Creating Institutional Change in Data Science

The Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington

Abstract

This paper is intended for academic leaders who seek to accelerate data-intensive discovery and data science education on their campuses. It is based on the experience of our three institutions (New York University; University of California, Berkeley; and the University of Washington) in the Data Science Environments effort sponsored by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation—a collaborative experiment intended to *transform the process of discovery and the institutional environments in which discovery takes place*.

We describe key elements we believe have contributed to our successes as well as some challenges we have encountered along the way. Our goal is to provide a menu of possibilities for those seeking to emulate aspects of our Data Science Environments.

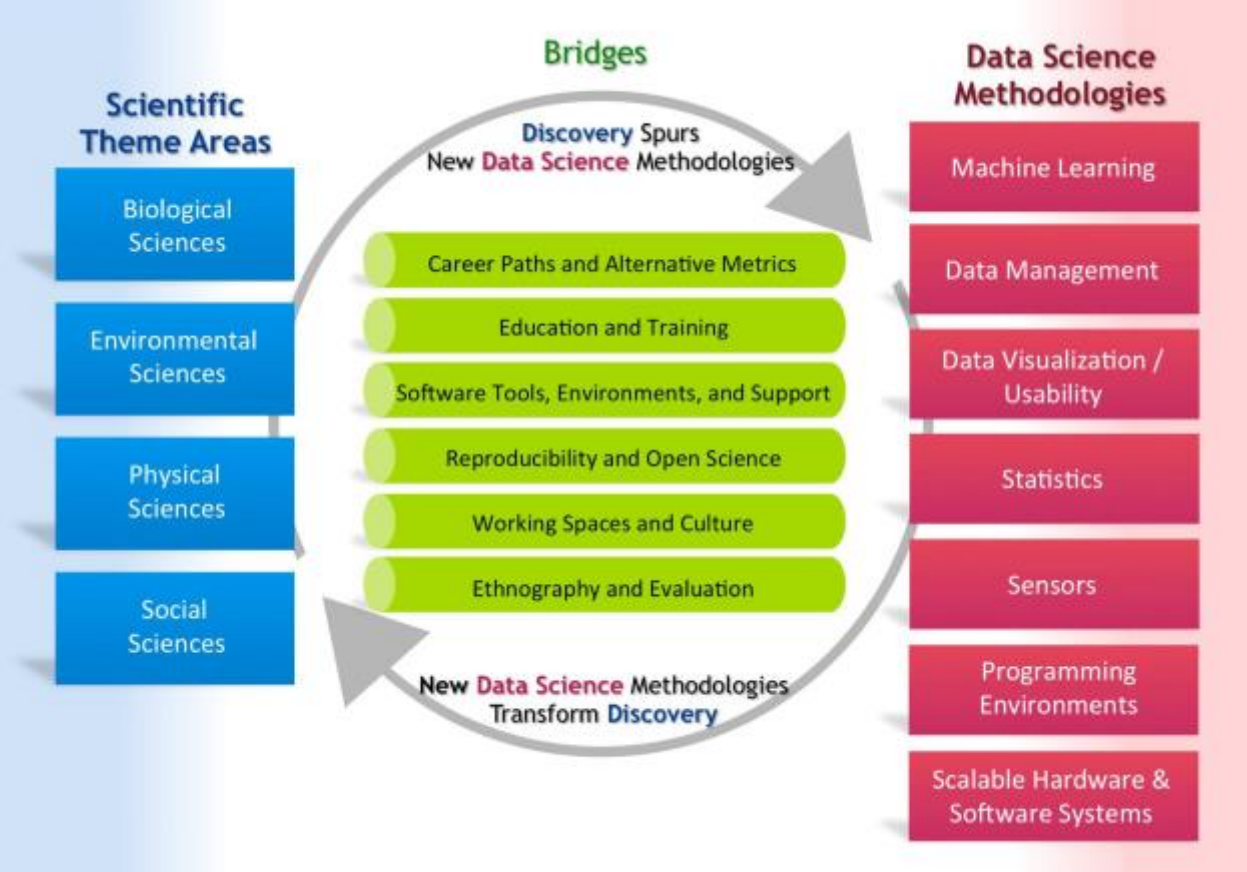
Introduction

Rapid advances in our ability to acquire and generate data have outpaced our ability to extract knowledge from these heterogeneous, noisy, and often massive datasets, creating a significant bottleneck to discovery in nearly every field of inquiry.

