missing or confusing content in lessons. Data Carpentry has posted curricula under CC-BY licenses online for ecologists, genomicists, biologists, and scientists working with geospatial data; and it has actively engaged scientists in helping to expand its workshops to new scientific domains. The instructor trainings for librarians in May and June 2017 were fully booked, and scientists have increasingly begun to acknowledge Data Carpentry in publications about resources for others seeking data-intensive resources and tools.[68,69]

---

66 Jordan, K. L. (2016). *Data Carpentry assessment report: Analysis of post-workshop survey results.* Retrieved from https://zenodo.org/record/165858#.WUfF_OvyttR

67 Jordan (2016).

68 See: Tippmann, S. (2014, September 11). My digital toolbox: Ecologist Christie Bahlai talks data hygiene [Blog]. https://doi.org/10.1038/nature.2014.15896

69 See: Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., … Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution, 1,* 0160. https://doi.org/10.1038/s41559-017-0160

The Julia Language

In October 2015, Julia Computing received its first installment of DDD grant funding. Initial work on the Julia language began in 2009, culminating in a 2012 public announcement by its four developers.[70] In 2015, the developers formed Julia Computing, allowing them to raise revenue by providing consulting and training to commercial clients, while retaining Julia as an open source language.[71] In May 2015, the developers reported over 5,000 github stars and almost 600 registered packages.[72]

Although data on usage of the Julia language from 2015 to 2016 were unavailable, **Julia saw dramatic growth in its user base from 2016 to 2017** (Exhibit 2.1) and recently the founders published a paper describing Julia for numerical computing (in preprint since 2014) in the prestigious *SIAM Review*.[73]

Exhibit 2.1: Growth in Julia Usage, 2016–2017

Metric	Increase
Downloads (based on Amazon's S3STAT logs)	161%
Number of published citations of two foundational papers on Julia[74,75]	121%
Number of questions posted to Stack Overflow	79%
github stars (across packages)	74%
Number of registered packages on pkg.julialang.org	72%
Number of JuliaBox users (a browser version of Julia currently in beta)[76]	71%

Source: Julia Computing (2017, February 3). Newsletter 2017. [Blog]. Retrieved from http://juliacomputing.com/blog/

As of May 2017, Julia was one of the top ten programming languages in active development on github with more than 8,500 stars and nearly 2,000 forks.[77] Evidence from its first two user conferences (JuliaCon 2014 and 2015) also illustrates growth in the range of scientific applications of Julia: at JuliaCon 2014, approximately 20 talks over two days included presentations on its use in natural language processing, image analysis, and statistics; a year later, JuliaCon 2015 featured more

---

70 Bezanson, J., Karpinski, S., Sha, V. & Edelman, A. (2012). Why we created Julia. [Blog]. https://julialang.org/blog/2012/02/why-we-created-julia

71 Novet, J. (2015, May 18). Why the creators of the Julia programming language just launched a startup. *VentureBeat*. Retrieved from https://venturebeat.com

72 Github stars function both as a bookmark for easy access and a sign of user appreciation.

73 Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59* (1), 65–98. https://doi.org/10.1137/141000671

74 Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). *Julia: A fast dynamic language for technical computing.* Retrieved from https://arxiv.org/abs/1209.5145

75 Bezanson et al. (2017).

76 Julia Computing (2017, February 3). Newsletter 2017. [Blog]. Retrieved from http://juliacomputing.com/blog/

77 Github stars allow users of a repository easy access and show appreciation. Forks allow programmers to add features or make contributions to a github project. See Claster, A. (2017, May 25). Julia ranks among top 10 programming languages developed on github. [Blog]. Retrieved from https://juliacomputing.com/blog/

than 60 talks with sessions on applications of Julia in astronomy (JuliaAstro), bioinformatics (BioJulia), econometrics, geosciences, numerical computing, statistics, and data visualization. By 2016, JuliaCon needed five days to accommodate its talks, workshops, poster sessions, and a hackathon; in 2017, JuliaCon featured applications of Julia in additional disciplines (such as ecology, evolution, systems biology, mathematics, machine learning, neuroscience, and quantum physics).

Interview data suggest that those who had adopted Julia valued its combination of a high-level syntax with speed,[78] but also recognized that because Julia has not yet reached version 1.0, "it's not ready for prime time" (according to one DDD Investigator) and is "still growing in adoption" (according to one *Practices* project lead). Nonetheless, researchers in neuroscience and biostatistics cited Julia as particularly well suited for big data:

> *There's R, and it's a great language, it's open source, academics are contributing a lot to it, there's a lot available in it but it's not necessarily scaled for big data, and when you start programming things, you easily create programs which run too slow. Julia is an environment which is almost like C programming, but it's still high level and it doesn't get slow. (Non-awardee in the DDD Investigator competition)*

> *Julia is very fast. Gone are the days where you might test an idea in one programming language and then have to implement it in a second programming language. It's a huge source of efficiency. Julia has been designed so that it operates in the end in terms of what the machine runs. And that's not true of other languages that don't have Julia's performance advantages. (Julia user)*

Despite these advantages, some interviewees explained their reluctance to use Julia. One postdoctoral researcher and one graduate student each cited as a barrier the large investment of time needed to learn any new language. Another cause for reluctance seems to be that Julia, as a relatively young language, still lacked some important functionality and was not yet entirely stable. A regular Julia user explained these concerns:

> *First of all, for a young language, sometimes there just aren't packages available and you end up having to write things that already exist in other languages. That's getting rarer all the time but it still happens. The second disadvantage is that … the core fundamentals of the language are changing over time. Every new incremental version of Julia requires you to rewrite or potentially update your code. Those version upgrades are welcome because every one of them has been what I would call a major step forward, but we have enough code that updating one version … to the next can easily be a week of effort.*

Despite these limitations, Julia Computing has made steady progress toward the release of version 1.0, the primary goal of its DDD grant:

- October 2015: Version 0.4 released;

[78] A high-level programming language typically abstracts away from machine language and allows a user to code using elements that are closer to human language. Unlike Julia, most high-level languages tend to manage CPU memory less efficiently than lower-level languages, resulting in slower computational times.

- October 2016: Version 0.5 released; and
- June 2017: Version 0.6 released.

In June 2017, Julia Computing received a $910,000 grant from the Sloan Foundation to support training, adoption, and documentation and to promote greater diversity in the Julia community.

Dask and Numba

In July 2016, Continuum Analytics received a DDD award to develop Numba and Dask sufficiently to release versions 1.0 of each package, to engage the academic research community at Python events using free and open source training materials, and to establish a formal steering committee and governance rules. Relatively few participants in the evaluation reported using either Numba or Dask. Three DDD Investigators reported they were beginning to explore it, and one reported using it routinely; one of these three had Numba for work that resulted in a best paper award.

We observed similar results for Dask. Most respondents were unfamiliar with it. A non-awardee in the DDD Investigator competition who completed a survey for the evaluation noted that "Dask has been having an increasing impact on geoscience." One postdoctoral researcher had explored Dask and believed it could have been useful for a project at the outset, but "at this point, it would take a lot of development to change over to using Dask." Notably, however, five data science fellows and affiliated faculty member at the eScience MSDSE recently co-authored a comparison of five database management systems (DBMSs)—including Dask—for large-scale scientific image management and analysis.[79] Other eScience Institute affiliates are using Dask as part of a project to execute rapid analysis of large, array-oriented datasets, a data format frequently encountered by earth scientists and others using remote sensing data.[80] At BIDS, another MSDSE affiliate member has explored features and limitations of Dask for machine and reinforcement learning in dynamic, real-time applications.[81] Finally, researchers at NYU's MSDSE working on scikit-learn have implemented a parallel processing implementation of scikit-learn in Dask (dask-sklearn).

2.3.3 MSDSEs' Development of Science-Enabling Software, Tools, and Resources

The development and dissemination of software and tools for science is also a key focus of the MSDSEs. Each MSDSE has a working group with an explicit focus on building and disseminating software tools for research, and researchers at the MSDSEs have engaged actively in multiple such efforts. They have developed both domain-specific tools to address particular research challenges, and more general-use tools with applications to multiple scientific domains. Although the development of some tools has occurred primarily within one MSDSE, researchers from across the three MSDSEs have jointly contributed to other tools—which is not surprising given the MSDSE

---

79 The four other DBMS were SciDB, Spark, Myria, and TensorFlow. See: Mehta, P., Dorkenwald, S., Zhao, D., Kaftan, T., Cheung, A., Balazinska, … AlSayyad, Y. (2016). Comparative evaluation of big-data systems on scientific image analytics workloads: Experiments and analysis. arXiv:1612.02485v1 [cs.DB] Retrieved from https://arxiv.org/pdf/1612.02485.pdf

80 http://escience.washington.edu/research-project/exploration-of-software-tools-for-geospatial-analysis/

81 Nishihara, R., Moritz, P., Wang, S., Tumanov, A., Paul, W., Schleier-Smith, … Stoica, I. (2017). Real-time machine learning: The missing pieces. arXiv:1703.03924 [cs.DC]. Retrieved from https://arxiv.org/pdf/1703.03924.pdf

emphasis on cross-institutional collaborations and the commitment to open source tool development. Some examples of the tools with applications across domains include:

- **Project Jupyter**: led by a BIDS co-PI and supported in part by BIDS postdoctoral scholars, one BIDS leader described as an "anchor" for their MSDSE.
- The **Julia** programming language: One of the language's lead developer is a research scientist at NYU's MSDSE, and senior personnel from both BIDS and eScience Institute also contribute to the core Julia language and develop specific packages (e.g., AstroJulia, BioJulia).
- **rOpenSci**: this collection of R-based tools to support access to scientific data and text, interactive data analysis and visualization, and efficient documentation and deposit of data in repositories was developed and led by BIDS data science fellow; it received grant support from the Sloan Foundation in 2013 and $2.9 million in 2015 from the Helmsley Charitable Trust.
- An **alternative to "rainbow" color maps in matplotlib**: introduced by two BIDS computational fellows at SciPy 2015, this alternative (viridis) has now become the new default color map in matplotlib, supplanting the original, data-distorting default option.
- **SciSheets**: eScience Institute personnel are developing a new spreadsheet tool for scientists that improves readability and implements formulas as python expressions (enhancing access to thousands of existing python packages).
- **VisTrails:** this is an open source scientific workflow and provenance management system, developed by researchers at NYU's MSDSE.
- **ReproZip**: this tool, also developed by researchers at NYU's MSDSE, allows scientists to package the data files, libraries, environment variables, and options associated with a project for others to explore on any machine.
- **Myria Big Data Management Service**: developed and operated by personnel at the eScience institute, Myria is a cloud-based service for big data management intended to make initial data manipulation (cleaning, filtering, joining, grouping, transforming, and extracting features) more efficient and automated.

The MSDSEs have also developed a range of tools to address more specific research challenges in machine learning, image processing, probabilistic modeling, with applications in astronomy, particle physics, neuroscience, and bioinformatics. Some examples include:

- **mst_clustering**: this software implements an estimator for minimum spanning tree clustering in python similar to scikit-learn, and was developed by a senior data scientist at the eScience Institute.
- **pomegranate**: this Python package, developed by an IGERT big data graduate student fellow and eScience Institute participant, implements a set of fast probabilistic and graphical models ranging from probability distributions, Bayesian networks, Markov chains, hidden Markov models, and general mixed models.

- **muscle_saxs**: this software for automated analysis of X-ray diffraction images from intact muscle cells borrows methods from astronomy for high speed x-ray imaging of proteins in contracting cells, developed by a WRF/MSDSE postdoctoral fellow at the eScience Institute with faculty mentors.
- **Genotet**: an interactive visual exploration tool for validating gene regulatory networks, developed via a collaboration between researchers at NYU's MSDSE and a computational biologist.
- **scikit-learn**: development of this Python package for machine learning—and **dask-sklearn**, for parallel processing implementations of machine learning methods in scikit—was led by members of NYU's MSDSE, with contributors from other MSDSEs.
- **scikit-image**: this compendium of algorithms for image processing was developed as part of a joint Image Across Domains (ImageXD) collaboration between BIDS and eScience researchers.
- **AstroPy**: a community-driven compilation to compile Python packages for astronomy and astrophysics, this resource has contributors from both BIDS and eScience.
- **astroML**: this Python module for machine learning and data mining for fast statistical analysis of astronomical and astrophysical datasets was developed by researchers at the eScience Institute.
- **Kira**: developed by a BIDS data science fellow, Kira is a distributed astronomy image processing toolkit for implementation on Apache Spark as a faster alternative to traditional scientific workflows for processing these very large datasets.
- **TextThresher/Annotator Content Analysis module**: developed by a BIDS fellow, this tool is designed to speed content analysis of text via crowd-sourcing and decomposition of the data and coding scheme into chunks that minimize cognitive load; the project has received funding from the Sloan Foundation to develop the prototype into an open source, github hosted Annotator Content Analysis (ACA) module.

2.4 The DDD Initiative's Role in Fostering Academic Environments That Nurture Data-Driven Research and Researchers

Here, we examine progress to date in the DDD initiative's efforts to create more nurturing academic environments for data-driven research. Three key findings, discussed below in more detail, have emerged:

- **The DDD initiative has played a major role in catalyzing academic institutions' provision of training opportunities to acquire data-driven skills for scientific inquiry at several scales**. Examples include:
  - Informal offerings such as meetings of student groups to exchange tips (e.g., Berkeley's The Hacker Within), more structured activities such as seminar series, "hackathons," (e.g., AstroHack Week at NYU's MSDSE), Python boot camps, or Data Carpentry workshops; and
  - New curricular offerings and new undergraduate and graduate educational programs (e.g., UC Berkeley's Data8 undergraduate course and new Division of Data Sciences;

UW's Advanced Data Science Option for doctoral students; NYU's new doctoral program in data science).

- **The DDD initiative has fostered robust collaborations between computational methodologists and domain-based scientists**.
  - Particularly notable are the MSDSEs' project-based collaborations (the Data Science Incubator Program at the eScience Institute, the Machine Shop program at BIDS, and SEED grants at NYU's MSDSE).
- **The DDD initiative has had a limited effect, to date, on promoting changes in institutional mechanisms for retaining data-driven scientists in academia**:
  - Although individual DDD Investigators, as well as MSDSE data science fellows, research scientists, postdoctoral fellows and graduate students have advanced professionally within academic research settings, the "host" institutions have not implemented formal changes in the criteria for evaluating data-driven scientists' contributions.
  - Many respondents cautioned that four to five years was insufficient time to implement changes to deeply embedded, long-standing tenure and promotion procedures or to detect much evidence that could point to progress.
  - MSDSE institutions have begun experimenting with alternative career pathways for data-driven researchers; limited data exist, to date, to assess the outcomes of these experiments.

2.4.1 Retention of Data-Driven Scientists in Academia

Because industry can offer individuals with the types of skills increasingly needed in data-driven inquiry lucrative job offers, a major goal of the DDD initiative is to retain data-driven scientists in academia. Two mechanisms that the DDD initiative has advocated for achieving this goal including: (1) promoting changes in the criteria used in academic research institutions to assess data-driven scientists' contributions and (2) encouraging these institutions to establish alternative career pathways in academia for data-driven researchers.

Evidence to date shows that **some DDD Investigators' institutions and all three MSDSE institutions are actively exploring options for improving retention of data-driven scientists**; **however, they have encountered ongoing challenges** both in persuading colleagues that there is any need for changes in tenure/promotion criteria and in establishing sustainable long-term alternative career paths for the types of personnel needed for data-driven scientific inquiry in academic research institutions:

- By March 2017 (the date surveyed), 10 DDD Investigators were tenured faculty and 4 others were in tenure-track positions, but most administrators at DDD Investigators' institutions reported no meaningful changes in institutional tenure and promotion criteria or procedures.
- At the MSDSE institutions there was little evidence that formal criteria for tenure and promotion review of data-driven scientists had changed, although some MSDSE faculty (tenure-track and not) or fellows with joint appointments believed that evidence of their MSDSE-related contributions in their review dossiers had helped them receive a promotion or tenure.

- The MSDSEs at each institution have each attempted to establish agreements on appropriate tenure criteria with academic departments for jointly appointed faculty, but data on the status of these agreements was unavailable.
- Data from the evaluation illustrate both the competition from industry that academic institutions face for researchers with experience working in a data-intensive scientific setting, and concern among interview respondents about opportunities for data-driven scientists in academia.
- Yet, other respondents suggested that the MSDSEs have provided unique opportunities to their fellows, postdocs, and graduate students that would make these individuals attractive candidates for academic positions.
- Senior administrative officials at the MSDSEs voiced strong support for finding ways to sustain alternative career options for data-driven scientists at their institutions, but also expressed concern about the long-term viability of these types of positions.

Below, we highlight some individual instances of data-driven researchers' successful professional advancement in academia; describe how MSDSE institutions are exploring options for retaining data-driven scientists; and illustrate respondents' concerns about persistent challenges of expanding these efforts and ensuring sustainable change.

Criteria for Advancement at DDD Investigators' and MSDSE Institutions

The majority of DDD Investigators (10 of 14) are associate or full professors with tenure and the other four are tenure-track assistant professors.[82] Although 9 of 13 DDD Investigators interviewed acknowledged not knowing exactly why they had received tenure or a promotion, three of these nine highlighted the external validation of the DDD award, and perceived that it had some influence on the tenure/promotion process.

> *Universities place a high value on external funding, and I think it is certainly true to say [that] receiving the DDD Investigator award was a crucial component of successfully going through promotion and tenure here at the university. Was it the only thing that made it go through? [That's] unlikely, but I think it was a key ingredient, because of the large dollar value ... and also the selectivity of the award process. (DDD Investigator)*

Administrators (department chairpersons, deans or vice provosts, research center directors) similarly characterized the DDD Investigator award as "something special" and "not just another ... [federal agency] grant." Another administrator credited the award with helping the recipient obtain early tenure, despite having a less traditional profile and fewer publications than typical for someone at that career stage.

