New Collaboration for New Education: Libraries in the Moore-Sloan Data Science Environments

Jennifer Muilenburg, Research Data Services Librarian, University of Washington, and Visiting Program Officer, Association of Research Libraries

Judy Ruttenberg, Director, Scholars and Scholarship, Association of Research Libraries

In 2014, the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation partnered to invest $37.8 million across three US universities to build what they called Data Science Environments in order to "demonstrate how an institution-wide commitment to data science can deliver dramatic gains in scientific productivity and lead to significant new discoveries."[1] The Moore-Sloan Data Science Environments (MSDSE) were a five-year experiment "to better understand how best to bring together interdisciplinary people within institutional environments in order to provide them with the resources, freedom, and interconnected networks necessary for science to flourish." As that five-year experiment and its funding wind down, how can these and other research institutions—and their libraries—advance data science education? What did the MSDSE sites learn about the pedagogical relationship between data science methods and traditional disciplines, and intra-institutional partnerships? ARL staff spoke with key personnel at the three MSDSE sites: New York University (NYU), UC Berkeley, and the University of Washington (UW).

Of course, these three institutions are not unique in dedicating resources to data science; many other universities in Canada, Europe, and the United Kingdom have also launched degree and research programs focused on data science and analytics.[2] This grant, however, structured as it was across three institutions, provided a unique opportunity to assess each university's goals against the outcomes achieved. This article highlights not only their individual learning, but also draws overarching conclusions of value to all research libraries engaged with data science or looking for a pathway to doing so.

Common Themes and Distinctive Paths

Each MSDSE institution focused on interdisciplinarity and collaboration—attributes that data science fosters and requires. Each also initially established or built up a data science education program outside of existing departments (but in collaboration with them) in order to encourage institution-wide cohesion. In addition to local endeavors, the three institutional groups also came together frequently, through focused meetings on a particular topic, and in an annual summit to talk about research, education, careers, and the progress of the grant itself.

There were core themes, goals, and structure to the three-site design. Each institution participated in six cross-institutional working groups focused on recognized challenges to data-driven science, including (1) careers, (2) education and training, (3) tools and software, (4) reproducibility and open science, (5) working spaces and culture, and (6) data science studies. NYU, UC Berkeley and UW approached curriculum-building differently—some starting with graduate and others with undergraduate courses—and each emphasized the six themes to greater or lesser degree. NYU's focus on openness and reproducibility; UC Berkeley's development of Data 8, an entry-level course designed for students in any major; and UW's widely adopted data science "options" across the curriculum are examples of their individual distinction.

> "Core to the development of data science at the three MSDSE universities was a posture of non-competitiveness with existing programs and disciplines...and a belief that the set of tools and methods that comprise data science are critical to unlocking new discoveries across all research domains."

All three institutions have made many of their research and education materials open to the public. For example, the Data 8 course can be viewed online and freely reused at http://data8.org/, UW has three massively open online

courses available via Coursera at https://escience.washington.edu/education/mooc/, and all three institutions contributed to the open access book *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*, available at https://www.practicereproducibleresearch.org/.[3] Core to the development of data science at the three MSDSE universities was a posture of non-competitiveness with existing programs and disciplines, particularly in computer science, engineering, and statistics, and a belief that the set of tools and methods that comprise data science are critical to unlocking new discoveries across all research domains.

Abt Associates conducted an assessment of the grant in 2015 (at the midway point), and published its findings in 2019. Abt found that the MSDSE funding strategy was effective at increasing positive conditions for data-driven discovery at academic institutions, and noted that "the culture of partnership and experimentation adopted by the program facilitated mutual learning and growth."[4] The Abt report also stressed the importance of strong management and staffing, the success at establishing promising career tracks, and new programs that led to collaboration. In a secondary review of 20 other data science entities, Abt also found that there is no one way to create a data science center: culture, administration, physical space, funding, and people all combine in different ways to work toward supporting data-intensive science on campuses.

> "There is no one way to create a data science center: culture, administration, physical space, funding, and people all combine in different ways to work toward supporting data-intensive science on campuses."

Finally, a concentration on ethics was not a formal objective of the funding agencies providing MSDSE funding, but NYU, UC Berkeley, and UW have all developed an emphasis on data science and ethics, or data science and the public good. As the sites developed informal and formal curricula, and started working with students, it was immediately clear how important ethics education is to these efforts.