Consider the scale and complexity of data sources coming online: simulations of scale and resolution unimaginable only a few years ago (e.g., global climate models, universe-scale n-body simulations), networks of tiny but powerful sensors (e.g., on the seafloor; in the forest canopy; in living organisms; in buildings, roads and bridges), high-bandwidth remote-sensing platforms (e.g., satellites like Terra and Aqua with Moderate Resolution Imaging Spectroradiometers, telescopes used for survey astronomy projects like the Sloan Digital Sky Survey and the Large Synoptic Survey Telescope), high-throughput laboratory instruments (e.g., gene sequencers, micro- and macro-scopic imaging equipment, flow cytometers, mass spectrometers), city-wide urban sensing platforms (e.g., connected vehicles, environmental sensors, ubiquitous cameras), repositories of open government data driven by a new culture of transparency, and social science data created in digital form (e.g., global economic indicators; social network data; consumer activities, including purchasing, mobile phone usage, and internet clickstreams). These advances share a common trait: they produce data with relentlessly increasing *volume*, *velocity*, and *variety*—data that must be captured, transported, stored, organized, curated, accessed, mined, visualized, and interpreted. *Data-intensive discovery*, or *data science*, is a cornerstone of 21st-century discovery.

Working in partnership with one another and with the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation, our three institutions—New York University; the University of California, Berkeley; and the University of Washington—are engaged in a collaborative experiment intended to *transform the process of discovery and the institutional environments in which discovery takes place*: the Moore-Sloan Data Science Environments (MSDSE) effort.

When we began this experiment three years ago, our universities, like so many others, were not well prepared for the data science revolution. The challenge we faced was to go beyond a small number of narrow successes, identifying and tackling a variety of impediments to the broad and sustainable adoption of data-intensive discovery. The impediments that we identified at the outset of our work are shown in green in the illustration below. Our Data Science Environment (DSE) effort seeks to tackle these impediments (in as general a way as possible while recognizing the need to accommodate the unique cultures of each of our institutions) and to establish an ever-accelerating virtuous cycle in which advances in data science methodologies enable advances in discovery, which stimulate further advances in methodologies.



Our collaboration has undertaken the challenge of blazing trails into new *methods*, new *software*, new *partnerships*, new *organizational forms*, and new types of *people* (plus the *institutional change* required to create new *career paths* and new *reward structures* for these people). We are pioneering the development of tools and software environments that are sustainable, reusable, extensible, and translatable across problem domains. We have started new curricula and new degree options for students in data science to empower the next generation. We have leveraged institutional commitments and other funding to hire faculty with deep expertise in data science methodologies and a domain science so they can help lead the

way through teaching and discovery as well as to support targeted data science projects in a broad range of domains ranging from astronomy and high-energy physics to neuroscience and urban science. We have created career paths for professional data scientists who pursue personal research agendas as well as software projects and consulting engagements. We have advocated and built tools to support reuse, reproducibility, and open science. We have created vibrant shared spaces that exist independently of academic departments where research collaborations can flourish across the lines of traditional institutional politics. Lastly, we support data science ethnographers who conduct social scientific research on data science practice and culture and provide reflection and guidance to help us remain innovative and responsive to stakeholders.

Our commitment to each other and to the Moore and Sloan Foundations has been to work in close collaboration and to openly share both successes and failures. The institutes that we have established and the initiatives that we have undertaken are seeds, not solutions. Between us, however, we have acquired considerable experience in what works and what does not work and have developed a deeper understanding of the role different institutional cultures play in determining appropriate approaches. In this paper, we describe key elements we believe have contributed to our successes and also some challenges we have encountered along the way. Our goal is to provide a menu of possibilities for those who may seek to emulate aspects of our DSEs.