Despite the effects of the DDD Investigator award for the individual recipient, administrators at DDD Investigators' institutions disagreed about the effect of the DDD Investigator award on

[82] Although eight of these were identified by the DDD team at the time of their application for the DDD Investigator Award as experienced researchers, and six as early-career researchers, the evaluation did not collect data on pre-post award changes in faculty rank or tenure.

their university's standard tenure and promotion review processes. Although the DDD Investigator award reportedly signaled the merits of an individual DDD Investigator for purposes of that individual's tenure or promotion case, most administrators interviewed saw no effect of the DDD Investigator award on tenure or promotion criteria. One of these administrators noted that existing criteria in the form of "intellectual service" *already* captured the contributions of data-driven scientists:

> *One value that's very strong in [this university] is that we care about someone's impact above how many citations or publications they've had. When I'm doing … reviews, the number one thing the whole college looks at is the notion of impact and that often comes in the form of a startup, or did a company take some of these things and use them in a product? … In [the DDD Investigator]'s work, we know there are thousands of others using the tools [the investigator] built, and the fact that that happens trumps [the] publication record. (University administrator)*

One contrasting view did emerge from a senior-level administrator at another DDD Investigator's institution. This administrator reported that the DDD Investigator award had catalyzed a meeting of senior university administrators and early career open science advocates who discussed the fact that traditional tenure criteria did not capture the impact of researchers providing new types of software to help "thousands" of other researchers. In this administrator's view, "Awards like the Moore award legitimized what [the DDD Investigator] is doing in a high-profile way, and that has been enormously helpful to effect … culture change on campus." However, we emphasize that this viewpoint was an isolated example, and that the institution had not made changes (at the time of our interview) to tenure criteria.

Little formal change in review criteria for promotion and tenure for data-driven scientists has taken hold at the MSDSE institutions. Individual faculty, both tenure-track and not tenure-track, have received tenure or promotion at each MSDSE, and some of these individuals believed that inclusion in their dossier of research products such as software and engagement in MSDSE-related activities had bolstered their case. Despite this perception that some review committees' practices looked favorably on these types of contributions, several MSDSE respondents saw challenges related to the departmental power to grant promotions and tenure. At one MSDSE, a respondent reported that the MSDSE lost a senior-level candidate for a jointly funded position because the relevant department was unwilling to hire the candidate with tenure. A member of the leadership team at another MSDSE noted concern about the role of external letters in the prospects for data-driven scientists with joint appointments:

> *Overall there is still a reluctance to have people outside the traditional box because it is hard to tenure people like that and the system is built on getting letters from people that have a reputation in pretty narrow fields and when people do stuff outside of those narrow fields it can be really hard.*

Due in part to these types of concerns, each MSDSE includes a Working Group focusing on career paths and alternative metrics for data scientists, and at each MSDSE, this group has worked to establish a set of agreed criteria for the tenure and promotion of faculty jointly appointed to a department and the MSDSE. Data from interviews and MSDSE annual reports provide little insight into the outcome of these efforts, beyond the few individual instances of MSDSE personnel who have

received promotion or tenure. Moreover, interviews with some administrators at the MSDSE institutions suggest a continuing belief in the efficacy of the existing practice of using external letters of support to evaluate a data-driven scientist.

> *The path to changing or amending [the tenure review] process is really at the level of getting faculty to help to educate their chairs on what they want to be evaluated on in their third-year review, in what should be included in the letter that requests from external evaluators. … [E]very high-quality institution of higher education relies on external letters of review as a key component of their evaluation process, and we send letters to those external evaluators enumerating the domains in which we want faculty to be evaluated. So open-science related, data-related products, tools, and datasets are part of the effort. (Administrator at an MSDSE institution)*

Career Pathways for Data-Driven Scientists in Academia

Data from the evaluation illustrate both the competition from industry that academic institutions face for researchers with experience working in a data-intensive scientific setting, and concern among interview respondents about opportunities for data-driven scientists in academia. Across a total of 47 former members of DDD Investigators' research teams who left these investigators' labs between 2014 and 2016, 19 (40 percent) had subsequently accepted a research or data scientist position in industry, and 12 of 13 DDD investigators reported that at least one former lab member had transitioned to an industry data scientist role.[83] Among the remaining 28 former lab members, 15 had accepted a faculty position (tenure-track, n=14, or non-tenure track, n=1); six, a postdoctoral appointment; four had entered a degree program, and the three others had taken some other non-academic, non-data-scientist position.

Although the majority (7 of 11) of current members of DDD Investigators' labs interviewed expressed interest in a job in academia, a few DDD Investigators expressed concern about their lab members' prospects for finding an academic position—they had experience, after all, with former members leaving for industry positions. One of these DDD Investigators, though concerned, also found reasons a more positive outlook:

> *The problem still happens at some universities where you have these amazing students. Really the smartest humans. They all said "we love him, we want to hire him, but we can't figure out what area he belongs in and who will write his tenure letters in seven years?" It hasn't totally clicked over yet. However, there are hiring committees who are willing to take risks. I think he will probably end up at [one of the MSDSE institutions] or somewhere open-minded like that. It looks like we have positions now we may not have had a few years ago.*

At the MSDSEs, some respondents also voiced concern about the career paths for data-driven scientists after they left the MSDSE:

> *Some of our fellows have been on the job market and ... for some of them, their involvement in [our MSDSE] was valued, but ... in some cases, it was simply ignored*

[83] Data on subsequent jobs of former members of DDD Investigators' labs come from responses by the 13 DDD Investigators to a survey conducted as part of the evaluation. See Appendix B for additional details.

> *... [or] it was seen with outright skepticism. [H]ow do we change the ... broader culture to recognize these things, since most of our fellows will not get a job here...?*

Other administrators and MSDSE leaders suggested that there may be enough growing appreciation across academia for the unique qualifications of MSDSE fellows, postdocs, and graduate students that these individuals could be attractive candidates for academic positions as they enter the job market. One MSDSE respondent pointed to astronomy as a field that has already noticed the importance of people with knowledge in the domain and in modern data management skills. She described a graduate student who had (prior to the MSDSE award), against the advice of his advisor, spent time learning data-driven skills that resulted in a thesis "that couldn't have been done about these modern techniques. ... it was clear [he] was doing just game-changing stuff ... and [he] got the postdoc precisely because [he] could do astronomy and modern data management." This respondent saw the potential for a similar dynamic to enable MSDSE participants to compete for top jobs in academia.

At another MSDSE institution, an administrator also saw signs of change in how departments and institutions viewed the interdisciplinary experiences provided to participants at the MSDSEs. Although he acknowledged the potential for such people to "fall between the cracks," he was optimistic about the potential for institutional changes in academia:

> *This work is a game changer. I'm going to a meeting [of academic leaders from multiple institutions] next week, and NYU, UW, and Berkeley are known for this experiment. I think it has already had an impact on the recruitment side, [and] I would not be surprised if it changes how we do business. Many disciplines are evolving where this is not just a nice add-on, but is essential.*

A data science fellow at one MSDSE thought that the opportunities he had had at the MSDSE, including the chance to focus on research, advise PhD students, write grants, and start new collaborations on his own, would made him more competitive than other candidates on the job market. One MSDSE made a similar point in its July 2016 renewal proposal:

> *[MSDSE] postdocs will be more competitive for modern faculty positions requiring extracting knowledge from large, heterogeneous, noisy datasets, and collaborations with researchers who develop the methods that enable this.*

Each MSDSE can also point to a few examples of former data science fellows or postdoctoral fellows who have transitioned to faculty positions in other academic research institutions (in a couple of cases, an individual has transitioned from one MSDSE to a more senior role in another). Nevertheless, to date, the MSDSEs have a relatively small number of "alumni" and thus little data available to examine the types of job placements and professional opportunities those former members have found.[84] Moreover, the types of opportunities for upcoming cohorts of researchers entering the job market may or may not resemble the positions available for earlier cohorts, as trends in industry and academia may shift—and there are some early indicators that the MSDSE host institutions are exploring new options for retaining these researchers.

---

[84] Further, our findings reflect data current as of early 2017, well before the annual cycle of academic (and other) hiring had been completed.

Not only are the MSDSEs are making efforts to prepare fellows for existing academic positions outside their host institutions, they are also exploring ways to establish sustainable alternatives to existing academic pathways for data-driven scientists. In particular, MSDSE respondents described efforts to make sustainable the new types of career positions, such as research scientist and data science fellows, that the Moore-Sloan support has enabled for the grant period.

There appears to be strong support among senior administrative officials at the MSDSEs for finding a way to sustainably support these "off-tenure-track" types of data-driven scientists in academia. One administrator observed that there is a necessary role for individuals who can bridge the divide between researchers producing work in the foundations of data science (e.g., statisticians, computer scientists) and researchers who need to apply that work to specific cases within different domains of science. Further, they argued that the data science enterprise needs such individuals in order for it to succeed as a legitimate field.

Leaders at two MSDSEs saw potential in creating partnerships with university research libraries especially as the mission of these service-oriented units has begun to evolve.

> *When the grant began, [the] campus independently was trying to restructure its approach to IT. Research IT found very good collaborators here [at the MSDSE]. It is extremely unusual with IT department to interact with researchers to the point of writing grants together. .... [W]e came just at the right time to amplify it. IT needed a place to reach out to actual researchers; otherwise, they would be knocking on doors department by department. Also the library, which was trying to push on the data management problem, which is now a requirement of every grant proposal and researchers don't know how to write it. The library has been hosting workshops here to engage with faculty. ... Both the library and IT ... began to engage in the intellectual part of the campus, and [the MSDSE] has catalyzed that.*

One of these MSDSEs was also exploring the possibility of modeling such positions after the kinds of clinical professorships used in some medical schools; another MSDSE was experimenting with a "salary buyback" program for their data scientists. Under this program, data scientists who obtained partial salary support from external grant funding would "get back" some of this salary from the university in the form of additional research funding. Other potential options mentioned by MSDSE respondents included joint support from external grants and university funding for two-year fellowships that would combine a light teaching load with a postdoctoral appointment; and the possibility of partnering with industry to support some of these personnel. One MSDSE respondent noted:

> *I don't think we leveraged the big companies enough. I think [some of these big technology companies] should pay for the fellows, that would be nothing for them. That would be huge for [us]. We should say, "you are hiring these people [because] we are valuable for you, [we are] training these people."*

These MSDSE leaders saw both a challenge and an opportunity to take risks in establishing a more permanent role for such individuals. "This [MSDSE] is not a department … and a departmental model is not a good fit. We should be brave and invent our own process." The challenge, according to this leader, was to give data scientists

> *the autonomy and prestige that they deserve ... If you are not careful on a university campus, ...since [data scientists] are not faculty, they will be used and not appreciated ... These people are much more important than that.*

2.4.2 Fostering Collaboration

The DDD initiative has nurtured data-driven science by facilitating productive collaborations. DDD Investigators have benefitted from the opportunities provided by the DDD initiative to network and build community. The MSDSEs have demonstrated their value to host institutions via (1) synergistic engagement with other research centers and grant-funded research initiatives at their universities and (2) structured, project-based collaborations between domain scientists and computational methodologists focused on solving concrete problems.

DDD Investigator Symposia

Twelve DDD Investigators highlighted the annual investigator symposium as valuable for community building, and seven reported collaborations with another DDD Investigator (or several), nearly all of which resulted from this annual symposium or from the DDD-sponsored weeklong "Barn-Raising for Data-Intensive Discovery" workshop:

> *The annual symposium was a great experience for me and exposed me to a lot of new tools that other investigators are using, which has enhanced my productivity. I have learned a lot about open science through those symposia. The connections with other investigators have also been great. Some of us have worked together on the side collaborating in other workshops. For instance, we know we can bounce ideas off of one another and seek advice from each other. We are connected to each other on a Slack channel. I would say there has been a lot more networking in a way that has been very beneficial that is associated with this grant, than other grants.*

Other DDD Investigators valued these events as opportunities to learn about how their colleagues had overcome the unique challenges that data-driven scientists sometimes encounter when seeking grant funding or attempting to publish.

Synergistic Collaborations at the MSDSEs

All three MSDSEs also reported many examples of successful collaborations with other research centers and engagement of MSDSE personnel in a variety of research initiatives. For example, BIDS staff and faculty work with the Berkeley Initiative in Global Change Biology (BIGCB), the Social Sciences Data Laboratory (D-LAB), the National Energy Research Scientific Computing Center (NSERC), and the Simons Institute for the Theory of Computing; BIDS personnel have also participated in several successful proposals for institutional or multi-institutional big data research or training initiatives:[85]

- BIDS collaborated with other UC Berkeley faculty and both UW and UC San Diego to operate NSF's Western Big Data Regional Innovation Hub one of four regional data science hubs.

[85] Look for information about BIDS' involvement in other UC-Berkeley big data educational and training initiatives later in this chapter.

- A BIDS Senior Fellow is the PI for the NSF-funded Data Science for the 21st Century (DS421) research training program which will apply data-driven approaches to research challenges in the interaction of human and natural systems.
- BIDS supported the proposal and will share a data science fellow with the U.S./China Clean Energy Research Center for Water-Energy Technologies (CERC-WET), a multi-institutional, bi-national collaboration with China targeting sustainable water use in power generation, climate impact of energy-water systems, and treatment of management of wastewaters, (U.S. funding from Department of Energy).
- **NIH BD2K Biomedical Training Grant.** For this doctoral training grant, BIDS partnered with faculty from biostatistics, computational biology, computer science, epidemiology, integrative biology, molecular and cell biology, neuroscience and statistics to support the proposal; BIDS will share a data science fellow and provide training workshops.

At the eScience Institute there are collaborations with the Applied Physics Laboratory, the Computational Neuroscience Center, the Human-Centered Data Science Lab, the Institute of Neuroengineering, and the Virtual Planetary Laboratory; eScience institute also contributes to (and benefits from) NIH Big Data for Genomics and Neuroscience Training Grant, the Center for Genomics and Public Health, and the Computational Molecular Biology Program. Notable collaborations in which eScience personnel are heavily engaged include:

- eScience Institute members along with other UW faculty collaborated with UC-Berkeley and UC-San Diego to operate **NSF's Western Big Data Regional Innovation Hub,** one of four regional data science hubs.
- multiple faculty and data science/postdoctoral fellows from eScience Institute work on the **Large Synoptic Survey Telescope (LSST)**.
- eScience Institute has several joint activities with **Urban@UW**, a partnership between an interdisciplinary group of faculty from 20 academic units at UW, city government and community stakeholders to use data analytics to address urban challenges in environmental, health and housing. These joint activities include:
  - eScience Institute's Associate Director and a Senior Data Science Fellow are overseeing UW's portion of the **Cascadia Urban Analytics Cooperative**, a collaboration enabled by a $1 million donation from Microsoft; and
  - Co-sponsorship (with the federal MetroLab Network) of a two-day **MetroLab Workshop on Big Data and Human Services.**

Centers and initiatives with which NYU's MSDSE collaborates include the Center for the Promotion of Research Involving Innovative Statistical Methodology (PRIISM); the Music and Audio Research Lab; the Global Institute of Public Health; and the Center for Neural Science. Other notable collaborations involving faculty, fellows, and research scientists at NYU's MSDSE include:

- Multiple research projects in statistical methods and software in support of the **DIANA/HEP project, a software development community for high energy physics** (with funding from NSF's Software Infrastructure for Sustained Innovation program) include MSDSE personnel (from NYU as well as BIDS and UW's eScience Institute):

  - Efforts to integrate ROOT, the data analysis framework for CERN, in Jupyter Notebook form;
  - Analysis preservation of Large Hadron Collider experiments; and
  - RECAST, a tool allowing theoretical particle physicists to reinterpret searches by evaluating the sensitivity of a published analysis to a new model; and ADAGE, designed to add a web-based front-end and a back-end system to RECAST.

- **Sounds of New York City (SONYC)**, a project using machine learning, big data analysis and visualization, and citizen scientists to address noise pollution, is a collaboration between NYU's MSDSE/CDS, the Center for Urban Science and Progress (CUSP), the Steinhardt School of Culture, Education, and Human Development, the Tandon School of Engineering and Ohio State University (funded by a $4.6 million NSF grant).
- **Open Space, an open source visualization tool to showcase NASA's astrophysics, planetary science and Earth sciences and engineering activities and results** for the general public, middle and high schools, and citizen scientists, is a collaboration between NYU and the American Museum of Natural History, funded by a $6 million NASA grant.
- **Vizier, an NSF-funded collaboration to streamline data curation** and proactively structure, validate and repair big data sets, enable fast interactive exploration of data that automatically tracks the provenance of interactive changes to the data.

Incubator Projects at the MSDSEs

Another mechanism that the MSDSEs have used to demonstrate the unique contributions of an environment dedicated to data-driven scientists and practices is that of project-based collaborations between scientists and data-driven methodologists. Each MSDSE has its own name for these collaborations:

- The Machine Shop and BIDS Collaborative at BIDS;
- Data Science Incubator program and Data Science for Social Good programs at the eScience Institute; and
- Data Science Seed Grant program at NYU's MSDSE.

In each, the MSDSE accepts applications from scientists (or teams) who propose a specific, domain-based problem that would benefit from consultation with experts in data-driven methods or tools. Successful proposals receive a small amount of funding to conduct a short-term collaboration, typically one to three months, to work toward a solution. Exhibit 2.2 lists a subset of notable or recent project-based collaborations at each MSDSE.

For the Data Science Incubator Program incubator program at the eScience Institute, applicants commit to twice-weekly working sessions at the Data Science Studio. According to multiple respondents, the program has facilitated collaborations across a wide range of scientific domains, as well as helping build community and support from across the institution.

> *A very important thing about the Moore-Sloan funding is that it has enabled us to ... say yes to things and to run experiments and to provide some level of free services. And that's important. Many of the people who come in here for assistance, it's almost*

> *like a statistical consulting service on a grander scale, which is that they've got a problem that needs solving now or in the next three months. … They don't have time to write a grant. They don't have time to get money to pay a research center for somebody. So in some sense the [incubator projects] we're giving them is access to our data science and research side, the use of the [Data Science Studio]. They sort of become our friends for life.*

The DSSG program has also demonstrated the value of the eScience Institute to a broader community. In its first year (2015), DSSG attracted 11 proposals and 140 applicants from students in 15 departments. The four accepted projects engaged the expertise of the data scientists and research scientists in GIS algorithms, machine learning, and data visualization, as well as graduate, undergraduate, and high school students. It also garnered local media attention, and one project, a data-driven effort to improve the allocation of resources and programs targeting homelessness, led to grants from the Gates Foundation totaling more than $460,000.