Research libraries are keenly interested in supporting data-intensive science under conditions of privacy, human subjects protection, and inclusivity. The commitment to ethics within the MSDSE and other academic data science programs bodes well for the library as a critical campus partner in the future of data science education.

NYU and the Center for Data Science

> As we transition to data-intensive scientific discovery, we have the opportunity to address these issues through software tools and practices that support the sharing, preservation, provenance tracking, and reproducibility of data, software, and scientific workflows.[5]
>
> —"Reproducibility and Open Science," Moore-Sloan Data Science Environments

Professor David W. Hogg (Physics and Data Science) served as executive director of the MSDSE at NYU in 2013–2015. Hogg has been involved with the project and the Center for Data Science (CDS) from the beginning. He emphasized interdisciplinarity as the main driver of the Moore-Sloan grant, and of the resulting transformation to data science education at NYU. While the NYU Libraries were not an original partner on the grant, Hogg acknowledged, he quickly learned that the library is "the most interdisciplinary place on campus," and credited the library for its influence on the CDS over the life of the MSDSE grant. In fact, in 2015, Vicky Steeves was appointed librarian for research data management and reproducibility, reporting jointly to the NYU Libraries and the CDS.

Steeves talked to ARL staff about how the norms and practices within data science—namely openness and reproducibility—are helping to transform disciplines that interface with CDS as well as the library. "Data science is a set of methods involving lots of computing power, lots of fast access to data, and the norms are open," said Steeves. Hogg also talked about openness as a cultural norm, and explained that by working with Moore and Sloan, along with NYU's lawyers and a legal team at UW, the CDS was able to construct an intellectual property

(IP) policy whereby any of the center's inventions would be released under an open license. UC Berkeley's Institute for Data Science (BIDS) and UW's eScience Institute also have working groups to advance reproducibility and open science. According to Steeves, the MSDSE has contributed to a wider conversation about openness at NYU in general.

NYU is currently launching an undergraduate Data Science course, which may lead ultimately to an undergraduate major. But Hogg pointed out that the MSDSE project did not set out to create formal degree programs at its three sites. Rather, it was used to create informal educational opportunities that would draw from and aid the disciplines. For example, one of the accomplishments Hogg is most proud of is CDS's deployment of Hack Weeks—informal opportunities to learn by doing in the company of experts. He was a co-author of "Hack Weeks as a Model for Data Science Education and Collaboration," along with colleagues from UC Berkeley and UW, which assesses this work at the three institutions in promoting interdisciplinarity, methods rigor, skill building, and positive attitudes toward open science in general.[6] Geohackweek, Oceanhackweek, and Waterhackweek were created and sponsored at UW's eScience Institute after the initial Astrohackweek offering.

The MSDSE provided all three site institutions the opportunity to experiment with different approaches to education and different partnership arrangements. While UC Berkeley and UW Libraries each housed the initial MSDSE sites, at NYU, the CDS started out in temporary space, and then worked closely with the libraries when it designed its new permanent space. The library, Hogg came to find out, had a deep understanding of how people across the campus and across disciplines use space—for writing, reading, interacting with technology, and collaborating. The library had experience with modularity and flexibility in space design that greatly informed CDS's new environment.

Likewise, the growth of data science throughout the university has influenced the library's collecting, such as purchasing more vendor-

produced data sets, responding to students' need for big data (for example, large social media feeds), and integrating APIs into their collection and discovery environment. The library has also ramped up its teaching of data literacy, and there is a new CDS-IT-library joint program called DS3 (data science and software services).

UC Berkeley and the Berkeley Institute for Data Science

> What faculty needed their future graduate students to know, no undergraduate program in the world taught.
>
> —David Culler, UC Berkeley

David Culler and Cathryn Carson were among the original co-investigators on the MSDSE grant to UC Berkeley. Culler is the interim dean for Data Sciences, and Carson is the faculty lead on UCB's undergraduate Data Science program. In 2012, when data science was still an emerging field, Culler, Carson, and their colleagues began an extensive inventory of related practice throughout the university, and began planning to launch the Berkeley Institute for Data Science (BIDS) in order to be competitive as an MSDSE site. Hundreds of faculty participated in the inventory that led to BIDS and eventually their successful proposal. "What faculty needed their future graduate students to know, no undergraduate program in the world taught," said Culler. The process, he added, "revealed a crack in the foundations of the research university. The world had changed and we had not gotten out ahead of it." Carson, a trained historian, was even more pointed: "UC Berkeley was ready when the Moore-Sloan funding was on the table." Undergraduates were moving in greater numbers to "real-world facing" computer science, statistics, and applied math programs. "Students were ready, and they were basically trying to do it on their own without the university," Carson said.