Institutes as catalysts within a university landscape

Data science is the great unifier. Its emergence is bringing about a major cultural shift in universities toward cross-disciplinary scholarship—not just between methodology fields and application fields but across methodology fields and across application fields. It depends integrally on faculty lines and research programs embedded in departments, schools, and colleges all across the university. These units are organized in different ways at different institutions, but everywhere, data science cuts across all of them. In a data science ecosystem where collaboration and agility need special support, we have found real value in creating data science institutes bridging different units.

At each of our universities, we have established institutes outside departmental structures that act as both catalysts and leaders, empowering others around campus to embrace data science in education and in research:

- New York University, Center for Data Science (CDS)
- University of California, Berkeley, Berkeley Institute for Data Science (BIDS)
- University of Washington, eScience Institute

These institutes do not pretend to “own” data science. Instead, they bring together researchers and educators from across campus to focus on common goals.

These institutes also serve as strong partners in a network of organizations that spur research and education across the university and as catalysts for larger institutional change. For example, BIDS collaborates closely with the Social Sciences Data Laboratory (D-Lab), the university’s Research IT organization, and the Data Science Education Program. Faculty members’ enthusiasm for creating BIDS was also a major signal to the university’s leadership of the transformational effect of data science across fields. Aligned with other powerful signals, the MSDSE at UC Berkeley has had a catalytic effect in creating the new Division of Data Sciences.

Broad-based, engaged leadership

An essential component to the successes we have realized at each of our institutes is the broad-based leadership of people who have strong reputations; a commitment to the institutional good; and a focus on trust, empowerment, and delegation. This leadership ensures visibility across campus, credibility for new programs and funding opportunities, and neutrality in the political landscape of a university.

Looking at the University of Washington eScience Institute as an example, the 11-person Executive Committee includes the Institute's co-executive directors (a highly successful two-person job-share arrangement—both with PhDs in Oceanography) and nine highly regarded faculty members from eight academic departments spanning four schools and colleges. This core team—which has been stable over a number of years—meets biweekly, and all decisions are made as a group. Strength, breadth, stability, and commitment to the institutional good all are important.

A dedicated space for collaborative activities

Having dedicated space for the DSEs has been important for a number of reasons, including programmatic (space for specific people and activities), intellectual (space for cross-disciplinary collaboration), and political (a space that represents "neutral turf" that is welcoming to all) reasons. Innovative data science is advanced through the creation of high-quality physical spaces that successfully cultivate "the water cooler effect" that used to take place in computer centers during the days of batch processing: the serendipitous exchange of information sparking innovation and collaboration.

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. Physically, libraries are no longer just "Shhhhh!" repositories for printed resources—they are vibrant hubs of activity. On our campuses, 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. See, for example, photographs of BIDS below.

Many design elements were carefully considered to animate our spaces and draw in users. The features that we consider most critical include the following:

- Drop-in workspace for small teams to hold technical discussions and office hours.
- Quiet workspace for permanent staff and in recognition that some people need to be able to escape “cafe noise” to think in quiet.
- Seminar and larger meeting room spaces for events.
- Casual lounge space for ad hoc interaction.
- Hack space for pair-coding, multi-group coding events, or new unknown modes.
- Reconfigurable spaces (e.g., classroom, active learning, roundtable) that include desirable technology, furnishings, amenities, and social conventions.

Professional data scientists

Our institutes are staffed by professionals with the working title “data scientist.” Typically, these are PhD-educated scientists with deep knowledge of data science methodology (one or more of data management, machine learning, causal inference, data visualization, cloud computing, etc.) and of a significant application domain (astronomy, neuroscience, oceanography, sociology, etc.) who have a genuine curiosity beyond their home disciplines and are energized by the opportunity to work across disciplines. These individuals are able to flexibly and nimbly move and adapt methods across domains and problems spaces, learn new tools and approaches quickly, and translate across different disciplinary languages to facilitate successful collaboration. In the aggregate, they form a data science community of practice to keep learning and developing themselves professionally.

Our data scientist positions are intended to be prestigious academic alternatives to faculty pathways—much like the research scientists at supercomputer centers. Our data scientists combine work on their personal research agendas with projects in software; training and education; reproducibility; and collaboration/outreach, including office hours, incubator programs (to be described later), and hackathons. They have been essential to the impact that we have achieved—on campus and beyond.