At BIDS, there are similar programs. For BIDS' Machine Shop program, a scientist or lab proposes a domain-specific problem requiring a software-based solution amenable to a 1 to 3 month development time and designates a project liaison to engage actively with the team. BIDS personnel select projects, allocate resources, and provide consultation and expertise needed to develop a proof-of-concept tool for release under an open source license. Undergraduates may apply to work on a machine shop project (currently through Berkeley's Undergraduate Research Apprenticeship Program), and receive training in software development and computational problem solving by working alongside BIDS postdoctoral and computational fellows. BIDS Collaborative projects are semester- or year-long projects focused on real-world problems. BIDS personnel matches teams of four students to work on problems proposed by a set of pre-vetted partners from campus researchers or administrative offices, local non-profits or Berkeley-affiliated startups. Students earn academic credit and receive training in necessary data science skills from graduate student and faculty members, while also building data-driven skills.

NYU's MSDSE hosts an annual "seed grant" competition, in which it matches scientists and methodologists interested in pursuing a collaborative project.

Arguably, the incubator programs at the eScience Institute have seen more success than their counterparts at the two other MSDSEs (rates of participation appear stronger at UW, and some completed projects have received awards, media attention, and even external research funding), although the reasons are unclear. One MSDSE respondent reported that these projects fulfilled a need that, without a structure like the MSDSE to support it, would otherwise go unaddressed: "Like software development for research purposes. Where do you slot that into a normal [academic] course? ... Where would people ... go if they have questions? It's not obvious."

2.4.3 Opportunities for Training in Data-Driven Skills and Methods

In addition to increasing the credibility of data-driven science, the DDD initiative has also fostered awareness across many domains of the need to give more people the skills to work with data more efficiently and more reproducibly. A wide range of respondents, including DDD Investigators, non-awardees who competed for a DDD Investigator award, postdoctoral and graduate students, university administrators, and MSDSE leaders, cited the need for graduate students, particularly those in the physical, earth, and life sciences, to have more opportunities to acquire software development,

computational, and statistical skills and knowledge. The DDD initiative has facilitated increased training opportunities for students to acquire data-driven research skills.

Training Opportunities at DDD Investigators' Institutions

DDD Investigators who have taken a lead role in educating graduate students in data-driven science skills noted the influence that their efforts have had on colleagues and administrators. One DDD Investigator has worked with a graduate student to develop and post online a modular, semester-long Data Carpentry course for graduate students in his field. After inviting others to download the materials and develop their own version of the course, the investigator had had multiple conference calls and was noticing growing indicators of interest on github. Other investigators highlighted increasing awareness among their colleagues:

Exhibit 2.2: Project-Based Collaborative Programs at the MSDSEs: Notable Projects

Machine Shop at BIDS
Cesium: a Machine Learning Time Series Platform with a library for time-series analysis and a web platform for non-expert users to interact with the library. Led by a DDD investigator and Project Jupyter's Lead
TextThresher: students can work on frontend or backend software tasks to develop this crowd-based text analysis software that enables researchers to more quickly extract data from large text corpora.
Inselect: a desktop application for automatically segmenting images of insect specimens from whole-drawer digital scans of museum collections for archiving and annotation.
BIDS Collaborative
University of California Wellness Project: Measure the effectiveness of university support services for food, housing and financial insecurity
Berkeley Research Development Office: Improve means to connect faculty with research funding opportunities
Brainspell: Help develop a database that matches 3D coordinates in the brain to 10,000 scientific articles
Data Science Incubator Program at the eScience Institute
REDPy (Repeating Earthquake Detection using Python) automatically detects potential repeating earthquakes using an online clustering algorithm in real time or with archived data
Scalable manifold learning: A software suite for scaling a broad class of manifold learning methods to very large data sets (such as spectroscopic data from the Sloan Digital Sky Survey).
Cloud-enabled tools for the analysis of subsea HD camera data: A framework for analyzing large streams of data from the NSF-funded Ocean Observatories Initiative
Data Science for Social Good (summer program) at the eScience Institute
Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks: Expand OpenStreetMap with sidewalk data to enhance pedestrian wayfinding, particularly for users with disabilities.
Predictors of Permanent Housing for Homeless Families: analyze homeless program enrollment data to identify predictors of successful re-location into permanent housing, and to investigate families' transition between different programs and episodes of homelessness
Mining Online Data for Early Identification of Unsafe Food Products: an exploration of the efficacy of text mining of food product reviews to aid in the identification and ranking of food safety issues measured against Food and Drug Administration enforcement reports.
Data Science Seed Grants at NYU's MSDSE
Standard Cortical Observer: joint work between an MSDSE fellow and faculty from Center for Neural Science to create a repository and tool to streamline computational modelling of brain responses to sensory inputs
Estimation of Multiple Tissue Compartments from Magnetic Resonance Fingerprinting Data: Used Magnetic Resonance Fingerprinting (MRF) reconstruction that accounted for the presence of multiple compartments in a voxel. The method was validated with simulated data, as well as with a controlled phantom experiment
Statistics Meets Transcriptomics: Time-Series Responses of Post-Transcriptional Regulation By Families of Conserved RNA Structures: a collaboration between computational/statistical biology and a domain expert in biology

> *Scientists [in my field] ... are beginning to wake up to the fact that [the field] is becoming an extremely data-driven, analysis-driven field and that students need help. It's about statistics, it's about computation. They've always seen the importance of data, it's more the scale of the data making them realize that for their lab to continue to function effectively, students need to learn new stuff.*

Faculty and administrators have started to recognize that graduate students need the right training in fundamental computational and quantitative methods; otherwise, they are unprepared to manage the large amounts of data that scientists can now collect. One university administrator described how faculty were borrowing the DDD Investigator's practice of using short, organized training events, rather than full-length courses, to train graduate students in software and other data science tools specifically applicable to their research domain. At another university, one department had asked its (non-tenured) DDD Investigator to teach the graduate-level course in the computational foundations, a task typically reserved for tenured faculty.

Despite raising awareness of the need for training graduate students in these methods, one investigator expressed some doubt that faculty would endorse a large data-driven science initiative:

> *I think [my colleagues] see the value in educating their students in these areas, but I'm not sure they would say that the best way to address them is through a data-driven science initiative. I think there's more enthusiasm for hiring quantitative faculty, let's say that, but not necessarily as a separate initiative.*

In contrast to this viewpoint, some faculty and administrators at the MSDSE institutions viewed these environments as providing a centralized home for graduate students to acquire a range of data-driven skills.

Training Opportunities at the MSDSEs

Although the primary mission of the MSDSEs is to cultivate data-driven research, each MSDSE also has an Education and Training Working Group charged with facilitating and promoting training opportunities for students and researchers. Some training takes place in structured workshops or boot camps hosted or sponsored by the MSDSE; other training occurs in the context of an incubator project, or through serendipitous interactions in a shared space. Faculty value the training that provides their graduate students with skills needed to contribute productively to research, and all three MSDSEs have faculty or other members who teach courses and who have participated in the development of new certificate or degree options for data science. All three MSDSEs host Data Carpentry and Software Carpentry workshops, as well as Python boot camps, hackathons (e.g., AstroHackWeek, GeoHackWeek, NeuroHackWeek), and various workshops specific to each MSDSE (BIDS' 2015 Multiphysics Object-Oriented Simulation Environment [MOOSE] Framework Workshop; a 2015 NSF-funded graduate Data Science Workshop at the eScience Institute; the 2016 Atlantic Causal Inference Conference co-sponsored by NYU's MSDSE). The eScience Institute at the University of Washington has partnered with Software Carpentry and Data Carpentry to serve as a regional hub for instructor trainings, and BIDS has a paid partnership with Data Carpentry and Software Carpentry.

In addition to events common to the three MSDSEs, BIDS also hosts weekly meetings of the university's The Hacker Within collaborative (started by a former data science fellow). A faculty member affiliated with BIDS received an NSF Research Training (NRT) grant for "Data Science for

the 21st Century" to support graduate student research training. BIDS has also played a central role in the faculty-led push for a Data Science education program (DSEP), including the introductory-level Foundations of Data Science (Data8) course launched in the Spring of 2016. Combining instruction in computational thinking and statistics concepts with hands-on exposure to real-world data from a range of fields, the course is open to all students regardless of intended major and is linked to connector courses that focus on a particular data domain (e.g., civil engineering; cognitive science) and more advanced extender courses. A BIDS Senior Fellow assists faculty preparing courses for Data8; in 2017, tutoring sessions for the course have taken place at BIDS; and a JupyterHub hosts all Data8 course materials, assignment submission, and grading tools (developed by BIDS personnel).

Spurred by efforts of some of the same faculty who advocated for Data8 and the Data Science education program, UCB approved the creation of a new Division of Data Sciences in 2016, and in May 2017 hired an interim dean. This dean will lead efforts to define this new unit, fundraise, recruit faculty, create an undergraduate program, and determine what graduate-level programs to offer.

> *What would have happened without BIDS is that we would have had eight to nine data initiatives, who may have been talking to each other, but would not [have] come together in quite this way. I don't think the commitment to create a data science division would have happened, certainly not on this time scale. It is very fast for a university to make a decision to do something like this. I would credit BIDS pretty directly for being a catalyst for it. (University administrator)*

At NYU's MSDSE, training opportunities in addition to Data Carpentry and Software Carpentry workshops and HackWeeks have included workshops on text-as-data workshop, a broad range of tutorials from NYU libraries in data science tools and practices, such git, github, Python, and ReproZip. The MSDSE's Education and Training Working Group helped sponsor the 2016 Atlantic Causal Inference Conference, with roughly 150 statisticians, data scientists, epidemiologists, and others attending. NYU's MSDSE has created a Junior Data Scientist position for master's level students who work on incubator projects under the mentorship of research engineers and faculty; and faculty have begun to incorporate teaching of reproducibility practices and tools in their courses. In addition, MSDSE has sought advice from UC Berkeley for development of an undergraduate Introduction to Data Science class at NYU and data science minor. NYU's CDS has a highly selective data science master's degree program (according to two respondents, about 15 times as many applicants as students admitted), and the university recently approved a new data science doctoral degree program; the MSDSE has played a key role in both programs. Master's students must complete a capstone project working with researchers at the CDS (many of whom are funded by or affiliated with the MSDSE): these projects come from real-world settings and require students to collect and process data, design a method to solve a problem, and implement and present the solution. For the PhD program, admitted students receive up to five years of tuition and stipend support.

The eScience Institute not only offers Software and Data Carpentry workshops, but also serves as a Pacific Northwest hub for joint Software Carpentry/Data Carpentry instructor training, hosts a weekly Python for Sciences seminar, and held a successful NSF-funded Graduate Data Science Workshop with more than 100 students. The Education and Training Working Group has developed new undergraduate and graduate courses to help students in biology and the social sciences gain data science and software engineering skills for data-driven discovery. The Institute obtained approval to launch a professional master's degree program in data science, offered by the Information School,

departments of Applied Mathematics, Biostatistics, Computer Science & Engineering, Human Centered Design and Engineering, and Statistics. The eScience Institute has also gained university approval for an Advanced Data Science (ADS) Option open to doctoral students in nine departments (Mathematics, Astronomy, Biology, Chemical Engineering, Computer Science and Engineering, Genome Sciences, Mathematics, Oceanography, and Statistics).

An administrator at UW pointed to these educational and degree programs as one of the biggest impacts of the MSDSE. One eScience affiliate noted that the Moore-Sloan funding enabled departments already working together under an IGERT grant from NSF to convince other people to get involved in the ADS Option, and that the Option might not have had permanence beyond the end of the IGERT funding. Instead, the MSDSE enabled the faculty to engage a broader range of stakeholders and gain university approval for it to be permanent: "Once [you] create an option, it has to be taught."

2.5 The DDD Initiative's Role in Scientific Discovery

Although accelerating scientific discovery is one of the ultimate goals of the DDD initiative, the more proximate goal of the DDD initiative is to facilitate the development of software, tools, practices, and other kinds of "research infrastructure" on which scientists—particularly those working with large or complex data—increasingly rely. One MSDSE leader highlighted the importance of the DDD initiative's focus on supporting this infrastructure: "It's just like roads and bridges ... nobody wants to pay for them, but when they collapse it's a mess."

The evaluation revealed clear evidence, described in Section 2.3, that the DDD initiative is meeting its immediate goals to support the people, practices, and institutions of data-driven science—its "roads and bridges"—but several grantees (DDD investigators and MSDSE leaders) cautioned that it was relatively early to look for signs of their role in new discoveries, especially given the lengthy process of peer review, revision, and resubmission. If the MSDSE, as one respondent proposed, is analogous to a university library in that it is a critical resource for research, but one whose role is not acknowledged formally then the MSDSEs' contributions to scientific advances may be difficult to quantify. As an example, this MSDSE respondent cited the discovery of gravitational waves:

> *The color maps and the charts in the paper were designed [here] ... and all of that was backed up in Jupyter Notebooks, so that people can actually explore the data ... on their own. But the discovery was the gravitational waves, which was not ours.*

Nevertheless, **the DDD Investigators and the MSDSEs provided robust publication records in their annual reports to the DDD team, and two *Practices* grantees track their projects' contributions to science online**. In their first annual reports, for example, investigators reported an average of seven publications and four software applications, packages, or similar tools since award receipt; a scan of *Google Scholar* shows continued productivity in subsequent months.[86] In our

[86] Annual reports for DDD investigators covered publications through January 2016 (data on publications were unavailable for one DDD investigator). Median number of new articles from January 2016 to May 2017 was four. These counts exclude preprints and articles in press or under review in annual reports. This scan is not comprehensive and may include some matching errors due to differences in titles and journal names between articles listed in annual report data and those found online.

interviews, a number of DDD Investigators cited particular publications or forthcoming works. These included a new method for recovering three-dimensional structure and shape of proteins from two-dimensional electron microscopy images;[87] a best paper award for solving long-standing challenges with asynchronous Gibbs sampling;[88] and an algorithm for quantifying transcript abundance in RNA sequencing.[89]

Two of the *Practices* grantees also publicize their contributions to scientific advancements on their websites. For example, in November 2016, a team of researchers used the Julia language in a parallel computing environment to increase the speed of analysis of astronomical images from the Sloan Digital Sky Survey by 225 times over prior iterations.[90] A 2014 *Nature* article profiled iPython Notebooks,[91] and Jupyter's github site maintains a list of academic papers that include Jupyter Notebooks to enable readers to reproduce the results.[92]

Likewise, each MSDSE provided extensive documentation of the publications, software, and other research products produced by its individual members and affiliated faculty in their annual reports; each MSDSEs' website also featured additional publications not yet produced at the time of its most recent annual report. From the long list of publications from members of NYU's MSDSE, three notable examples include:

- A cover article in *Science* describing a computational model that classifies visual objects with human-like performance;[93]
- A paper in *Physical Review Letters* describing a Large Hadron Collider experiment that takes advantage of a novel statistical method for detecting sub-atomic particles;[94] and

---

87 Katsevich, E., Katsevich, A., & Singer, A. (2015). Covariance matrix estimation for the cryo-EM heterogeneity problem. *SIAM Journal on Imaging Sciences, 8 (1*), 126–185. https://doi.org/10.1137/130935434

88 De Sa, C., Olukotun, K., & Re, C. (2016). Ensuring rapid mixing and low bias for asynchronous Gibbs sampling. arXiv:1602.07415 [cs.LG]. Retrieved from https://arxiv.org/abs/1602.07415v3

89 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods, 14,* 417–419. https://doi.org/10.1038/nmeth.4197.

90 http://juliacomputing.com/press/2016/11/28/celeste.html

91 See: Shen, H. (2014). Interactive notebooks: Sharing the code. *Nature, 515,* 151–152. https://doi.org/10.1038/515151a

92 See: https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#reproducible-academic-publications, accessed May 18, 2017.

93 Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350,* 1332-1338

94 G. Aad et al. (ATLAS Collaboration) (2015). Search for Dark Matter in Events with Missing Transverse Momentum and a Higgs Boson Decaying to Two Photons in pp Collisions at s√=8 TeV with the ATLAS Detector. *Phys Rev Lett. 115*, 131801

- A paper in *The Astrophysical Journal* describing a data-driven computational model that derives the spectroscopic profile of "new" exemplar stars based on a small training set of reference objects;[95]

From BIDS members, notable publications include:

- A paper in *PLoS One* describing a novel method derived from information theory to estimate the number of species in a large spatial area too large to census directly;[96]
- A *Nature* paper that uses data-driven methods to generate a detailed semantic mapping from fMRI data collected while subjects listened to narrated stories;[97] and
- A *Science* paper that reveals a strong correlation between slow, semi-periodic seismic deformations and subsequent large earthquakes over an 80-year period.[98]

Finally, from members of the eScience Institute:

- Published in *Nature*, a new method of analyzing neural firing rates that revealed cortical evidence contradicting a widely-accepted theory that decision making occurred via gradual accumulation of evidence;[99]
- A paper in *Physical Review D*, showing a relationship between the number of voids in galaxy redshift surveys and the equation of state of dark energy;[100] and
- A paper in *Water Research* using quantitative PCR assays to describe differences in microbial communities codigesting waste restaurant oil combined with wastewater sludge to the microbial communities digesting wastewater sludge alone.[101]

Despite evidence of lengthy publication records from DDD grantees, a few caveats are appropriate when assessing evidence of the initiative's contributions to scientific advances:

---

95 Ness, M., Hogg, D.W., Rix, H.W., Ho, A.Y.Q., & Zasowski, G. (2015). *The Cannon*: A data-driven approach to stellar label determination. *The Astrophysical Journal, 808*. Retrieved from http://iopscience.iop.org/article/10.1088/0004-637X/808/1/16/pdf

96 Harte, J. and Kitzes, J. (2015) Inferring regional-scale species diversity from small-plot censuses. *PLoS One, 10*, e0117527. https://doi.org/10.1371/journal.pone.0117527

97 Huth AG, de Heer WA, Griffiths TL, Theunissen, FE & Gallant JL (2016). Natural speech reveals the semantic maps that tile the human cerebral cortex. *Nature, 532*, 453-458.