Once the university launched BIDS, based in Doe Library, as a new center of data science activity, other programs became involved, including the College of Engineering and the I School. With MSDSE investment, UC Berkeley built a data science curriculum from the first-year-student class up. "As we saw it, this was a moment to rethink

at a fundamental level what every educated person must know about quantitative reasoning: how to effectively understand, process and interpret information, to inform decisions in their professional and personal lives and as citizens of the world in the 21st century,"[7] said A. Paul Alivisatos in testimony to the US House of Representatives Science Committee in 2017. From the outset, the library was considered a key partner by virtue of their convening and collecting functions, enabling easy, broadly distributed opportunities for students to experiment with data and obtain peer consulting.

Data 8

The UC Berkeley Foundations of Data Science course combines three perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. It delves into social issues surrounding data analysis such as privacy and design.

—"Data 8: The Foundations of Data Science," http://data8.org/

Both Culler and Carson described the MSDSE experiment at UC Berkeley as transformative, and indicated that its enduring effects transcend data science and serve as a model for how faculty can come together around transdisciplinary research. Having established mechanisms for transdisciplinary practice as a new academic structure through BIDS, data science will become part of a larger engine at Berkeley. Culler observed that just as "Moore-Sloan wanted to create new career paths, UC Berkeley created new institutional paths." Carson hopes that "the new integrative structure is a good model for other [programs], that are conventionally schools and colleges—to build core strength [in those programs] and have solid connections outward." Might a new school or college of data science emerge at Berkeley or the other MSDSE sites? Perhaps, Carson conceded, but they would be

different: less competitive and with strong connections throughout the university.

In spring 2019, the UC Berkeley Libraries launched a Library Data Initiatives Plan. The plan will accelerate the library's focus and resources on data literacy, data acquisition, and housing data in a local repository. "For the Library, one constant goal is to demystify data science for the campus community, building new pipelines into the field from all directions."[8]

The University of Washington and the eScience Institute

> People come here [to the library] and are more likely to collaborate across disciplines than they might if they were all going to somebody's particular department.[9]
>
> —2017 Abt Associations Site Visit

ARL staff spoke with Sarah Stone, executive director of the University of Washington's eScience Institute—a campus unit that describes itself as UW's "hub of data-intensive discovery on campus."[10] Jennifer Muilenburg, a visiting program officer at ARL, has worked closely with Stone for the past five years as a data librarian and member of the eScience Institute Steering Committee at UW. Stone describes the library's role as central, and more aligned with the university's formal educational programs than the informal opportunities created by the MSDSE. They spoke in similar terms to UC Berkeley about the new program's non-competitiveness on campus, and its mission to integrate with all disciplines. These latter features are also characteristics of the library, which provide both physical space and expertise to the endeavor.

> Partnering with our libraries has been an important component of these efforts. For example, at UW several librarians who specialize in data management are data science fellows, and efforts toward a new institutional data repository, clarification of intellectual property rights concerning software and data, and a proposed Open

> Access Policy for faculty publications have all been informed by DSE efforts.[11]

Outside of a professional Master's Program in Data Science, geared toward retraining students returning to the workforce, the UW MSDSE did not set out to create a separate degree program. Rather, it took the approach that data science education should reside within the departments. At both the undergraduate and graduate levels, departments have ownership of the tools and techniques of data science that enhance their programs, and have the options and flexibility to implement what is most beneficial in their curricula. In contrast to UC Berkeley, UW started developing its data science curriculum at the doctoral level.

At UW, there was a pre-existing category referred to as a "transcriptable option," similar to a minor, that could be recognized on a student's transcript. This option, used at both the graduate and undergraduate level, allows a student to take a small number of data-focused courses in key data-science areas (such as statistics, machine learning, data management, data visualization, and ethics/privacy) in addition to their chosen major. Upon completion, the student receives a statement on his or her transcript—for example, "Bachelor of Science in Bioengineering, Data Science Option."