Even with this success, we have faced several challenges in implementing these data science positions: identifying and attracting individuals who meet our very high standards while asking them to accept a university salary (given the enormous industry demand for data scientists); achieving the right balance between independence/autonomy and organizational activities; offering them enough job security, prestige, and status (e.g., titles reflecting their competencies and abilities); enabling them to initiate and manage grants and contracts (i.e., “principal investigator status”); and ensuring the sustainability of these new careers beyond the MSDSE initiative. Nonetheless, our institutes have succeeded in becoming leaders in attracting and retaining top professional data science talent in a university environment. Instituting “data scientist” as a formal job title in our universities’ personnel systems—with appropriate levels, job descriptions, and salary ranges—is an important work in progress.

We have augmented our full-time professional data scientists with fractional time from research scientists hired by other units on campus who have capabilities of broad value and allow us to “buy” some of their time to participate in various organizational activities. These researchers seek us out because they want to have broad impact and have found that their skills transcend their own field. We have astrophysicists working on homelessness, oceanographic microbiologists working on transportation analytics, etc. These partially funded positions have been phenomenally successful at building our data science community and extending our visibility and impact across campus.

Through our Careers Working Group, we have also learned a lot about the kinds of careers that data scientists would like to have. We have completed in-depth surveys (designed at UC

Berkeley and administered at all three sites) of those associated with our units to learn how people define data science, what they would like to get out of data science positions, and what they see as major impediments to their success. In addition, detailed ethnographic work completed at all three locations has provided a complementary understanding of how people find their way and construct their careers within current DSEs. Based on this research, we know that there are people who want data science careers on university campuses and that they envision many ways to become involved. We are now engaged in the challenging task of determining how to adapt current university rules and regulations regarding ladder-rank faculty members, adjuncts and lecturers, research scientists, and research staff to create rewarding career paths for data scientists. Although we have had successes in developing some new career paths, at the moment, we have a much better picture of the needs than of the solutions.

Outreach: Consulting

Our dedicated collaboration spaces provide a home for a key outreach activity: scheduled, broadly advertised office hours/consulting appointments where researchers can receive tier-one advice on data science challenges they are facing. We have offered these services within the footprint of our institutes together with campus partners, using forms of staffing that build on different constellations of local expertise and institutional resources. Along with helping researchers push the frontiers of domain research forward, these services help remove obstacles that can stymie even the best-informed data scientists and generate positive attitudes toward data science research.

At the University of Washington, for example, office hours are held by the eScience Institute's professional data scientists; data science librarians from the UW Libraries; and consultants for MathWorks, the Statistical Consulting Service of UW's Center for Statistics and the Social Sciences, Tableau Software, UW Information Technology, Amazon Web Services, Google CloudPlatform, and others.

We have everything from shorter-term consulting discussions where problems are solved quickly to intermediate-duration collaborations, longer-term consultations, deep joint research engagements, and joint proposals. Consultants can point researchers in one domain to solutions already known in another, allowing researchers to stand on each other's shoulders across domains. We have also seen the value of thinking carefully about matching forms of consulting to particular populations and needs: for instance, for fledgling data science practitioners in need of encouragement versus experts seeking high-level advice or for guidance on challenges around broadly used packages, tools, and infrastructure versus domain-focused consulting where academically trained staff can address questions about data-intensive research design.

Outreach: Incubator and "machine shop" programs for researchers and students

Increasingly, domain scientists are faced with the challenge of working with large, heterogeneous, and noisy datasets. Many of these researchers have little or no experience with modern data science techniques and technologies. These individuals bring deep knowledge and important intellectual problems from their field, and they are looking for opportunities to innovate, to bridge beyond their current skillset and current disciplinary approaches. They have a problem that needs solving *now*. This situation creates wonderful "teach a person to fish" opportunities that pay multiple dividends.

In response, each DSE has created a version of a program designed to bring together domain researchers and our professional data scientists and research scientists to work on focused,

intensive, collaborative projects of intermediate duration (~10–15 weeks). While we also support shorter-duration interactions (e.g., office hours) and longer-term collaborations (leading to independently funded research projects), we find that there is a sweet spot where we can provide mentoring in data science skills via direct consulting on projects in progress. We provide access to tools and, just as important, expertise about using them, and we create opportunities for researchers to grow. Our goal is that at the end of the collaboration the researcher will return to his or her lab with a solved problem, new data science skills, knowledge of modern software techniques and technologies, and a desire to spread this knowledge among his or her colleagues.