98 Uchida, N., Iinuma, T., Nadeau, R.M, Burgmann, R. & Hino R. (2016). Periodic slow slip triggers megathrust zone earthquakes in northeastern Japan. *Science, 351*, 488-492

99 Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C. & Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature, 520*, 220-223

100 Pisani, A., Sutter, P.M., Hamaus, N., Alizadeh, E., Biswas, R., Wandelt, B.D., Hirata, C.M. (2015). Counting voids to probe dark energy. *Physical Review D, 92*, 083531.

101 Ziels, R.M., Karlsson, A., Beck, D.A.C., Ejlertsson, J., Yekta, S.S., Born, A., Stensel, H.D. & Svensson, B.H. (2016). Microbial community adaptation influences long-chain fatty acid conversion during anaerobic codigestion of fats, oils, and grease with municipal sludge. *Water Research, 103*, 372-382.

1. First, some portion of the publications listed by grantees likely originated prior to receipt of DDD funds. As is typical for many recipients of research grants (from any funder), many DDD Investigators and MSDSE researchers had well-established and productive research programs pre-DDD, but their publication data do not distinguish publications that were already in preparation pre-award from those that began post-award.

2. Second, even with a longer period of elapsed time, it is difficult to attribute particular scientific advances to *any* research funding program, including the DDD initiative. Because grant decisions often rely on prior evidence of merit, individual and institutional grant recipients (like the DDD Investigators and the MSDSEs) tend to have multiple sources of external funding, making it difficult to track the unique contributions of each funding source.

3. For DDD grantees in particular, however, it would be difficult to distinguish publications derived from "data-driven" scientific methods from those not: it is plausible that experts in the relevant research domains would disagree over such classifications as "data driven," as well as over discussions of which publications represented bona fide scientific "discoveries." Moreover, as described below (see Chapter Three), there are not current standards in widespread use for citing software and other tools produced by data-driven scientists that may enable discoveries by others. Although the adoption of such standards in the future may allow tracing data-driven scientists' contributions of software or other resources to scientific findings, at present these contributions remain difficult or impossible to track.

2.6 Synergies Between the DDD Initiative's Three Strategies

Examining the extent to which grantees in each of the three DDD initiative strategies have capitalized on the activities and outcomes of other grantees is an important question for the evaluation; evidence of synergies between the three strategies might suggest that there were critical interdependencies between the results of each strategy. At approximately mid-term, it appears that the strategies have been mutually reinforcing. Data from interviews and annual reports reveal a network of links between the *People* (DDD Investigators) and *Practices* strategies (in particular, Data Carpentry, Jupyter, and Julia); the *People* and *Institutions* (MSDSE) strategies; and the *Institutions* and *Practices* strategies.

Some DDD Investigators are affiliated with an MSDSE at their institution, leading to a much broader network of like-minded colleagues; DDD Investigators at other (non-MSDSE) institutions have collaborations with both DDD Investigators and other researchers at an MSDSE. Several DDD Investigators reported collaborations with *Practices* grantees. One investigator reported collaborations with two *Practices* grantees and another with a data science fellow at an MSDSE. At least two of the four *Practices* grantees are tightly connected to the MSDSE communities, and there are important links between a third *Practices* project and the MSDSEs.

One DDD Investigator described how connections to other DDD participants had resulted in software, a co-authored paper, and a collaboration with another DDD Investigator:

> *I've certainly been collaborating with Jupyter. Our software … was built by folks who were lead programmers in Jupyter. In addition, I co-wrote a paper with [another DDD Investigator] and my postdoc on the status of code in the field. And my students … have been … collaborating with [other DDD Investigators]. [One student] got in touch with them due to the … event [that some DDD Investigators]*

were involved in organizing … [that] was focused on postdocs and PhD students. [My student] flew to [another state] to work with [a DDD Investigator]…recently.

2.6.1 Links between DDD Investigators, Non-Awardees, and MSDSEs

Eight DDD Investigators either have collaborated with researchers at an MSDSE or have an affiliation with an MSDSE and actively participated in its community at the institution. Three other DDD Investigators reported delivering a talk or workshop at an MSDSE. For those investigators who share an institution with an MSDSE, the center has enhanced their visibility and expanded their network for potential research collaborations. Two DDD Investigators noted that the MSDSE at their institution had helped connect them to various data science events, and one noted that an affiliation with the MSDSE provided both training opportunities and opportunities to interact with other data-driven colleagues on campus.

Another DDD Investigator appreciated being in a resource-rich environment:

And I think that's the critical point, to have that critical mass to extend beyond a lab, and that has been the most successful way of promoting and changing culture. When …[I've] traveled to other universities, I really do see other schools [with] a lot of interest, or lamentation that they don't have the same level of collaboration. We have the envy of our peers [laughs]; I guess that's how you know you made it.

Non-awardees also benefitted from having an MSDSE at their institution.

It has ... attracted students and faculty to working on research projects in my lab. It has also created a home for my students and collaborators interested in data science. It has brought visitors to [my university] that I have since collaborated with. It has also facilitated relationships across campus that otherwise would have taken many more years to establish. These relationships have led to new programs and projects within my lab and my department. In sum, the DDD initiative has been hugely beneficial to my research, teaching, and quality of life.

The MSDSEs seem to provide a rich environment for investigators to build collaborations. Indeed, the problem for these investigators may be that there are multiple demands on their time. At least two investigators acknowledged having insufficient time to engage with all the activities going on at their respective MSDSEs. As one said:

I wish I had been able to take better advantage of the opportunities on campus through [the MSDSE] and take part in their seminar series but there is just not enough time.

2.6.2 Links between the *Practices* Grantees, DDD Investigators, and MSDSEs

Across the three strategies, two *Practices* grantees—Data Carpentry and Jupyter—appear to be the most deeply embedded throughout the DDD ecosystem. Each project has ongoing relationships with MSDSEs in addition to collaborations with individual DDD Investigators and other *Practices* grantees.

Data Carpentry

Data Carpentry is deeply integrated throughout the DDD initiative. Both BIDS and the eScience Institute have formal partnerships (Gold or Silver tier) with Data Carpentry that provide each MSDSE with an instructor training to build local capacity and on-site workshops. Data Carpentry has participated in several workshops and conferences with BIDS and eScience. For example:

- For the ImageXD (a joint project of BIDS and eScience researchers) inaugural event in June 2016, Data Carpentry worked with software developers and experts in computer vision, astronomy, earth science, and neuroscience to develop a blueprint for a Data Carpentry workshop on open source image processing.
- In January 2017, Data Carpentry held a hackathon at BIDS to develop a Data Carpentry workshop using Jupyter Notebooks to teach reproducible research practices.

DDD Investigator Ethan White is a co-founder (with Tracy Teal) of Data Carpentry and currently serves as board member, along with BIDS data science fellow Karthik Ram. The three have co-authored a journal publication and have active collaborations with Data Carpentry.[102] White has worked with Data Carpentry on the development of ecology data science workshops; Ram has worked with Data Carpentry to develop workshop content for the ROpenSci project (a software collective developing open source R packages to enable access to and integration of datasets, full text of journal articles, analysis, and visualization tools), which he founded at BIDS. Data Carpentry is also working with Ariel Rokem at the eScience Institute to develop a workshop on processing neuroimaging data. The eScience Institute is a Pacific Northwest Hub for the instructor trainings offered jointly by Software Carpentry and Data Carpentry, and data science fellow Ariel Rokem at the eScience Institute is a key contributor and trainer for these instructor trainings. Data Carpentry also partnered with the members of the eScience Institute and BIDS on a successful grant proposal for UCB, UW, and UC-San Diego to operate NSF's West Big Data Regional Innovation Hub (BD Hub: West).

Of the 13 DDD Investigators we interviewed, eight mentioned that they or members of their research groups had attended Data Carpentry workshops or its instructor training course. A non-awardee also commented on Data Carpentry's integration in the DDD ecosystem:

> *The [MSDSE] is actually running [Data Carpentry] workshops, and I'm a PI with [some other] faculty with a … training grant. So it was important, in order to get that training grant to use the environment on campus with [the MSDSE], so that's why we got involved with [the MSDSE] so that the students could get the extra training beyond the courses we will offer in the curriculum. So, I hear good things.*

Jupyter Notebooks

Jupyter Notebooks is also deeply integrated into the DDD initiative. BIDS provides Jupyter staff with a home (including office space and connections to the BIDS community), and it has several collaborations with BIDS senior fellows and other members. As one MSDSE member reported,

[102] Teal et al. (2015).

> *Jupyter anchored [BIDS]. If we have those three things: the space, the fellows, and Jupyter, we would feel like BIDS still exists even if other parts of the program we have been trying out were gone.*

Jupyter is collaborating with UCB's Data Science Initiative to enable Jupyter Notebooks and JupyterHub for use in data science courses and curricula. BIDS has also partnered with the Berkeley Research Computing program and Pacific Research Platform project to host a high-performance computing resource using JupyterHub. Individual members of the MSDSEs have featured Jupyter in several projects:

- A graduate teaching assistant and BIDS member implemented Jupyter Notebooks on a central JupyterHub so that 220 students enrolled in a computational science course could complete and submit course assignments without having to download and install software on their local machines; to manage and grade assignments, she developed and used a new package called nbgrader.
- This same graduate student developed a package (nbflow) integrating Jupyter notebooks with Scons to enable reproducible workflows (presented at SciPy 2016).
- Jake VanderPlas, the Director of Research in Physical Sciences for the eScience Institute, wrote both a bestselling book on data science and a free companion report providing scientists with an introduction to Python using Jupyter Notebooks; he has posted these notebooks on github.[103,104]
- An eScience Institute faculty member has developed a database for big data management, MyriaDB, that provides a Jupyter Notebook interface for analysis.

Jupyter also has a collaboration that includes DDD Investigator Matt Turk, who is co-PI on an NSF grant (the "Whole Tale" project) to build a pipeline for scientists to link code, data, and other information to online scientific publications, with Jupyter Notebooks serving as a front-end model.[105] In addition, Jupyter has a collaboration with the developers of Dask and iPython Parallel.

Finally, Jupyter provides DDD Investigators with a useful tool for teaching and collaboration. Ten DDD Investigators and several postdocs and graduate students working in DDD Investigators' labs have used Jupyter Notebooks for various purposes. One investigator described using Jupyter Notebooks to collaboratively prepare a manuscript for submission:

> *Jupyter Notebooks … is a really useful tool for collaborating on code. Right now I have a paper in review at a journal that details that a model that I developed in collaboration with an undergraduate, and we used Jupyter Notebook as an essential tool in that collaboration.*

---

103 VanderPlas, J. (2016a). *A whirlwind tour of Python.* Sebastopol, CA: O'Reilly Media.

104 VanderPlas, J. (2016b). *Python for data science handbook: Essential tools for working with data.* Sebastopol, CA: O'Reilly Media.

105 See: http://wholetale.org

A postdoctoral researcher in another DDD Investigator's lab reported using Jupyter Notebooks "all the time":

> *I'm in a [university] hack night group that meets weekly and I use it for that. It's more friendly for sharing with people not used to looking at a giant command line.*

The Julia Language

Based on data from interviews and annual reports, the Julia language appears to be less deeply integrated into the DDD network than is Jupyter. Nonetheless, there are active groups at all three MSDSEs contributing to Julia's development or who are developing domain-specific tools integrated with Julia. One of its lead developers (Stefan Karpinski) is a part-time research engineer at NYU's MSDSE. At BIDS, Kyle Barbary, a data science fellow, has developed packages for JuliaAstro, and Jupyter and Julia have teamed up to produce IJulia, a browser-based notebook interface for Julia. One of Jupyter's co-PIs, Fernando Perez, was a featured speaker at the June 2017 JuliaCon.

Dask and Numba

There is less information available to assess the role of Dask or Numba given that these tools received a DDD initiative grant relatively recently (in July 2016). Consequently, below we summarize what these two tools are intended to accomplish and any early indications that their efforts are gaining traction. The primary goal under the DDD grant to Continuum Analytics—divided about equally between the two tools—is for each package to release a stable version 1.0 by the end of the two-year grant period.

Dask is designed specifically for "people with custom or irregular computational problems that need parallelism. These are scientists, quants, [and] algorithm developers," according to lead developer Matthew Rocklin.[106] An example of its use comes from climate science, where the data might be "the temperature, air pressure, and wind speed measured every square kilometer of the Earth, for various altitudes, going back 50 years." When these kinds of datasets are too large for NumPy, Dask gives climate scientists a way to work with them, without sacrificing speed.

Dask was first prototyped at the BIDS Data Structures for Data Scientists workshop in 2015. In addition to its support from the Moore Foundation, Dask has received funding from the Defense Advanced Research Projects Agency (DARPA) and from the financial services industry. When industry funded, the developers agree to provide their client with a custom, proprietary solution—but the client agrees to contribute a percentage toward the development of Dask as an open source tool for others. By being open source, Dask benefits from its exposure to users who test it and identify limitations or bottlenecks for the developers to address. A key early area of focus has been improving types of analyses and computations that the Dask dataframes allow. Dask has been progressing toward its two-year goal of releasing a stable version 1.0 (as of May 2017, version 0.15 is the most recent release). Individuals at UW's eScience Institute have also been contributing to efforts to connect Dask with scikit-learn, a popular open source machine learning tool.

---

[106] Mayo, M. (2016, September). Introducing Dask for parallel programming: An interview with project lead developer [Blog post]. Retrieved from http://www.kdnuggets.com/2016/09/introducing-dask-parallel-programming.html

Numba is designed to speed the performance of NumPy, a widely used numerical computing package for Python. Although early versions of Numba existed in 2012, the modern version dates to 2015, after a major refactoring. Since the July 2016 DDD grant, Numba has been progressing toward version 1.0 (as of May 2017, Numba version 0.33 is the most recent release), and one of its developers reported a greatly improved debugging tool, a key milestones toward reaching its DDD goal. The DDD grant has made a significant contribution to this aspect of Numba: developing a robust debugging tool is not only difficult, but also the type of general improvement that a commercial client is likely unwilling to support.

3. The DDD Initiative in the Data Science Landscape

3.1 Introduction

The year 2012 marked a turning point in the era of "big data." The White House announced a $200-million Big Data Research and Development initiative, with major new funding programs in the National Science Foundation, the National Institutes of Health (NIH), and other federal agencies,[107] and philanthropic foundations began directing resources toward challenges presented by "data-rich, discovery-poor" sciences. 2012 also marked the launch of the DDD Initiative. Between 2012 and 2013, the volume of scientific publications featuring big data (coming from computer science, followed by materials science, computational biology, optics, biotechnology, biochemical research methods, statistics, and remote sensing) jumped more than 400 percent.[108] As attention to the challenges of extracting knowledge from large and complex datasets grew, efforts to promote data sharing intensified, and new voices began to advocate for shared code alongside shared data—in short, for transparent and reproducible research.

Within this ecosystem, the DDD initiative made a timely entrance in an environment where others were poised to act. Establishing the MSDSEs may, in and of itself, have prompted other institutions to move ahead with nascent plans for similar environments. Although the data are inconclusive, all three MSDSEs reported inquiries from multiple other institutions. The Moore and Sloan Foundations were also at the forefront of funders with early investments in data-driven science and scientists. Finally, the DDD initiative continues to be one of the few (or only) sources of grant funding for organizations focused on developing the kinds of open source tools and resources most needed by academic research scientists. The fact that the initiative focused directly and explicitly on tool development per se—and not in support of some other, primary goal, is a hallmark of the *Practices* strategy, and something that its grantees cited as unique among funders.

This chapter examines trends in the data-driven science "ecosystem," and the role of the DDD initiative within this landscape. Increased interest is evident in the emergence of new *data-driven science initiatives in academia* (i.e., the three MSDSE institutions and beyond) and new investments in data-driven science by key *funders in government, philanthropy, and industry*. Changes in the environments for conducting data-driven science are reflected growing momentum surrounding *open science and reproducibility* and in explorations of options for *sustainable careers in academic data-driven science.*

---

107 Office of Science and Technology Policy. (2012, March 29). Obama administration unveils "Big Data" initiative: Announces $200 million in new R&D investments [Press release]. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

108 The full paper describes the search terms and algorithm used to examine the growth of publications over 2009–2015 as indexed in the Science Citation Index Expanded, the Social Sciences Citation Index, Arts and Humanities Citation Index, Conference Proceedings Citation Index (Science), and Conference Proceedings Citation Index (Social Science & Humanities). See: Porter, A. L., Huang, Y., Schuehle, J., & Youtie, J. (2015). Meta data: Big data research evolving across disciplines, players, and topics. *Proceedings of the 2015 IEEE International Congress on Big Data* (pp. 262–267). Retrieved from http://ieeexplore.ieee.org/document/7207228/

3.2 Key Findings

The broader landscape for data-driven science has clearly changed over the past five years, based on evidence of progress in several key areas. First, the increasing prevalence of localized data science initiatives at major research universities suggests that the research and education terrain is becoming more hospitable to data-driven science, and there is evidence that the MSDSEs may have catalyzed some of these initiatives:

- All three MSDSEs reported multiple inquiries from other academic research institutions about their data science environments.
- Eight of the 15 universities invited by Moore Foundation and the Sloan Foundations to compete for an MSDSE award have since launched new "big data" or data science initiatives.
- In contrast to the MSDSEs, some of these university initiatives appear to focus more narrowly on research in applied sciences (e.g., biomedicine) rather than data-driven basic research in the life, physical, or earth sciences.