In looking for ways to expand data science techniques into the disciplines, UW has put much of its energy into its Incubator Program and Data Science for Social Good. Both programs match a researcher who is studying an existing problem with a group of data scientists who can help with the data-intensive part of the work. And both programs bring together a diverse set of participants who learn from one another, across disciplines, to help solve problems.

Conclusion

Each of the three MSDSE sites has had significant interaction and collaboration with their university library. At both NYU and UW, the newly designed data science centers were established in library

spaces. As stated in the Abt report, at UC Berkeley, by being in a library space, "The large investment in renovation and the choice of popular location signaled a commitment to data science by the administration and elevated the status of BIDS." At UW, "The open and modular layout of the space was intended to signal inclusivity to the university community and to foster collaboration." Libraries are seen as discipline-agnostic and welcoming to everyone, a feeling that carried over into the motivations of the data science centers themselves.

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

As part of the evaluation, Abt also looked at 17 other universities that have data science efforts underway, and found that each shared some characteristics with UC Berkeley, NYU, and UW, including the fact that nearly all were established in a location outside an established department. This helped signify inclusiveness and openness to collaboration. It's not known how many of these involved their library in the decision on location, but collaboration between data science entities and the libraries has proven beneficial for both parties in the MSDSE context, as well as the campus communities as a whole.

In *Creating Institutional Change in Data Science*, the authors emphasized the library's unique position with respect to the disciplines in hosting data science endeavors in their space:

> We have found that the ideal space is untethered to any department and is in a central location on campus where students and faculty habitually visit. In these respects, former library spaces are excellent choices. Our university libraries and librarians have been integral

partners to our efforts. Intellectually, libraries and librarians are in the information business.

Libraries at NYU, UC Berkeley, and UW, and many other research libraries in Canada and the US have incorporated data science education in their collecting and service frameworks. "Library spaces have been transformed—among other things—into campus centers for data science research, training, and services, with open floor plans and furnishings that are adaptable to a range of activities that promote and support data science research and learning.[12]

Endnotes

1. "About," Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/about/.

2. A partial list is available under "Other Data Science Centers" on "Environments," Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/environments/#others.

3. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*, ed. Justin Kitzes, Daniel Turek, and Fatma Deniz (Oakland, CA: University of California Press, 2018), https://www.practicereproducibleresearch.org/.

4. Luba Katz, *Evaluation of the Moore-Sloan Data Science Environments: Final Report*, (Abt Associates, 2019), iii, https://web.archive.org/web/20190501190246/http://msdse.org/files/ABT_MSDSE_Eval_Report_Feb2019.pdf.

5. "Reproducibility and Open Science," Moore-Sloan Data Science Environments, accessed June 26, 2019, http://msdse.org/themes/#reproducibility.

6. Daniela Huppenkothen, Anthony Arendt, David W. Hogg, Karthik Ram, Jacob T. VanderPlas, and Ariel Rokem, "Hack Weeks as a Model for Data Science Education and Collaboration," *Proceedings of the National Academy of Sciences* 115, no. 36 (Sep 2018): 8872–8877, https://doi.org/10.1073/pnas.1717196115.

7. *Hearing: STEM and Computer Science Education: Preparing the 21st Century Workforce, before the Research and Technology Subcommittee, House Committee on Science, Space, and Technology*, 115th Cong. 12 (2017) (testimony of A. Paul Alivisatos, executive vice chancellor and provost and vice chancellor for research, University of California, Berkeley), https://docs.house.gov/meetings/SY/SY15/20170726/106330/HHRG-115-SY15-Wstate-AlivisatosA-20170726.pdf.

8. Virgie Hoban, "Library Launches Initiative to Boost Data Science Expertise, Services at UC Berkeley," *UC Berkeley Library News*, April 25, 2019, https://news.lib.berkeley.edu/data-initiative.

9. Katz, *Evaluation of the Moore-Sloan Data Science Environments*, 27.

10. "About Us," University of Washington eScience Institute, accessed June 14, 2019, https://escience.washington.edu/about-us/.

11. *Creating Institutional Change in Data Science: The Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington*, Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/files/Creating_Institutional_Change.pdf.

12. *Creating Institutional Change in Data Science*.

To cite this article: Jennifer Muilenburg and Judy Ruttenberg. "New Collaboration for New Education: Libraries in the Moore-Sloan Data Science Environments." *Research Library Issues*, no. 298 (2019): 16–27. https://doi.org/10.29242/rli.298.3.