Projects are selected through a lightweight proposal process that ensures feasibility (e.g., availability of necessary data), commitment (e.g., an understanding on the part of the researcher that we are providing a hand, not a handout), and mutual benefit (e.g., enthusiasm for the project on the part of one or more of our data scientists and/or research scientists).

We do not attempt to recover the cost of the consulting and collaboration provided by our data scientists and research scientists. The researchers who participate in our incubator or "machine shop" programs need help *now*. They may not have the funds for a recharged consulting service, and they certainly do not have the time to write a joint proposal. We fund these activities out of our core budget to get research moving and diffuse knowledge. Importantly, our data scientists and research scientists provide free *advice* rather than free *labor*: they offer expertise in state-of-the-art techniques and technologies, but each project lead must be deeply engaged in order to learn by doing and create a real partnership. Hosting multiple collaborations like this simultaneously provides opportunities for beneficial cross-pollination among projects in different fields—the sort of cross-disciplinary water cooler effect discussed in the "Space" section above.

One successful incubator program coming out of the DSEs is the eScience Institute's Data Science for Social Good (DSSG) program. DSSG projects, all of which have an urban emphasis, are sourced from both inside and outside the university and are carried out during the summer. Each project is staffed by four students in addition to the project sponsor (advised by data scientists and research scientists). We highlight this program—adapted from similar programs at the University of Chicago and the Georgia Institute of Technology—because it has generated great interest on the part of project sponsors, student participants, and the press. Topics have included modeling gentrification and equity, assessing the level of automotive cruising (typically, looking for parking) in the urban core, improving transit service using ridership data, assessing the efficacy of various homelessness interventions, route planning for people who use mobility assistance devices like wheelchairs, and many more.

Across our universities, we have noted steadily rising student enthusiasm for incubating data science teams to work on practical projects. Programs like DSSG give students the chance to use the data science skills they have learned in the classroom. In the process, the studies are able to try out experiences they may hope to have in their careers. Whether teams are targeted at scientific research or at institutional and public service, student cohorts look for teamwork, mentorship, and leadership development as key parts of the experience. Student cohorts are likely to keep expanding, and that fact raises critical questions about scaling. Where the research incubator or "machine shop" model is rate-limited by the availability of high-level expertise, a broad-based student program faces other demands: a sufficient supply of good projects, a strong coaching program and quality assurance practices, and possibly a cascading mentorship structure where students gain experience guiding their peers.

Outreach: Software training and support with a focus on open source and harnessing the commercial cloud

We offer broad foundational support for learning, using and building science software and reproducibility tools at each of our campuses. Activities include Software Carpentry workshops and domain-specific hackweeks and hackathons, new courses (see more on education below), brief training events (e.g., Git/GitHub), and software journal clubs where research developers from across campus gather to discuss published software and conduct code reviews (similar to traditional journal articles in domain sciences). In addition, our staff make significant contributions and provide leadership on a number of globally utilized software projects.

Our engagement with Software Carpentry, in particular, has strengthened our ties to the campus. Several times each year, we run Software Carpentry events in both R and Python, and the demand continues to be high. To keep up with the demand for qualified instructors, we offer Software Carpentry instructor training courses, which expand our capacity to offer local events unilaterally. Software Working Group meetings have been an opportunity for our community to gather and participate in a wide range of activities, from discussions to talks to group code reviews, with a rotating lead drawn from our team of data scientists.

We are firmly committed to the open source software movement. Software that is freely available and openly developed lies at the core of data science, both in academic research and in the technology industry. However, it is widely acknowledged that this essential infrastructure is particularly difficult to maintain and sustain. In particular, academic incentive structures tend to encourage the creation of new and novel software systems (if they encourage software activities at all!), but we believe supporting, maintaining, and incrementally developing existing libraries better serves the community of researchers/ software users.