Second, although there is now more research support targeting data-intensive science research—primarily from federal sources, but also from philanthropic, and industry sources the DDD initiative was at the forefront, filling a funding need relatively early and it continues to play a prominent role in supporting data-driven science:

- The National Institutes of Health's (NIH's) Big Data to Knowledge (BD2K) program launched in 2013 with a focus on enhancing the utility of biomedical data with new standards for data sharing, support for research and development of software, methods, and tools for biomedical analyses, and enhanced training for using these tools.
- The National Science Foundation likewise launched a multi-component "big data" initiative to support environments enabling data-driven discovery and collaborations between domain scientists and methodologists.
- However, the Moore and Sloan Foundations appear to have been early leaders in establishing the MSDSEs as academic research centers for data-driven scientific research, and the DDD initiative remains one of relatively few sources of funding for individual researchers approaching basic (i.e., non-applied) scientific inquiry with a data-driven lens.

Third, there is increasing traction in the progress toward open science and reproducibility and more widespread recognition that open science translates into better science. The DDD initiative has fully embraced this movement and serves as an example by:

- Supporting researchers who have made important contributions to open science and reproducibility;
- Funding the development and dissemination of tools such as Jupyter Notebooks that enable reproducible research practices; and
- Including an explicit focus on reproducibility as one of six key themes of its MSDSEs.

Finally, there is suggestive evidence that some higher education administrators and faculty, as well as other scientific leaders, are beginning to recognize the value of multiple, sustainable career pathways for data scientists beyond the traditional tenure-track option.

3.3 Data-Driven Science Initiatives in Academia

The respondents interviewed for this evaluation perceive strong evidence of increasing interest in data-driven science. Across interviews with DDD investigators, non-awardees, and administrators, new data-driven science initiatives were reported in 10 of 12 institutions. At some DDD Investigators' institutions, the DDD Investigator award appears to have catalyzed campus-wide data science initiatives:

> *I thought of [the DDD Investigator] as the centerpiece … to help me get things started … and develop a concept for a campus-wide data science institute and … after a couple of years of work, we now have the beginnings of one. …[The DDD Investigator] was a good example for me to say, "Look, we are bringing in talented people and the awards they are getting—there is so growth potential in this area." So [the DDD Investigator] helped me … to grow the whole area across the campus. [He] was an important part of that. (Administrator, DDD Investigator's institution)*

Two DDD Investigators and one non-awardee in the DDD Investigator competition pointed to the MSDSEs as a catalyst for their institutions' respective data-driven science initiatives: "The [university's] data science initiative feels like it's almost in response to the Moore[-Sloan] centers." Another DDD Investigator elaborated:

> *Looking at timing—[the] DSEs came online and then the next year the next crop of institutions starting to talk about something. Could it have been that Moore and Sloan happened to beat [other institutions to the] leading edge? Possibly. Some folks may have had ideas independently but having the DSEs happen is the sort of thing that can get wheels moving within the administration. Before, they would go to dean and provost, and they would say, "interesting." But with DSEs, that turns into "Let's build this right now. Get to work."*

We also found explicit reference to BIDS and NYU's MSDSEs in a working paper on the structure of an interdisciplinary data science initiative at one university.[109] Finally, respondents across the three MSDSEs reported inquiries from other institutions that wanted to learn about their model for establishing data science environments.

> *We have a steady stream of universities visiting. … So when people come we tell them, that is how we do it and that is how you can do it, too. They use our reports when they write their proposals. We don't know whether they will manage to get there … but they are certainly coming here for advice. … In some parts of the world, they have a lot of money. In Scandinavia and China, these places can build these kinds of things and they are coming to us to ask what they can do. (MSDSE leader)*

[109] http://voyteklab.com/wp-content/uploads/UCSD-DataScience_in_the_SocialSciences2017.pdf

And from another MSDSE:

> *If you think about it, [the Moore and Sloan Foundations] had an amazing vision before other people. Now you see a bunch of other universities who are trying to replicate what we have here, people are reaching out....*

Among the 15 universities invited by the Moore Foundation and the Sloan Foundation to participate in the MSDSE selection process (Exhibit 3.1), some had data science programs with a related or closely aligned mission already underway (e.g., Rensselaer Polytechnic Institute's Data Science Research Center, Columbia University's Data Science Institute). Others have since formed new or significantly expanded data science initiatives (e.g., Brown University, California Institute of Technology; Caltech). Some universities beyond the 15 considered for an MSDSE award have also formed data science initiatives (e.g., the University of Florida, Michigan State University).[110] However, given the early nature of most of these universities' data science initiatives, their ultimate structure and the roles of researchers from different scientific disciplines were not yet fully established. Some of these initiatives lack the MSDSEs' explicit focus on supporting data-driven approaches to fundamental discovery across a wide spectrum of domains in the natural sciences. Some focus on data-driven solutions to problems in applied sciences, such as biomedical or health sciences research (e.g., the University of Chicago's Center for Data-Intensive Science); others focus primarily on mathematics, statistics, or computer science, or on applications of these fields to finance or social sciences (e.g., MIT's Institute for Data, Systems and Society).

Moreover, even at institutions with efforts to create a critical mass of data-driven researchers, not everyone agreed that these initiatives were taking hold. One non-awardee in the DDD Investigator competition lamented that "coordination has not been great," resulting in data science initiatives in three separate academic divisions at her institution. DDD Investigators and administrators from two other institutions saw their universities as focused on building capacity for data-driven science within, but not across, scientific domains. Some of these different perspectives may reflect idiosyncrasies of particular institutions; for example, inter-departmental competition for resources or status may impede some institutions' efforts to build cross-domain collaborations between the methodological fields and scientific domains.

Our interview data also suggest disciplinary variation in levels of enthusiasm for data-driven science. DDD Investigators reported particular enthusiasm in some domains (such as computational biology, bioinformatics, and ecology). Two others (a DDD Investigator and one non-awardee) perceived a tension between those who perceived that statistics had always been data driven versus those who perceived new interest in data-driven inquiry as an opportunity to demonstrate their relevance—or to ensure that the field of statistics would not be "left behind" in any institutional shifts. Nonetheless, the majority of interviewed respondents reported that their universities had active initiatives that they characterized as similar in purpose (if not in scope or structure) to the MSDSEs.

However, we do not have systematic data to attribute newly created data-driven science centers elsewhere to the DDD initiative, either at universities considered for an MSDSE or those that have appeared since elsewhere. Attributing any organizational change at a university to a causal antecedent

[110] Data science degree programs and initiatives exist at many academic institutions, but this discussion excludes those that focus on business analytics or health informatics, instead emphasizing those with a basic science research focus on enabling new discoveries.

is extremely challenging given both the desire for universities to distinguish themselves from competitors in the higher education marketplace and the fact that academic research institutions are complex systems subject to multiple external and internal influences. What we can say is that the establishment of the MSDSEs preceded the emergence of similar initiatives at some academic research institutions (including some of the 15 universities considered during the MSDSE selection process), and not all of the "data science" initiatives at these institutions appear to be focusing as directly on data-driven discoveries in the natural sciences as the MSDSEs.

3.4 Funding for Data-Driven Science

Other signs of increased interest in data-driven science come from new grants or grant programs at the NSF, NIH, philanthropic foundations, and industry research initiatives. DDD investigators, non-awardees, and *Practices* project leads described the following agencies as providing funding for data driven research: NIH, NSF, the Defense Advanced Research Projects Agency (DARPA), the U.S. Air Force Office of Scientific Research, the National Institute of Standards and Technology (NIST), and the U.S. Department of Energy (Exhibit 3.2). A few of the *Practices* leaders mentioned investments from philanthropic organizations other than the Moore Foundation, including the Helmsley Charitable Trust and the Sloan Foundation; two *Practices* leaders as well as a few MSDSE respondents noted industry engagement in academic data-driven research. Below, we briefly summarize the main thrust of programs supporting data-driven science in federal, philanthropic, and industry sectors.

3.4.1 Federal Funding

At the National Institutes of Health, the Big Data to Knowledge (BD2K) program launched in 2013 with a focus on enhancing the utility of biomedical data by

- Setting standards for sharing biomedical data and other digital assets;
- Supporting research and development in software, methods, and tools for biomedical data analyses;
- Enhancing training in the use of these methods and tools; and
- Supporting a data ecosystem that enables discovery.

BD2K established 11 Centers of Excellence for Big Data Computing at institutions across the country. It has a strong emphasis on training, with online and short course offering, webinars, and institutional training grants. Its K01 grants are specifically for early-career scientists developing data-driven tools and methods.[111] Its Training Coordination Center organizes activities across the BD2K Training Consortium and develops software to enable more efficient discovery of educational resources and institutional training grants. BD2K recently supported the development of shared principles for the sharing and management of scientific data (namely, data must be "Findable,

[111] https://datascience.nih.gov/bd2k/faqs/k01

Exhibit 3.1: Selected Academic Research Institutions with Data Science Initiatives

Institution	Initiative	Launch Date	Key Characteristics
University of Illinois at Urbana-Champaign	Illinois Data Science Initiative (iDSI)	2017	▪ Devoted to identifying challenges and opportunities through campus summits focused on core research themes ▪ Working towards articulating a plan for integrating and elevating Data Science on Illinois campuses ▪ Steering Committee includes members from agricultural, consumer and environmental sciences; applied health sciences; biology; business; education; computer science; electrical and computer engineering; history; information sciences; political science; mathematics; statistics; and urban planning ▪ Focus on using data science to serve Illinois community and educational, government, and industry partners
UC San Diego	Halicioglu Institute for Data Science	2017	▪ Will include computer science, cognitive science, mathematics and other (as yet, unspecified) fields ▪ White paper from UCSD's Division of Social Sciences, "Defining the Interdisciplinary Future of Data Science" explicitly cites faculty fellows at BIDS and independent postdoctoral fellowships at NYU's MSDSE
Brown University	Data Science Initiative	2016	▪ Over 2012-2017, researchers from Brown participated in the Intel Science and Technology Center on Big Data (based at Massachusetts Institute of Technology) ▪ Core departments: Biostatistics, Computer Science, Mathematics, Applied Mathematics, with other natural sciences represented: Biomedical Informatics, Computational Biology, Physics, Brain Sciences ▪ Master's degree program
University of Florida	Informatics Institute	2015	▪ Research Opportunity Seed Fund awards available to PI-eligible faculty/research staff to conduct research needed to improve a proposal for external funding. Thematic areas include core data science techniques or informatics and big data analytics in biomedical and life sciences, engineered systems and physical sciences, social sciences, education, humanities, and agriculture ▪ Postdoctoral and graduate student fellowship program to provide joint funding with another department ▪ Annual symposium
University of Michigan	Michigan Institute for Data Science (MIDAS)	2015	▪ Grew out of faculty-led development process over 2012-2014 ▪ Initiative in data-intensive learning, transportation, biomedical, and social science research. Researchers from statistics, biostatistics and mathematics, computer science and engineering, information science ▪ Graduate data science certificate program; workshops in Python, R, GIS, geopandas ▪ Seminar series of invited speakers
Michigan State	Computational Math, Science, Engineering Department (CMSE)	2015	▪ Jointly administered by College of Natural Science and College of Engineering ▪ Includes research in numerical methods and algorithms with applications to physical, biological, and engineering sciences (plasma physics, mathematical modeling, computational biology) ▪ Joint focus on development of data science and scientific computing methods ▪ PhD program; master's degree program projected to enroll first cohort in Fall 2019
MIT	Institute for Data, Systems and Society	2015	▪ Primarily focused on decision sciences, social sciences and applications to finance, energy systems, urbanization, social networks and health ▪ The IDSS's Statistics and Data Science Center is developing new academic programs, from a minor to a PhD in statistics and data science ▪ Hosts a variety of events and brings together researchers of a variety of disciplines

Institution	Initiative	Launch Date	Key Characteristics
Stanford	Stanford Data Science Initiative (SDSI)	2015	▪ Research projects with cross-domain collaborations between computer science, sociology, and linguistics; genetics, medicine, engineering, and statistics; computer science and electrical engineering ▪ Industry sponsors from technology, finance and insurance sectors bring funding and visiting scholars to SDSI ▪ Annual data science workshops, including a 2016 Workshop on Data Science for Biomedicine ▪ Retreats for SDSI partners to learn about each other's research projects ▪ Data Commons project executed by Stanford faculty and staff
California Institute of Technology (Caltech)	Center for Data Driven Discovery (CD3)	2014	▪ Started with a partnership with the Center for Data Science and Technology at the Jet Propulsion Laboratory; that center and CD3 formally merged in June 2015 ▪ Focus on applications of data-driven computation to problems in astronomy, biology, physics, geosciences, and cross-domain transfer of data-driven methods ▪ Research in data visualization, computational plant biology, disaster response
Columbia University	Data Science Institute	2012	▪ Administered by College of Engineering 2012-2017; to become university-wide research center in 2017 ▪ Research centers in foundations of data science (statistics, mathematics, computer science, engineering), cybersecurity, health, smart cities, materials science ▪ Data Science Interdisciplinary ROADS Provost Ignition grants ▪ Moore-Sloan funded Interface of the Natural Sciences and Data Sciences grants: partnerships of two pairs of faculty-doctoral student teams, one each from natural sciences and data sciences ▪ Industry affiliates program to identify productive collaborations with companies ▪ "Institute Industry Innovation" seminar series
Rensselaer Polytechnic Institute	Data Science Research Center	2010	▪ Facilitates collaborations among researchers in computer science, biology, engineering, mathematics, physics, environmental science, library, and social sciences ▪ Methodological researchers focus on core problems in data acquisition and storage; data complexity; modeling and knowledge extraction; simulation and visualization; security and privacy ▪ Partnerships with industrial laboratories ▪ Linked to the Institute for Data Exploration and Applications: http://idea.rpi.edu/
Johns Hopkins	Institute for Data Intensive Engineering and Science (IDIES)	2008/2013	▪ 2013 expansion to university-wide initiative. Arts & Sciences, Engineering, Libraries, Schools of Medicine and Public Health ▪ Fosters education and research in the development and application of data intensive technologies in physical and biological sciences and engineering ▪ Provides seed funding to JHU faculty interested in data-intensive computing projects
	JHU Data Science Lab (JHUDSL)	2012	▪ Initiated by faculty in biostatistics, now includes collaborations with Computer Science, Biology, Biomedical Engineering, Medicine, and The Center for Teaching and Learning ▪ Includes software developers, outside collaborators, and a few students ▪ Creates open-source online courses on various topics and platforms

Institution	Initiative	Launch Date	Key Characteristics
University of Chicago	Center for Data Intensive Science	2014	▪ The Center focuses applications of big data in biology, genomics, biomedicine, and health care ▪ Develops algorithms, statistical models, software and infrastructure; to combine cloud computing technology with large scale data commons to support scientific research ▪ Hosts and helps operate: the Open Science Data Cloud, the Genomic Data Commons, the Bionimbus Protected Data Cloud, the Environmental Data Commons
	Computation Institute	2000	▪ Established as a joint initiative with Argonne National Laboratory ▪ Over 100 researchers and staff ▪ Projects and research in bioinformatics, biomedicine, neuroscience, genomics, metagenomics, energy and climate, astronomy and astrophysics, computational economics, molecular engineering

Sources: University websites and news archives:
http://idsi.illinois.edu/
https://illinois.edu/calendar/detail/7?eventId=33268888
http://triton.news/2017/06/3676/; http://www.newswise.com/articles/alumnus-taner-halicioglu-kicks-off-campaign-for-uc-san-diego-with-75-million-gift
http://voyteklab.com/wp-content/uploads/UCSD-DataScience_in_the_SocialSciences2017.pdf
https://news.brown.edu/articles/2016/10/dsi
https://informatics.institute.ufl.edu/about-us/michailidis/
http://midas.umich.edu/about/
https://cmse.msu.edu/news-events/news/general-news/msu-department-of-computational-mathematics-science-and-engineering-to-hold-inaugural-workshop/
https://stat.mit.edu/, https://stat.mit.edu/about/
https://sdsi.stanford.edu/
http://www.caltech.edu/news/caltech-jpl-team-take-big-data-projects-47037
http://engineering.columbia.edu/news/data-science-university-wide-institute
http://www.dsrc.rpi.edu, https://news.rpi.edu/content/2013/09/06/harnessing-petabyte-data-science-research-center-explores-cloud-computing-and/
http://idies.jhu.edu/
http://president.jhu.edu/meet-president-daniels/speeches-articles-and-media/institute-for-data-intensive-engineering-and-science/
https://cdis.uchicago.edu/
https://www.ci.uchicago.edu/about/mission

Accessible, Interoperable, and Reusable"), calling them the FAIR Guiding Principles.[112] It is initiating a pilot phase of an NIH data commons. One DDD investigator is a co-PI on a BD2K award.

The National Science Foundation likewise launched a multi-component initiative to support environments enabling data-driven discovery and funding primary research. Programs include: the Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) program; the multi-institutional Big Data Regional Innovation Hubs (BD Hubs); and the Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes).

BIGDATA funds researchers investigating fundamental theories and methods motivated by big data challenges with broad applicability across domains and collaborations between methodologists and domain scientists seeking innovative applications of new methods to solve specific problems in one or more science domains (one DDD Investigator and an MSDSE co-PI received a 2013 BIGDATA award).[113,114]

Exhibit 3.2: Federal Funding Initiatives in Data-Driven Science

Agency	Program(s)
National Institutes of Health (NIH)	▪ Big Data to Knowledge (BD2K)
National Science Foundation (NSF)	▪ Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) ▪ Big Data Regional Innovation Hubs (BD Hubs) ▪ Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes)
National Institute of Standards and Technology (NIST)	▪ Data Science Research Program ▪ Data Science Evaluation Plan ▪ NIST Big Data Public Working Group
Defense Advanced Research Projects Agency (DARPA)	▪ Information Innovation Office (I2O): Data-Driven Discoveries of Models (D3M)
U.S. Department of Energy (DOE)	▪ Advanced Scientific Computing Research

The BD Hubs program was designed to facilitate public-private partnerships. Notably, the grant solicitation explicitly mentioned the desire to build on the momentum of the Data to Knowledge to Action event—at which the Moore-Sloan Data Science Environments was one of several big data collaborations announced. With the BD Hubs program, NSF established a regional network of consortia that include academic institutions, industry, non-profit, foundation, and state and local government partners to stimulate regional partnerships to exploit big data to solve societal problems and advance scientific discovery.

NSF does not fund Hub research directly; rather it provides the means (staffing, networking activities) to coordinate multi-stakeholder engagements organized into regional themes. For example, the West Big Data Innovation Hub's themes include urban data science (smart cities, transportation, housing),

---

[112] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3,* 160018. https://www.nature.com/articles/sdata201618

[113] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

[114] NSF award # 1251274 to J. Bloom & F. Perez from the University of California –Berkeley

precision medicine, natural resources management and disaster response, big data technology (e.g., cloud computing, storage, visualization), and data-enabled scientific discovery and learning. The BD Spokes program extends the regional Hubs by funding partnerships aimed at supporting activities in one of their theme areas.

The NIH and NSF programs clearly intersect with the DDD initiative. They have shared goals—to stimulate the development of methods, tools, and technologies motivated by the challenges of big data and to promote productive collaborations between data methodologists and domain scientists. They also share grantees. For example, two MSDSE participants in UW's eScience Institute (Bill Howe, Magdalena Balazinska) are co-PIs on a BIGDATA grant that has contributed to publications with DDD investigator Jeff Heer[115] and with MSDSE data science fellows Jake VanderPlas and Ariel Rokem; one publication compared the performance of several big data systems (Dask, Myria, SciDB, Spark, and TensorFlow) for scientific image processing.[116]

In addition, UCB and UW were selected by NSF, along with the San Diego Supercomputer Center at the University of California-San Diego, to operate the West Big Data Innovation Hub; several of the co-PIs are BIDS and eScience Institute personnel.[117] Project Jupyter is one of the key collaborators of a BD Spokes project funded by the Hub. That project—Network for Computational Modeling in Social and Ecological Sciences (CoMSES Net)—focuses on enhancing the usability of the Computational Model Library to enable open and reproducible scientific computation.

In April 2017, NSF announced a new initiative—Harnessing the Data Revolution for 21st Century Science and Engineering—as one of four opportunities to explore "convergence," described by the National Research Council as an "expanded form of interdisciplinary research" characterized by sustained interaction of disciplines that results in a shared set of concepts, methods, goals, even a shared language, that is applied to solve a complex problem.[118] Indeed, BIGDATA postponed its deadline for 2017 proposals to incorporate proposals focusing specifically on Harnessing the Data Revolution for 21st Century Science and Engineering (HDR) workshops.

3.4.2 Foundation Funding

During the course of our evaluation, interview respondents also mentioned the importance of foundation funding that supports data-driven science and organizations affiliated with promoting more reproducible research practices. In addition to the Moore Foundation and the Sloan Foundation, other notable philanthropic players include the Helmsley Charitable Trust and the Simons Foundation (Exhibit 3.3). Several respondents at MSDSE institutions mentioned prior or concurrent support from one or more of these foundations. They have also supported other organizations working to make research data and computational code more accessible and reproducible, through either development

---

115 Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., & Heer, J. (2016). Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics, 22* (1), 649–658. http://ieeexplore.ieee.org/document/7192728/

116 Mehta et al. (2016).

117 See: http://westbigdatahub.org/people/.

118 National Research Council. (2014). *Convergence: Facilitating transdisciplinary integration of life sciences, physical sciences, engineering, and beyond.* Washington, DC: The National Academies Press. doi:10.17226/18722

of tools (rOpenSci) or practices to promote open sharing of research products (e.g., Make Data Count, FORCE11). Some foundations have also established research centers with strong engagement in data-driven research and inquiry; examples include The Allen Institute for Brain Sciences and the Simons Foundation's Flatiron Institute.

Exhibit 3.3: Key Foundation Investments in Data-Driven Research or Science-Enabling Tools and Resources

Foundation / Initiative	Selected Grantees
Helmsley Charitable Trust: Biomedical Research Infrastructure program*	▪ Project Jupyter ▪ rOpenSci ▪ Center for Scientific Integrity ▪ FORCE11: Mapping the Scholarly Landscape ▪ Mozilla Science Foundation: Building Capacity for Open Source Research
Simons Foundation: Flatiron Institute (includes: Simons Center for Data Analysis (SCDA) Center for Computational Astrophysics Center for Computational Biology Scientific Computing Core)	▪ arXiv ▪ MAGMA ▪ Simons Institute for the Theory of Computing (UC Berkeley)
Alfred P. Sloan Foundation: Data and Computational Research Scholarly Communication Sloan Digital Sky Survey	▪ arXiv ▪ FORCE11 ▪ Make Data Count ▪ NumFOCUS ▪ Project Jupyter ▪ Moore-Sloan Data Science Environments (MSDSEs)

* Note that this initiative closed down as of 2016

3.4.3 Industry Funding

Investment in data-driven science from federal and private funding sources has demonstrably increased over the past five years. Perhaps not surprisingly, information about the levels and targets of industry funding is quite limited. Most DDD stakeholders were aware of industry *interest* in data-driven skills: DDD investigators, non-awardees, and administrators all bemoaned having to compete for postdocs and other research staff with companies such as Google, Amazon, Facebook, and Microsoft. However, few mentioned industry as a primary funding source.

From an institutional perspective, however, there are industry partnerships in data science or related initiatives at academic research institutions (e.g., Amazon Web Services, Google, Hitachi, IBM, Intel, Microsoft Research, and VMWare). Intel's Science and Technology Center program, for example, has supported work by two DDD participants at UW, including one DDD investigator (Jeff Heer) and one eScience Institute leader (Bill Howe), both of whom participated in a 2012-2017 research working group on big data visualization through this center. NVidia is a founding partner of the Center for Data Science at NYU, and BIDS has industry partnerships (e.g., Seimens, State Street) that have generated approximately $350,000 used to support fellows, selected events and travel. Google has made some research awards to academia, including a Google Faculty Research Award to a faculty member at NYU's Center for Data Science. A DDD investigator was also a past recipient, with colleagues, of an unsolicited Focused Research Award from Google. These few examples may not reflect the true extent of industry support for data-driven research in academia, but interviews and search of extant data revealed little in the way of DDD participants' interactions with industry.

The chief source of funding for academic research in the life and physical sciences, and in the computational, statistical, and mathematical sciences, is federal agencies, with NSF and NIH playing lead roles. These agencies' investments in data-driven science appear to have lagged that of the Moore Foundation, the Sloan Foundation, and a handful of other foundations. At least with respect to the MSDSEs, the Moore Foundation and Sloan were at the forefront of funders with early momentum.

NSF's and NIH's big data share some surface-level similarities to the DDD initiative, but they do not appear to align entirely with the DDD initiative's goals. In terms of similarities, NIH's BD2K program includes both single-investigator awards (e.g., BD2K Career Development K01, K22 awards) and institutional and multi-institutional training grants (NIH's BD2K Centers of Excellence for Big Data Computing and Training Coordinating Centers BD2K Training Coordination Centers, and Training Grants), just as the DDD initiative includes single-investigator and institutional-level grants. Similarly, NSF big data programs include research awards for collaborations between domain-based scientists and methodologists (NSF's BIGDATA Innovative Applications award track) as well as larger, multi-institutional BigData Hubs to foster partnerships. These programs are also supporting some of the same themes that permeate the DDD initiative, such as an emphasis on developing methods and tools for application to domain-specific research challenges, as well as advancing foundational computational and statistical methods (e.g., NSF's TRIPODS program).

Yet, there are key differences between these federal funding programs and the DDD initiative. Chief among these differences is the MSDSEs' explicit goal to prompt cultural changes in the way academic data-driven scientists are evaluated and rewarded. Moreover, the *Practices* strategy appears to have no parallel, to date, with other federal big data funding programs, as grantees themselves attested, and as a scan of the available federal programs affirms: the *Practices* strategy's explicit funding for projects and organizations that develop tools and resources for general application and adaptation by *other* scientists (i.e., not by the *Practices* grantee organizations themselves) is unique.

3.5 Open Science and Reproducibility

Across its funding strategies, the DDD initiative purposefully promotes transparent and reproducible scientific practices, especially critical for researchers working with large and complex data, and supports the development of tools and resources to enable adoption of these practices. Participants in the DDD initiative have made important contributions to open science and reproducibility.

Although the movement toward more transparent practices reflects concern that much of the published scientific literature was largely inaccessible to that public,[119] another motivating factor stems from highly visible retractions of high-profile articles found to have errors in analysis, as well as alarm over apparent non-replicability of published and widely accepted findings.

> *We're talking … about how to … support open science and reproducible research in a way that changes the publication models, by allowing scientists... to see the data behind things. Right now there's very active debates across multiple scientific disciplines about retractions and results that are not reproducible … there's very*

[119] Accessibility to scientists was also limited by delays related to the peer-review and publication process, topics not addressed here.

> *heated debates about … the way the statistical analyses are done. I'm not saying those issues are magically settled if you have access to the code, data, … but I certainly think it would be useful for those debates if the code and data and tools were available. (MSDSE respondent)*

Greater appreciation for the benefits of open source software has also played a role in adoption of reproducible practices. One interviewee, a user and contributor to a *Practices* project, described a shift in her research field:

> *In the past, with a lot of the code being proprietary, there's a lot of heuristics and things like that that are built into those codes, a lot of ... parameters and those types of things. Opening those up, showing where they come into play, and also being able to talk about them in an interactive environment has promoted a lot of positive conversations in terms of reproducibility. (*Practices *project user)*

Increased data sharing and use of open source software has become intimately intertwined with open access to publications. As a non-awardee pointed out, these trends have begun to shift scientific norms:

> *Open source software and open datasets has just fired up the world. … Journals that want to associate papers with code and data are winning, and so the journals and culture of the profession is changing.*

Grantees across the three DDD strategies have engaged actively in open science and reproducibility practices, contributing to a shift in the cultural norms for scientific dissemination. DDD investigators routinely post data and software to accompany their publications; others disseminate new research-enabling tools they have developed, or training and course materials for others to borrow and adapt. DDD grantees have also published articles in scholarly journals or conference proceedings where they advocate for sharing of software and adoption of standards and best practices for open science.[120,121,122] One investigator has advocated for examination of data produced by other researchers by creating an annual "research parasitism" award for the best example of this type of scholarship.[123]

Each of the four *Practices* grantee organizations assessed by the evaluation is actively engaged in efforts to foster reproducibility. Specifically, Dask, Julia, Jupyter, and Numba all provide free and

---

120 DDD investigators Ethan White and Jeffrey Heer: Mislan, K. A. S., Heer, J. M., & White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology and Evolution, 31* (1), 4–7. https://doi.org/10.1016/j.tree.2015.11.006

121 DDD investigator C. Titus Brown and BIDS data science fellow Karthik Ram: McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A. Lin, J. … Yarkon, T. (2016). Point of view: How open science helps researchers succeed. *eLife, 5,* e16800.

122 eScience Institute data science fellow Ben Marwick: Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P. … Poline, J. B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience, 20,* 770–773.

123 Greene, C. S., Garmire, L. X., Gilbert, J. A., Ritchie, M. D., & Hunger, L. E. (2017). Celebrating parasites. *Nature Genetics, 49,* 483–484. https://doi.org/10.1038/ng.3830

open source software, packages, and applications. Data Carpentry posts its curricular materials online, for anyone to use—even though its workshop fees represent a key revenue source. Among investigators and non-awardees, Jupyter is both widely known and often valued for its contributions to reproducibility. Data Carpentry's founders emphasized that among their objectives was to "enable [researchers] to retrieve, view, manipulate, analyze, and store their and other's (sic) data in an open and reproducible way in order to extract knowledge from data."[124] One graduate student credited her interest in reproducible research as a factor that influenced her decision to become an instructor, noting that she is

> *… able to better collaborate now with people because I can work more reproducibly. I know how to write code that other people can understand, and I know how to document things well. So that's been a big help, especially as I moved through my PhD and now … into my postdoc, there's more and more collaboration, so that's been a really helpful skillset.*

The MSDSEs also have active commitments to fostering open science and reproducibility. A current senior data fellow at the eScience Institute and a former data science fellow from BIDS both serve on the editorial board of the *Journal of Open Source Software*, a new journal that accepts short papers describing software that the authors have already released and documented (thereby enabling citation of the software). Starting with a workshop hosted by BIDS, members of all three MSDSEs collaborated to publish a book of case studies of reproducible research in data-driven science.[125] The BIDS Reproducibility and Open Science Working Group also created a graduate course in reproducible and collaborative statistical data science, which is a core requirement for the Data Science for the 21st Century research training grant. NYU's MSDSE faculty have also described plans to develop a required course module on reproducibility to for postdoctoral and graduate student researchers. A senior leader at one MSDSE described the enthusiasm for reproducibility:

> *We have a critical mass, several faculty members … who are interested in reproducibility. We have a research scientist who is working exclusively on software for reproducibility. We have a joint hire … who has been amazing in evangelizing the whole university about open science and reproducibility. She made a huge impact. We asked our fellows and postdocs to write a report in the end of the year, and most of their papers are reproducible— they have the code and the data.*

The eScience Institute has worked with the Center for Open Science[126] to grant "badges" to researchers who make their code and data available not just for journal articles, but also for posters or conference presentations. A faculty leader at NYU's MSDSE served on a committee of the Association for Computing Machinery (ACM) to establish a system of badges to mark papers for which code, data and other artifacts are available (enabling examination by others), have been

---

[124] Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data Carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation, 10*, 135–143. https://doi.org/10.2218/ijdc.v10i1.351

[125] Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2017). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences.* Retrieved from http://www.practicereproducibleresearch.org

[126] See: https://cos.io/

evaluated by reviewers, or where the key results have been validated independently of author-supplied artifacts.[127]

Recent developments suggest that important stakeholders throughout the scientific ecosystem are engaging productively (perhaps cautiously, in some instances) to further the adoption of these principles and practices. Two related developments include:

- Adoption of incentives to encourage sharing, such as the Center for Open Science's badges, and *Science* and *Nature* journals' endorsement of the Transparency and Openness Promotion (TOP) guidelines, which establish a common standard for assessing journal practices and policies surrounding the citation of data or code, sharing of data and code, and other transparent research practices;[128,129] and
- Movements to establish common standards for scientific data management and data citation practices, such as the adoption by NIH of the FAIR standards for data sharing and the May 2017 launch of the Make Data Count project (with support from NSF and the Sloan Foundation).

The importance of these changes cannot be understated, because practicing reproducible research supports a normative shift that increasingly recognizes the contributions of those who share data and software. That is a necessary precondition for another important goal of the DDD initiative, namely shifting the reward structures in academia to place greater value on research products and practices; as such, it represents another avenue through which the DDD initiative is tackling changes in the broader academic research landscape.

3.6 Academic Careers in Data-Driven Science

One key distinction between the DDD initiative and other efforts to support data-driven inquiry is its deliberate spotlight on research products that enable new discoveries and on the establishment of sustainable alternatives to the tenure track for data-driven researchers in academia. Retention of talented data-driven scientists in academia faces two related challenges. One concern, discussed in Chapter Two, is that current promotion and tenure criteria at major research universities cannot adequately capture data-driven scientists' contributions in general—regardless of whether they pursue traditional tenure-track faculty positions. The second concern is about the longer-term career prospects for research staff who provide the labor —and much of the intellectual capital— that support scientific research in academic institutions. (Criteria for assessing contributions to research by those staff are no less important, but first they need a pathway in which such criteria would apply.) This section examines these two concerns and how the DDD initiative is tackling changes in the status quo.

3.6.1 Metrics for the Research Contributions of Data-Driven Scientists

The DDD initiative is demonstrably committed to expanding the career pathways for data-driven scientists; it does so by supporting efforts to recognize the value of data-driven scientists' research

[127] See: http://www.acm.org/publications/policies/artifact-review-badging

[128] Announcement: Transparency upgrade for Nature journals [Editorial]. (2017, March 16). *Nature, 543*, 288.

[129] McNutt, M. (2016). Taking up TOP [Editorial]. *Science*, 352, 1147.

contributions and to broaden the metrics by which academic faculty achieve important career milestones. Both of these approaches center on the importance of data and software in scientific research. Typically, the current promotion and tenure process includes candidates' submission of promotion/tenure portfolios about their research achievements, satisfactory teaching and successful production of new doctorates, service to the institution, and evidence, often provided by letters from reviewers outside the institution, of impact on the field. The most problematic of these criteria for data-driven scientists centers on the assessment of research achievements and impact on their respective field(s).

Some administrators, both at DDD Investigators' and MSDSE institutions, argued that this process was sufficiently flexible to give due weight to the tools, methods, and practices developed and shared by data-driven researchers. The lack of shared norms or formal standards for counting (let alone assessing quality thereof) products such as software, packages, visualization tools, and the like, however, makes these researchers vulnerable to department chairs, provosts, and external reviewers who may not appreciate the value of these products either to the candidate's own research portfolio or more generally.

Nonetheless, there are some indications of changes in how such data scientists' contributions are perceived. One neurobiologist described a shift in his field:

> *I would say the field as a whole is getting more professional in their development strategies and more aware of the value of writing and sharing reusable code. There's more reward out there for people who engage in these kinds of activities and invest their time in tools like Julia and other tools out there, and I think there's more awareness of the importance of numerical and data-driven reproducibility and posting your datasets online and documenting the steps to reproducing your research.*

This perception is shared, for example, by the Future of Research Communications and e-Scholarship (FORCE11), a community of researchers, funders, publishers, and librarians, which believes that the research paper is no longer sufficient as the appropriate unit of scholarly publication. Instead, FORCE11 contends that the modern unit of knowledge is a research object set containing datasets, workflow, software, mathematical models, *and* papers that result from an investigation. As a result, assessing research productivity accurately demands new metrics.[130]

Since its 2011 manifesto, FORCE11 has had some notable accomplishments, including endorsement of its 2014 Joint Declaration of Data Citation Principles by more than 100 organizations, including 40 data repositories, 28 professional associations, 23 publishers (e.g., Elsevier, Nature Publishing Group), and more than 250 individuals at top research universities worldwide (e.g., Columbia, Harvard, NYU, Massachusetts Institute of Technology, Rensselaer Polytechnic Institute, UCB,

[130] Bourne, P., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E., & Shotton, D. (Eds.), on behalf of the FORCE11 community. (2011). *FORCE11 white paper: Improving the future of research communications and e-scholarship.* Retrieved from https://www.force11.org/about/manifesto

University of Michigan, UW, etc.).[131,132] (Signatories include Kyle Cranmer, of NYU's Center for Data Science, and DDD investigator Ethan White; FORCE11 is supported by the Moore Foundation/DDD and the Sloan Foundation, Elsevier, PLoS, and The Network Institute.)