We are deeply engaged with several communities that use and maintain open source software for data science. BIDS is home to the Jupyter project and ROpenSci, which are central to research applications using the open source languages Python and R, respectively, and CDS is one of the hubs of development for the Julia language. Faculty, students, and data scientists across our three centers lead and contribute to myriad discipline-specific software (in fields ranging from economics to astrophysics). An interdisciplinary group with ties to many different software communities serves as an excellent environment for the development and diffusion of common practices around the testing, documentation, and distribution of open source software, where usually no other university-based organization has clear incentives to engage. Centers that take on and incentivize these activities (“ecosystem gardening”) can accelerate the development and use of software and help mitigate the risks inherent in a focus on essential but less glamorous software infrastructure work.

Data science relies far more on *intellectual infrastructure* (new methods, new tools, new partnerships, new people) than on physical infrastructure (local compute services—inevitably subsidized and promoted by institutional leadership). Encouraging researchers to harness commercial cloud services from Amazon, Google, and Microsoft where appropriate has been an important focus. There is far more to say on this topic than there is space for here, but it is worth noting that the University of Washington has eliminated a significant and nonsensical cost distortion by waiving indirect cost on outsourced cloud services. UC San Diego has followed suit, and other UC campuses are expected to do the same.

Outreach: Seminar series and other outreach activities

Our institutes provide a wide variety of seminar series. Because participants come from diverse disciplinary backgrounds and technical training, topics span science, methods, and technology across the mission of our institutes. As an example, one of our seminar series is focused specifically on reproducible and open research, with monthly presentations that include invited speakers working on tools for incentivizing and facilitating reproducible and open research, developing portals for sharing code and data, or launching open science journals. We also host broad data science topics in both formal and informal settings. Less formal seminar series provide a venue for graduate students, postdocs, and local faculty to give presentations that encourage discussion among the group. More formal university-wide seminar series bring distinguished speakers from across the nation to discuss topics related to data analysis, visualization, and applications to domain sciences. These high-profile speakers draw large crowds and provide an opportunity to expand our data science campus community and outreach.

In addition, we regularly reach out to the greater campus community with open events and opportunities designed to increase awareness of data science and our institutes on campus. For example, the UW eScience Institute runs an annual all-campus data science poster and networking session; with >100 posters presented every year, this two-hour event is an opportunity for the campus community and regional partners to present their activities and connect with others engaged in data-intensive discovery. The UW eScience Institute also runs an all-campus Data Science Recruiting Fair where companies interview undergraduates, graduate students, and postdoctoral fellows with data science expertise. BIDS has had ongoing success with its Data Science Faire and is now partnering with other organizations to expand the event. BIDS has also hosted large student poster events for the Data Science Education Program.

Engagement: Designation of “data science fellows”

The title “data science fellow” is given to highly engaged members of our data science communities who are recognized for accelerating the development and/or application of new methodologies as well as for their participation in various institute programs and working groups. (To clarify—“data scientists” are professionals that we employ, “data science fellows” are researchers from across the campus.) The goal of this designation is to build community, increase engagement, expand visibility, and provide a stamp of data science authority to researchers in a broad range of fields. By making this engagement explicit on their CVs, web pages, and public presentations, these individuals gain recognition for their data science expertise, and our institutes gain visibility. The data science fellow designation began at Berkeley, met with great success, and was adopted by NYU and UW.

Engagement: Accelerating data science–savvy faculty hires

As field after field transitions from data poor to data rich, grounding in both data science methodology and a significant application discipline is likely to be an increasingly valuable attribute for faculty members and for graduates at all degree levels.

At the faculty level, we have developed a variety of protocols to influence hiring priorities and evaluation criteria. This is difficult, and while we are making a start, significant change will take time. Change is possible, though. For example, it is becoming common to consider citation counts and technology transfer when evaluating research impact and to give full weight to collaborative efforts. Widely used innovative software, research that is open/reproducible, and

high-impact datasets must join the list. Exceptional appointments can serve as beachheads and signals. The appointment in 2017 of Fernando Perez (a BIDS Fellow) in a tenure-track faculty line in the Department of Statistics at UC Berkeley indicates that change is possible in far-sighted departments whose faculty advocate for new criteria of merit.