In 2016, a FORCE11 working group issued software citation principles, motivated in part by the need for academic researchers to get credit for products they develop, "particularly when those products enable or further research done by others."[133] A related publication is "the Dagstuhl Manifesto" on engineering academic software, which describes problems and proposes specific actions to ensure that software is cited properly and to promote viable career pathways for research software engineers;[134] The participants in the workshop leading to this manifesto included a former BIDS data science fellow, a DDD investigator, and an eScience Institute data science fellow.

Major funding organizations, including NSF, NIH, the Wellcome Trust, and the Sloan Foundation have also sponsored workshops to foster discussions of credit and citation of software.[135] Two other influential groups in this realm are the Software Sustainability Institute and the Working Towards Sustainable Software for Science: Practices and Experiences (WSSSPE),[136] both based in the United Kingdom.

A primary (though not the sole) motivation behind efforts to promote standards for citing data and software is ensuring that their creation can be appropriately quantified—and ultimately, acknowledged and rewarded. Without some mechanism for assigning credit for contributions to software, data, and other research products, the value of such work is obscured. This in turn provides little incentive to disseminate these products (i.e., so that other researchers might benefit from them). The incentive structure for investing the time and the opportunity costs of research software and other science-enabling tools also become misaligned with the advancement pathways of individual researchers, the topic to which we turn last.

3.6.2 Academic Career Paths for Data-Driven Researchers

An underlying motivation of the DDD initiative is to transform the employment landscape for individuals who bring data-driven approaches to scientific inquiry. The DDD initiative provides an opportunity both to individual DDD investigators and to the three MSDSE institutions to experiment

---

131 Data Citation Synthesis Group. (2014). Joint declaration of Data Citation Principles–Final. M. Martone, Ed. San Diego CA: FORCE11.

132 See: Endorse the Data Citation Principles [Website], https://www.force11.org/datacitation/endorsements. Retrieved June 16, 2017.

133 Smith, A. M., Katz, D. S., Niemeyer, K. E., & FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science 2*:e86. https://doi.org/10.7717/peerj-cs.86

134 Allen, A., Aragon, C., Becker, C., Carver, J., Chiș, A., Combemale, B. … Vinju, J. J. (2017). Engineering academic software: Manifesto from Dagstuhl Perspectives Workshop 16252. Retrieved from http://drops.dagstuhl.de/opus/volltexte/2017/7146/pdf/dagman-v006-i001-p001-16252.pdf

135 For example: National Science Foundation Advisory Committee for Cyberinfrastructure, Task Force on Software for Science and Engineering. (2011). *Final report, March 2011*. Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_Software.pdf

136 The DDD initiative has provided support to the WSSSPE.

with alternative types of *positions*. It remains to be seen if these positions turn into alternative *career pathways*, however.

The three MSDSEs have created both tenure-track faculty positions and some version of a non-tenure-track research scientist position. All three hired postdoctoral fellows jointly supported by an academic department and two of them made a number of tenure-track joint hires with academic departments. A small number of joint-hire faculty or research scientists have received (or been hired with) tenure. It is simply too early to know about the longer-term career trajectories of most early-career individuals currently (or recently) employed at the MSDSEs.

Likewise, for individuals working in DDD Investigators' research groups, it is too early to observe career outcomes. Some DDD Investigators used their funding to hire software engineers or computational methodologists whose positions will likely depend on subsequent (post DDD awards) grant funding. And while at least two DDD Investigators observed that federal agencies have historically been less supportive of such positions, FORCE11 and similar groups may succeed in persuading federal agencies to change practices and policies about whether such positions are indeed supported.

The majority of postdoctoral researchers and graduate students working with the DDD Investigators aspired to an academic research position, and most of these to a tenure-track faculty position. However, the tenure-track route is not universally appealing, as some DDD Investigators' research group members were seeking alternative academic roles. The BIDS' MSDSE-wide career path survey provided additional evidence that alternative academic pathways were appealing to some data-driven researchers.

Several MSDSE respondents expressed concern about whether non-tenure-track positions would prove sustainable in the long term, however. It is possible that the MSDSEs have demonstrated enough success— vis-à-vis joint hires, and seeding cross-domain collaborations, for example—that those three institutions will continue to invest resources, and conduct additional fundraising, to support a staffing model friendly to data-driven scientists. At UCB, for example, the new Division of Data Sciences bears watching, as it represents a significant institutional commitment. The long-term outlook for UW's eScience Institute, beyond the Moore-Sloan and Washington Research Foundation funding, may well depend upon Washington State's budgetary health. NYU may use at least a portion of the revenue from its data science master's degree program to sustain the research arm of the CDS that resulted from the MSDSE.

However, due to a less than optimal environment for academic scientists and engineers in general, creating career paths for data-driven scientists is challenging. The past three to four decades have witnessed an overall decline in the percentage of science and engineering doctorates employed in academia, and increasing reliance on adjunct and other term-limited or contingent positions.

Moreover, between 2010 and 2014, federal funding for research and development dropped by 11 percent (measured in 2016 dollars; the drop was 17 percent when accounting for inflation).[137, 138]

Looking more broadly, and not solely at the three MSDSE institutions, the growing number of data-driven science institutes may signal potential for new career options inconceivable a few years ago. The large investments by NSF in regional big data hubs, and by NIH in its BD2K coordination centers, also point toward a transformed landscape for data-driven science. Another signal comes from Steven Hyman, of the Broad Institute, former provost of Harvard University, and former director of the National Institute of Mental Health; he argues that science needs both faculty and allied research lines:

> *The staff-scientist model is a win for all involved. Complex scientific projects advance more surely and swiftly …. This model empowers non-faculty scientists to make independent, creative contributions, such as pioneering new algorithms or advancing technologies. … A scientific organization should be moulded to the needs of science, rather than constrained by organizational traditions.*[139]

Hyman champions the staff-scientist model, particularly in the context of research institutes such as the Broad Institute or Allen Institute for Brain Science, where the scope of projects benefits from faculty and non-faculty scientists collaborating to solve common challenges. Similarly, a professional association of research software engineers in the United Kingdom points to parallel research institutes there (the Alan Turing Institute, the Francis Crick Institute) as places that have successfully established career paths for software engineers.[140]

We turn next to exploring the sustainability of the multiple interrelated efforts funded through the DDD initiative.

---

137 National Science Foundation. (2016). *Science and engineering indicators 2016.* Arlington, VA: Author.

138 National Research Council, National Academy of Engineering, & Institute of Medicine. (2014). *The arc of the academic research career: Issues and implications for U.S. science and engineering leadership: Summary of a workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18627

139 Hyman, S. (2017). Biology needs more staff scientists. *Nature, 545*, 283–284. https://doi.org/10.1038/545283a

140 See: Brett, A., Croucher, M., Haines, R., Hettrick, S., Hetherington, J., Stillwell, M., & Wyatt, C. (2017). *Research software engineers: State of the nation report 2017.* https://doi.org/10.5281/zenodo.495360

4. Sustainability, Remaining Challenges, and Potential Opportunities

4.1 Key Findings

As the DDD initiative moves toward the end of its initial phase, several questions arise about how to sustain positive outcomes, what challenges or unmet needs remain, and what potential opportunities have emerged. Key findings about the sustainability of key successes, and unmet needs include the following:

- Sustaining some of the key successes of the DDD initiative will likely require additional external funding after the DDD grant period concludes, in particular, funding for DDD Investigators and MSDSEs to support research software engineers, research scientists and data science fellows.
- All three MSDSE host institutions signaled enduring commitment to the data science environments, but respondents also raised concerns about continuity of funding.
- Formal career pathways in academia for research software engineers may present a potential test case for the viability of alternative career paths for data-driven researchers.
- Survey respondents representing 30 academic research institutions perceived multiple unmet needs for data-driven research at their institutions, including space to meet with colleagues from multiple domains; access to other data-driven faculty and to data scientists and software engineers; and educational initiatives to build capacity of students to contribute to data-driven research.

Potential opportunities for advancing the goals of the DDD initiative include:

- Implementing an institutional level "Challenges in Data-Driven Science" program to unite data-driven domain scientists and computational methodologists at non-MSDSE universities around a shared problem that they propose. If feasible, such a program could present an opportunity to demonstrate the value of data-driven science at institutions without the impetus or resources to establish an MSDSE-like data science environment.
- Exploring ways to further engage academic research libraries and/or research computing in data-driven research.
- Supporting small-scale, cross-domain and cross-institutional community-building events for data-driven investigators or early career scientists to network and learn from each other.

4.2 Sustaining the DDD Initiative's Successes

4.2.1 Sustaining Key Positions

The DDD initiative demonstrated the value of data-driven researchers both at DDD Investigators' institutions and at the MSDSE institutions; yet, both **individual investigators and data science environments will likely continue to require external funding after the end of the DDD grant period to sustain productive programs of data-driven research and opportunities for training and cross-disciplinary collaboration.** About a third of the DDD Investigators hired software

engineers and/or research scientists above the level of a postdoc; all three MSDSEs hired research software and computational methodologists; and at least three of the four *Practices* grantees included in the mid-term evaluation hired research software engineers.[141] DDD Investigators will need to fund positions that added data-intensive expertise to their research teams, and MSDSEs will likewise need to support a critical mass of researchers with such expertise, as well as administrative staff necessary to run the programs.

Like most faculty at top tier academic research institutions, the DDD Investigators will need to compete successfully for grants to continue their research programs. Given the prestige of the DDD Investigator award and their record of publications and grant funding, they are well-positioned to compete for these grants. The risk, as about a third of DDD Investigators (and half of the non-awardees in the DDD Investigator competition) indicated, comes from the relative scarcity of funding for research that bridges disciplines. As one DDD Investigator said:

> *I am not convinced there will always be an award like the Moore award that so nicely straddles communities, which means I am going to have to go back to disciplinary funding. So how do I make sure I'm doing the most exciting work I can do...so that when I have to apply for disciplinary funding, which is again silo-ed, that I'm in a good, strong position....I hope other funding agencies will broaden their portfolio ...for those of us who are really working at these kinds of intersections.*

Not only do these researchers face stiff competition for limited federal research dollars, but they also face the constraints of limited funding for interdisciplinary research, continued competition for software engineers or computational methodologists from industry, and the fact that grant funding the contingent nature of grant funding mitigates against longer-term job security for their research staff.

Although the salary differential between industry and academia presents hiring challenges, the relative saliency of salary compared to other factors varies among doctorate level scientists. Other factors affecting preferences for academic versus industry jobs include freedom to choose research projects; opportunities to publish and attend professional conferences; ability to collaborate with other institutions or organizations; peer recognition; level of responsibility; and access to resources, especially innovative technologies. [142] Consequently, DDD Investigators may want to focus less on *competing* with industry on salary and instead explore other potentially appealing features of academia, including giving these research staff some autonomy to pursue their own research interests. (For example, the DDD Investigator awardees might negotiate with their institution to certify a desired candidate as eligible to apply for their own grant funding as a principal investigator.) Another possibility is to identify potential university-industry partnerships. Some companies offer scientists some of the perquisites traditionally open primarily to academic researchers, such as publishing opportunities, conference attendance, and in some cases, direct collaboration with university scientists. For example, JuypterLab has benefitted from contributions by software developers at

---

[141] Arguably, the DDD grant allowed the Data Carpentry to hire an individual with data-driven expertise in bioinformatics as its Executive Director.

[142] Roach, M. & Sauermann, H. (2010). A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry. *Research Policy, 39*, 422-434.

Bloomberg and Continuum Analytics, who are paid by their employer and not by Project Jupyter's grant funding.

At MSDSE institutions, despite concerns about funding, there were clear signals pointing to continued institutional commitment to the data science environments after the end of the DDD grant period. Interviewees voiced concerns about the availability of post-DDD funding to sustain a critical mass of fellows, research scientists and other personnel at the data science environments (DSEs) and apprehension about possible changes to the structure of the MSDSEs. When the DDD award period ends, it is unclear whether the DSEs will continue to operate as independent entities outside traditional academic boundaries. If there is sufficient consensus that a physically distinct space and shared, inter-departmental or provost-level governance of the DSE are necessary to reap its benefits (e.g., training opportunities, research collaborations, incubator projects), then we can expect the search for new funding sources to be vigorous.

At UCB, the role of BIDS in relation to the university's new Division of Data Sciences is yet to be determined, although administrators point to BIDS as a key catalyst for this new division. When contemplating the future sustainability of BIDS, its leaders prioritized three components for preservation:

1) The dedicated physical space for BIDS, viewed as an important neutral territory that has been critical to bridging cross-departmental boundaries;

2) The critical mass of research scientists and data science fellows who have contributed software development and computational expertise to scientific research challenges and training opportunities for students; and

3) Project Jupyter, to which BIDS contributed working space and a robust intellectual community of like-minded colleagues, and from which BIDS benefited; these benefits have included opportunities for data science fellows and postdocs to contribute to Jupyter's suite of tools and extensions, and a critical computing backbone for data science education at the university.

The cross-institutional excitement for undergraduate data science education, and the new Division of Data Sciences could bode well for the sustainability of BIDS, especially if BIDS is viewed as serving a necessary, distinct role in the institution's data science landscape. If faculty and administrators view BIDS and the new Division as mutually complementary, BIDS will likely endure as a separate entity. If BIDS is seen as duplicating offerings available elsewhere, then it may be absorbed into the new Division, re-organized, or dissolved. Still, data from interviews suggest that this MSDSE has demonstrated its value at UCB as an incubator for a range of collaborations and a wellspring of training opportunities, both formal and informal.

At UW, a few respondents noted uncertainty about the funding required to sustain the positive outcomes of the MSDSE award for the eScience Institute:

> *There is a nice model here where we need a lot of external money, and a lot of it to show that it works. And now that we have shown that it works, it should be absorbed into [the university] and a case ... made to keep this. We are in the process of this*

conversation, and people agree a lot, but will that translate to opening their pocket books? It remains to be seen. (MSDSE leader)

Nevertheless, evidence points to a strong university commitment to sustaining the MSDSE-enabled enrichment of the eScience Institute. Not only did the eScience Institute have a pre-MSDSE corps of committed faculty, but also the provost displayed support for the MSDSE experiment by allocating a number of faculty half-lines and research scientist positions. The university has likewise demonstrated its commitment by approving two new formal educational programs (the data science master's degree program and the Advanced Data Science option for doctoral students). The successes of the Data Science Incubator Program and Data Science for Social Good summer program at demonstrating the return on investment to the broader campus community also suggests that university will strive to preserve key elements of the data science environment post-DDD.

At NYU as well, the MSDSE experiment appears to have accumulated sufficient momentum to endure the end of DDD funding. Like the eScience Institute, NYU's MSDSE also received several half-faculty lines and two research scientist positions for the CDS from the provost, and MSDSE leaders anticipated that these positions would remain after the MSDSE award period has ended. In addition, the CDS has a highly selective data science master's degree program, and just launched a data science doctoral program as well. Respondents expressed optimism that the CDS has a sustainable, revenue-generating model, although they had not yet determined how to continue funding for some research scientist positions.

4.2.2 Sustaining the Development of Science-Enabling Tools and Practices

To sustain momentum in the adoption of tools and practices for data-driven discovery, the scientific community needs a standard set of principles for acknowledging and citing software. First, citation will help drive further adoption of these tools; and second, citation standards will highlight the contributions of the software developers to scientific inquiry—which may be one link closer to a better system for rewarding and retaining these individuals in academia.

More than 90 percent of scientists agree that software plays an important role in their research, and nearly 70 percent that their research would be infeasible without software.[143] Scientists are also the primary developers of research software.[144] Unfortunately, software is not cited consistently, and informal acknowledgments often lack crediting information.[145] Standardized citation of software will encourage scientists to acknowledge the contribution of software to their research. In turn, citing software in research reports will enable assessments of the role of software in scientific discoveries. Recent developments suggest cautious optimism for the propagation of software citation practices. Building on the success of data citation efforts, proposed guidelines for citing software have emerged

---

143 Hettrick, S., Antonioletti, M., Carr, L., Cheu Hong, N., Crouch, S., De Roure, D., Emsley, I., ... Sufi, S. (2014). *UK Research software survey 2014* [Data set]. doi: 10.5281/zenodo.14809.

144 Groble, C. (2014). Better software, better research. *IEEE Internet Computing, 18*, 4-8. Retrieved from http://ieeexplore.ieee.org/document/6886129/

145 Howison, J. & Bullard, J.A. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology, 67*, 2137-2155.

in the past two to three years.[146,147] The attention of research funders and publishers to data and software citation issues also suggests that the scientific community may soon converge on a set of such principles.[148,149]

Standards for data and software citation will allow scientists to demonstrate their contributions to their field and the impact of these critical research products more transparently. In turn, enabling science to track the role of software could plausibly influence a reward structure in academic science better aligned with the unique role of data-driven scientists. Concurrently, as more scientists share their code, the demand for standard citation norms will grow.

4.2.3 Bolstering Institutional Exploration of New Career Pathways for Data-Driven Scientists in Academia

The three DDD initiative strategies have begun to challenge traditional definitions of who is valued in academic research settings, as well as how institutions can recruit, hire, and reward people whose expertise sits outside traditional roles in academic research, but establishing new career pathways for data-driven researchers in academia remains a difficult challenge. Some DDD Investigators and all three MSDSEs have hired research software engineers; and *Practices* projects have collaborated with industry, yet retained their tools' open source status. Still, these individuals' contributions to science are too often unrecognized.