At the University of Washington, the Provost provided permanent 50% funding for a number of faculty positions to incentivize the recruiting of candidates who meet this profile. A detailed MOU is associated with each of these appointments, but in general, these faculty devote 50% of their teaching to data science courses of interest beyond their home department and 50% of their service to campus data science outreach, each overseen by the leadership of the eScience Institute. A representative of the eScience Institute contributes to their annual reviews and promotion considerations (which includes advocacy for forward-looking assessment criteria). Under this program, truly extraordinary faculty have been appointed in applied mathematics, astronomy, biology, computer science and engineering, mechanical engineering, sociology, and statistics. These appointments create and highlight an interdisciplinary core of superb faculty (and faculty mentors) engaged in data-intensive discovery, help build strong ties across disparate campus units, and make important educational and outreach contributions. Participating departments, in turn, are happy to have gained such incredible talent and are willing to push the agenda of data science.

At UC Berkeley, the Provost accepted the 2016 recommendation of a Data Science Faculty Advisory Board to invest roughly two dozen faculty lines in data science appointments, including a broad range of application domains as well as computing and statistics. These appointments will be made in consultation with departments and deans across the university, with a leadership role and coordinating function given to the new Dean of Data Sciences.

Engagement: Postdoctoral programs

Our postdoctoral programs are designed to prepare the next generation of data science researchers in a broad range of application domains. The specifics vary at each of our institutions, but here is the general idea:

- We recruit locally and nationally to attract outstanding interdisciplinary researchers to these positions, researchers with expertise in the methods of data science and in a physical, life, or social science.
- We leverage multiple funding sources, and we have named the postdoctoral awards in accordance with these funding sources, increasing the “CV status” of these positions. For example, UC Berkeley has developed partnerships with domain sciences, with industry, and with other nearby universities (e.g., UCSF) to fund data science fellows, and the University of Washington received substantial support from a local foundation for postdoctoral fellows.
- The postdocs can have mentors from multiple fields—for instance, one mentor in a methodology field, and one in an application field. (It is worth noting that these dual-mentoring situations, while potentially of great benefit, work best when the two faculty members have a pre-existing relationship).
- We offer a variety of programs to build a scientific community—for example, weekly lunches with a program component.
- We offer many opportunities for the postdocs to build their skills for success on the job market, such as career fairs and small-group discussions with faculty on how to land a job.
- We encourage our postdocs to host speakers in our high-profile seminar series. The postdocs gain one-on-one time with these distinguished visitors as well as campus-wide

visibility. More recently, we have added Distinguished Young Academic Data Scientist (DYADS) speakers to this seminar series. These are invitation-only postdoctoral-level speakers who have been selected by the executive directors at each of our institutes for their outstanding research and leadership in the data science community. In addition to a high profile speaking engagement, this program gives them an invited talk on their CV and the opportunity to network at a different institution.

- We have found that best practice involves taking organizational responsibility for our postdocs and ensuring strong mentoring and support while recognizing that they occupy a middle ground in between the nurturing environment provided for graduate students and the independent operator status of faculty members.

Formal educational programs

Prior sections have described a variety of approaches to outreach, engagement, and informal education: office hours, intermediate-duration collaborations, software training and support, seminar series, poster sessions, and others.

Here, we discuss approaches to formal education: for-credit courses and degrees. Across our institutes, we have pursued a variety of approaches, influenced by institutional context and faculty interests. Thus, this section is considerably more detailed than others—"Your context may vary!" All of the approaches we describe have met with success: there is tremendous demand for data science education, and context matters!

We begin by enumerating some common themes and then discuss some local details.

Content: There are some obvious components of a data science curriculum: statistics and inference, machine learning, data management, scalable/cloud computing, data visualization, and ethics (including security, privacy, and human contexts of practice). These can be taught in the abstract, or in the context of real scientific data and its analysis.

Depth of coverage: The required depth of knowledge differs for individuals who need to utilize data science techniques and technologies in an intelligent way in their work, for individuals who seek to become practicing data scientists, and for individuals whose career will involve advancing the forefront of data science techniques and technologies. For all of these categories, the various methodology disciplines (computer science, statistics, etc.) can recognize that one can succeed and indeed thrive as a data science practitioner without the equivalent of a full degree in any of these fields.