Although there are multiple challenges of providing career pathways for "pi-shaped" individuals who have both domain knowledge and methodological expertise, one prominent theme throughout the evaluation has been the need for better software development in scientific research. **One suggestion worth exploring is that academic institutions establish formal positions and career pathways for "research software engineers" to improve the software on which most scientific research now depends**.[150,151,152] Career challenges for data-driven researchers are not limited to software engineers

---

146 Smith, A. M., Katz, D. S., Niemeyer, K. E., & FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science 2*:e86. https://doi.org/10.7717/peerj-cs.86

147 Gent, I., Jones, C., & Matthews, B. (2015). *Guidelines for persistently identifying software using DataCite*. [Report.] Swindon, UK: Science & Technology Facilities Council. Retrieved from https://epubs.stfc.ac.uk/work/24058274

148 White, O., Dhar, A., Bonazzi, V., Couch, J., Wellington, C. (2014). *NIH Software Discovery Index Meeting Report*. [Report]. Bethesda, MD: NIH. Retrieved from http://www.softwarediscoveryindex.org/

149 Stodden, V., Guo, P., Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE, 8*, e67111. Retrieved from https://doi.org/10.1371/journal.pone.0067111

150 Groble, C. (2014).

151 Jiménez, R.C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M. Capella-Gutierrez, S., ... Crouch, S. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research, 6*, 867. Retrieved from https://f1000research.com/articles/6-876/v1#ref-4

152 Brett, A., Croucher, M., Haines, R., Hettrick, S., Hetheringotn, J. Stillwell, M. & Wyatt, C. (2017). Research software engineers: State of the nation report 2017. Southampton, U.K.: University of Southampton (on behalf of the Research Software Engineer Network).

working in academia; nevertheless, the challenges of finding a rewarding career path as a software engineer in academia illustrate an important test case.

In 2015, the U.K.-based Engineering and Physical Sciences Research Council (EPSRC) solicited applications for a Research Software Engineering (RSE) Fellowship that specifically targeted early career doctorates providing software "that is used as a research tool in science and engineering" in academic institutions. The purpose of the fellowships was

> *to provide long-term funding for those individuals working as Research Software Engineers in universities, to give them the resources to develop their careers and their skills. It has the additional aim of encouraging universities to recognise the role of the Research Software Engineer in supporting research.* [153]

Notably, the solicitation required host universities to describe mentoring and career development opportunities that the institution would provide plans for the Fellow after the end of the fellowship period and confirmation that the Fellow would be eligible for Investigator status in grant proposals to funding entities. Seven of 201 applicants received an RSE Fellowship; one RSE Fellow described the challenge of research software experts in academia:

> *We don't fit the normal 'money-in, papers-out' model of many academics...Many RSEs [research software engineers] are on short-term contracts with low salaries. In short, [we] get much of the grief of working in academia without any of the benefits. Little wonder, then, that many of the best in the community choose to work in industry.*[154]

The EPSRC has since funded a Research Software Engineering Network (RSEN) to coordinate communication and sharing across various RSE groups. An inaugural conference of the U.K. Research Software Engineering Association organized by the RSEN in 2016 drew more than 200 attendees, including funders, academic researchers, industry representatives and research software engineers from 14 nations (attendees included a DDD Investigator, non-awardee, BIDS data science fellow, and other DDD stakeholders). Attendees' active social media presence[155] and publication of a 2017 report from the RSEN has amplified the impact of this conference.

As the dilemma of research software engineers demonstrates, there are challenges for establishing rewarding career opportunities in academia for some of the individuals who play a vital role in data-driven scientific research; yet there are early indicators that funders and academic institutions have begun to acknowledge the need for innovative career pathways. The fact that the EPSRC's investment in career pathways for research software engineers is a rare experiment to date calls for a tempered optimism; nevertheless, the RSE fellowships provide an example of different stakeholders uniting in recognition that changes are imperative if academic research institutions are to retain this type of expertise.

---

153 EPSRC (2015). Research Software Engineer (RSE) Fellowships, Invitation for proposals [Funding announcement]. Retreived from https://www.epsrc.ac.uk/files/funding/calls/2015/rsefellowships/

154 Croucher, M. (2016). EPSRC Research Software Engineering Fellow: Mike Croucher. [Blog]. Retrieved from http://www.walkingrandomly.com/?p=6033

155 See https://storify.com/ResearchSoftEng/world-s-first-rse-conference and Brett, et al. (2017).

4.3 Unmet Needs in Data-Driven Science

DDD Investigator and non-awardees in the DDD Investigator competition who participated in a survey for the evaluation perceived several unmet needs for data-driven research at their institutions. When asked to hypothesize about how they would allocate time to discuss potentially unmet needs for data-driven research at their institution, the 44 survey respondents, representing approximately 30 different academic research institutions, allocated the largest blocks of time, on average, to the need for their institutions to:

- Hire more full-time, permanent data scientists or software engineers.
- Hire junior faculty with data-driven expertise.
- Incorporate additional training in data-driven methods or tools into existing degree programs.
- Create interdisciplinary centers for data-driven research.

More than 60 percent of survey respondents (combined across the two groups) indicated that *all* of the unmet needs listed on the survey deserved some attention from university administrators (Exhibit 4.1). Although these findings reflect data from a small sample of accomplished data-driven

Exhibit 4.1: Survey Respondents' (N=44) Perception of Unmet Needs for Data-Driven Research at Their Academic Institutions

Potential Unmet Need for Data-Driven Research at [Your Institution]	Number Who Would Devote Time to Discussing Each Unmet Need With Their Institution's Administration		
	All Respondents (n=44)	DDD Investigators (n=13)	Non-Awardees (n=31)
Hire junior faculty with expertise in data-driven research	41 of 44	12 of 13	29 of 31
Hire full-time, permanent data scientists or software engineers	39 of 44	13 of 13	26 of 31
Incorporate additional training in data-driven methods or tools into existing degree programs	37 of 44	12 of 13	25 of 31
Establish degree programs in data-driven research	33 of 44	10 of 13	23 of 31
Hire senior faculty with expertise in data-driven research	31 of 44	9 of 13	22 of 31
Offer salaries for data scientists that are competitive with industry	31 of 44	12 of 13	19 of 31
Create interdisciplinary data-driven research centers	30 of 44	10 of 13	20 of 31
Provide physical spaces for data-driven researchers to work and gather	30 of 44	10 of 13	20 of 31
Invest in computing infrastructure	27 of 44	6 of 13	21 of 31
Other - please describe	8 of 44	0 of 13	8 of 31

Notes: DDD Investigators (N=13); Non-awardees (N=32, Missing= 1).

Source: Survey of DDD Investigators and non-awardees. (Item D2. How many minutes (out of 100) would you use to talk about the following unmet needs for data-driven research at your institution?)

researchers at top research universities, the fact that the majority of respondents perceived unmet needs across several elements of their working environments suggests that these elements function synergistically, requiring institutions to consider a portfolio of coordinated initiatives to effect change. These needs, each of which was seen as worthy of some institutional attention, appear to fit into three clusters: space to meet with colleagues from multiple domains; access to other data-driven

faculty and to data scientists and software engineers; and educational initiatives to build capacity of students to contribute to data-driven research.

4.4 Potential Opportunities

Based on the information we have collected and analyzed towards the end of the DDD initiatives' six-year initial grant cycle, however, we find several potential opportunities for continuing—and enhancing—progress thus far.

4.4.1 Suggestions From Interview Respondents

Institution-Level Challenge Projects

One interview respondent at an MSDSE institution wondered whether a funding challenge to non-MSDSE institutions that centered on a concrete research problem requiring a cross-campus (or cross-institutional) collaboration of domain scientists and computational methodologists would advance the DDD initiative's goals—perhaps as a complement, or an alternative, to the approach taken in funding the MSDSEs. This suggestion is somewhat reminiscent of the incubator programs at the MSDSEs. Under the MSDSEs' incubator programs, scientists propose a specific research challenge, typically requiring new software or computational methods, and a panel of reviewers (data science fellows) selects proposals that seem amenable to short-term development of a prototype.

In contrast to asking non-MSDSE institutions to develop a proposal for a data science environment, it seems plausible to consider inviting these institutions to compete for the opportunity to tackle a data-driven challenge. NSF has experimented with this approach by identifying "grand challenges" and targeting them with specific programs.[156] However, an alternative would be *to invite institutions to propose a specific problem* that would require a similar concentration of attention and resources and collaborations of cross-disciplinary teams of domain scientists and computational methodologists.

The incubator projects at the MSDSEs resulted in proposed solutions—some worthy of additional development, some not—to concrete problems that brought together domain scientists and methodologists. If these types of incubator projects can engage the concentrated attention and resources of small teams of individuals to stimulate and even accelerate action toward a solution, it may be interesting to pursue a similar approach on a larger scale at non-MSDSE institutions. Characteristics that appeared to contribute to the successful implementation of the MSDSE incubator projects may be scalable to institution-level opportunities. These characteristics included: an intermediate-length, closed-ended project duration that neither overburdens participants with long-term, open-ended engagements, nor limits the time needed for executing plans and evaluating the results; a sufficient number of personnel with the right mix of expertise (within-domain and in cross-cutting methods), time, and administrative support to select and execute the portfolio of projects; and mutual agreement that the project is a true collaboration—not the provision of technical assistance—for which participants are equally responsible. There may also be relevant literature on similar "grand challenge" approaches have helped frame the development of some research funding programs.

For universities not able, or not willing, to commit to an initiative at the scale of an MSDSE-like data science environment, an opportunity to pursue a more time-limited but cross-departmental project-

[156] See for example, NSF's cross-cutting Science, Engineering, and Education for Sustainability (SEES) portfolio. Available at https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504707

based collaboration could demonstrate the benefits gained from domain scientists and methodologists collaborating around a shared goal. Because the commitment of the participants to each other would be time-limited (for example, one year), participants would likely view their investment as having the potential for a large reward in exchange for relatively modest risk. The problem could engage researchers' graduate students as well, extending the benefits to the next generation, just as the MSDSEs offered students opportunities to learn from peers in other disciplines. Such projects have the potential to generate intellectual excitement, stimulate skill development (and application of specific skills in often novel ways) and collaborative ventures, and increase awareness of data-driven science, practices and tools.

Engaging Research Libraries and Research Computing Centers

The MSDSEs illustrate that research libraries and universities' centers for research computing—in contrast to campus information technology units—have the potential to play a key role, in at least two complementary ways: first, research librarians and technology staff have *specialized* knowledge that is particularly valuable in a data science-rich context. They understand digital information management and storage; librarians in particular also recognize the importance of tracking research productivity, and the methods and practices that enable information dissemination in a digital era. Second, librarians have *broad interdisciplinary* knowledge based upon their day-to-day experiences working with faculty and students across multiple domains of science and across numerous dissemination avenues. Leveraging such resident staff expertise represents another opportunity for the DDD initiative to continue to strengthen the infrastructure supporting data-driven science. At some institutions, there may also be opportunities to engage research computing staff in a manner that expands (and rewards) their role from technicians who facilitate access to resources to more substantive participation in solving challenges in scientific research labs and identifying potential cross-domain opportunities.

Community-Building Events

The evaluation revealed broad support for small, focused gatherings for researchers from multiple institutions. Although the annual MSDSE summits, the 2016 Data Summit, and institution-specific events like Data Science Fairs showcase the intellectual vibrancy of data-driven science for early career and established researchers, students, and university and industry colleagues DDD Investigators, postdocs, and graduate students who attended the smaller investigator and early career symposia universally praised these events. They saw these gatherings as valuable opportunities for potential collaborations and for exchanging lessons learned for navigating the search for research funding and hospitable publication venues for data-driven research. Participation in such events has resulted in new collaborations, opportunities for informal mentoring, and learning about new tools that other researchers have used, developed, or tested. These meetings may also help investigators identify and recruit the next postdoctoral fellow, research scientist, or graduate student to join their research group. Especially for individuals early in their careers as data-driven researchers, spending a day or two with a small community of like-minded researchers, postdocs, and students, with a set of shared talks and workshops, provides meaningful exchange and reassurance that their work is valued.

4.4.2 Other Avenues

We conclude with two other potential opportunities for gaining additional traction toward some of the DDD initiative's institutional change goals. First, examples of successful university-industry partnerships suggest that thoughtful exploration of similar arrangements may prove fruitful. The

participation of software developers from the technology sector in Project Jupyter has provided essentially "free" expertise and labor to expand the capacity of the team. Intel and Microsoft Research have had a robust presence in academic research settings, including embedding employees in laboratories to contribute to particular projects. Examining such partnerships more explicitly to identify the conditions that support effective partnerships could yield insights into important preconditions or ground rules that enable productive partnerships to flourish.

Second, it may be useful to convene senior administrators and academic leaders from across the DDD initiative, and potentially beyond, to take stock of progress at the MSDSEs (in real time) with respect to career pathways for data scientists within the academy. Possible candidates for this group include key program staff from the Moore Foundation as well as Sloan, leaders from the MSDSEs, and relevant academic deans and research vice provosts from 10 to 15 universities with strong data science initiatives.

Concluding Remarks

As the initial funding phase nears conclusion, it is already clear that the DDD initiative has had a strong imprimatur on data-driven science. The initiative has been at the forefront of interest and engagement in this area at academic institutions and research funders, and has made common cause with associations of scientists advocating for more transparent and reproducible research practices. The initiative has also filled a gap by devoting resources for fundamental tool development to enable scientific inquiry; this investment has demonstrated the need to support organizations to devote full-time effort to making such tools user-friendly and accessible to scientists, and ensuring that they meet production quality standards to ensure their reliability.

Despite signs of increasing attention to the needs of data-driven science, the DDD initiative remains unique in its orientation and strategies. Advancing the initiative's goals further may require continued attention to maintain an emphasis on two critical needs: (1) viable career pathways in academia for research scientists, particularly software developers and computational specialists; and (2) support for those organizations (or emerging organizations) that focus on providing a broad suite of tools and resources for data-driven science. While a five- to six-year investment strategy may seem long at conception and its outset, it is also a relatively short amount of time in which to achieve goals that only now may be gaining momentum. It is also clear that it is too soon to take full measure of the DDD initiative's cumulative and collective investments, as its longer-term investments have yet to come to fruition.

Limitations of the Evaluation

As with any single research study in isolation, the mid-term evaluation of the DDD initiative is subject to certain constraints that limit the interpretation of its reported findings. Noting these limitations encourages appropriate caution and may help inform subsequent assessments.

First, the evaluation design limits its ability to address questions of causal attribution. This challenge of *selection bias* is not unique to the evaluation of the DDD initiative, but is typical for evaluations of research funding programs. Selection bias occurs when characteristics that affect selection for a grant (or into a program) are also correlated with outcomes of the grant program. For example, the DDD team likely used prior evidence of contributions to data-driven science or science-enabling methods as a factor in selecting grantees. The merit that contributed to grantees' selection for DDD funding means that these grant recipients would likely continue to advance professionally, achieve further success in their research, even in the absence of the DDD award. Evaluation designs that mitigate selection bias were not feasible. These designs would have required random assignment of individuals or institutions to receive (or not receive) DDD funding, or a quasi-experimental design with a comparison group. Most quasi-experimental designs require large samples sizes or have other assumptions not met by the DDD initiative.

As with any program evaluation, time and resources constrained the scope of data collection. For example, we conducted interviews with administrators, postdoctoral fellows and graduate students from just five of the fourteen DDD Investigators' institutions. Similarly, the survey is subject to the limitations of a small sample size: the initial sample included just 93 DDD Investigators and non-awardee finalists and semi-finalists in the DDD Investigator competition, and analyses drew on the 45 responses received. Due to the timing of the evaluation, there was no opportunity to collect "baseline"

data pre-award, and because the grant periods are still ongoing, some outcomes of interest may not have yet emerged.

Moreover, in an effort to limit the burden on the MSDSEs and their institutions from two evaluations (with two Abt teams of researchers)—the mid-term evaluation of the entire DDD initiative and an ongoing developmental evaluation of the MSDSEs—Abt and the Moore and Sloan Foundations agreed that the mid-term evaluation of the DDD initiative would leverage data from interviews collected to date as part of the developmental evaluation. Thus, the mid-term evaluation drew both on the Year 1 report of the MSDSE evaluation and a sample of interviews conducted in the spring of 2017 as part of the MSDSE evaluation. This sample included interviews with MSDSE administrative personnel, faculty, and Working Group leaders, along with key administrators (e.g., academic deans, vice provosts for research) at each host institution. Nevertheless, evidence bearing on the MSDSEs came from multiple sources, including the shared and individual MSDSE annual reports, MSDSE websites, research publications and github postings, a rich repository of information. The two evaluation teams included some of the same individuals at Abt who met regularly to ensure that information from these sources accurately reflected the status of the MSDSEs to date.

Strengths of the Evaluation

The use of multiple sources of data gives the mid-term evaluation credibility and strengthens the reliability of its reported findings. By integrating data from interviews with an online survey (including, but not limited to, some of the interview respondents), annual reports, and other independent sources, we were able to check for consistency and corroborate assertions. For example, the survey findings largely confirmed themes reported by interview respondents. This agreement lends credence to conclusions made from such data. Moreover, independent sources of information in the scholarly literature, news sources, and github postings also substantiated observations of DDD grantees, their colleagues, and non-awardees who applied for a DDD Investigator award.

The evaluation also incorporates viewpoints from a large number of respondents (48) selected to represent a diverse set of roles in the landscape of data-driven science, providing a diversity of perspectives that also strengthens the evaluation. When respondents from different vantage points agree with one another, trust in the veracity of individual reports improves. When respondents provide alternative or conflicting reports about the same topic or set of events, they signal that conclusions must be contextualized, tempered, or suspended until further inquiry is possible. From academic research institutions, the evaluation included the perspectives of faculty (both tenure-track and not), department chairs, directors of research centers, deans and provosts, as well as graduate students, postdoctoral researchers, and research scientists. From the MSDSEs, the evaluation also benefited from the perspectives of administrative leaders. To assess the outcomes of the four *Practices* grantees included in the evaluation, data came both from leaders of these organizations and from individual users of these grantees' tools and services.

Finally, the individuals interviewed for the evaluation are not mere spectators, but are active participants either in the DDD initiative's funded activities or in the broader landscape in which the initiative was implemented. As such, they are the relevant experts about their disciplines, their institutions, and their colleagues. These respondents were those who have either experienced the DDD initiative directly, or who have had an opportunity to observe its effects.