Appropriate courses: It is tempting to believe that existing courses can be assembled into an effective curriculum. Often this is not the case. For example, a data management course designed for a computer science major—which, in caricature, focuses on how to build data-management systems—is not appropriate for a data science curriculum. Even something targeted at novices, such as introduction to programming, is best served differently when the target audience is future computer scientists versus data scientists. Additionally, there can be huge gains from integrating material together in new courses, acknowledging that computation is changing the very way we think about statistical inference, for instance. Some data science application domains have a strong interest in offering courses that adopt or inflect data science methodologies for their own fields.

As noted above, institutional context has influenced the choices that each of us has made. Here is a highest-level abstraction (many details elided!) of the approaches each of us is pursuing.

The University of Washington: At UW, our curricular offerings are guided by the philosophy that at the bachelor's and PhD levels, data science must become integrated pervasively into existing degree programs instead of being a stand-alone degree. (Master's programs are different; master's degrees in data science make a great deal of sense given industry demand.) UW's data science education efforts differ from Berkeley's in this regard—the MSDSE project is truly a distributed collaborative experiment.

We have taken advantage of a pre-existing UW designation called a "transcriptable option"—a specialization (lighter weight than a minor) that is recognized on the student's transcript. A student who successfully completes the courses required for our transcriptable option receives a notation on his or her transcript—for example, "Bachelor of Science in Biology, Data Science Option." We have three such options—two at the graduate level and one at the undergraduate level. At the graduate level, the Advanced Data Science Option is available to students who complete a small number of intensive courses in key data science areas (statistics, machine learning, data management, data visualization, and ethics/privacy), and the Data Science Option is available to students who complete more introductory-level courses in these areas. In both cases, the courses must offset courses otherwise required for the student's degree (i.e., "instead of" rather than "in addition to"). In both cases, students are required to participate in the eScience Institute's inter-departmental seminar series to help them become integrated with the broad data science community on campus. The undergraduate-level Data Science Option is similar but with greater flexibility in how participating departments choose to cover the various key data science areas. A separate professional Master of Science in Data Science is offered jointly by a coalition of six departments that the eScience Institute assembled: Applied Mathematics, Biostatistics, Computer Science & Engineering, Human Centered Design & Engineering, the Information School, and Statistics.

In addition to the courses that serve the above programs, we have developed multiple specialized courses at the graduate and undergraduate levels; examples include "Data Science for Biologists," "Big Data and Population Processes," and "Software Development for Data Scientists." Emulating our Berkeley colleagues, we have recently developed a pair of large undergraduate introductory courses on "Data and Society" and "Introduction to Data Science" in addition to expanding our introductory "Data Programming" sequence. Our goal is to expose a majority of UW undergraduates to data science early in their college careers, complementing the more in-depth upper-division Data Science Option.

We find that many of the new courses that we develop in support of data science are best offered jointly by different combinations of departments. A highly visible example is the course "Calling Bullshit: Data Reasoning for the Digital Age," co-developed and co-taught by faculty from Biology and the Information School.

UC Berkeley: Berkeley's leadership has invested in developing a broad-based, comprehensive data science curriculum from the ground up. Our view is that data science is revolutionizing many fields and professions, and it is incumbent on the university to make it broadly available and integrated into the fabric of the undergraduate experience. In fact, at the foundational level, data science thinking needs to become part of the mindset of an educated citizen of the 21st century. Berkeley's Data Science Education Program starts with a broadly accessible freshman-level class on the "Foundations of Data Science," familiarly called Data 8, which is taught hands on in Jupyter Notebooks with real-world examples and is projected to enroll 2,000 students in 2017–2018. The Foundations class is accompanied by several dozen entry-level "Connector Courses" that bring the Foundations material into direct contact with a wide range of specializations (sciences, engineering, social sciences, and humanities as well as computer

science and statistics). All these courses were newly created in the space of two years and amount to the fastest growing program in the university's history. Their course materials are freely available online and leverage a common computational platform that can serve several thousand students at once.

With the data science curriculum continuing to build from the bottom upward, faculty have now created new integrative "gateway" courses at the junior level, which have begun feeding students into pathways into a broad range of data-enabled courses in application domains. Berkeley is in the process of creating data science major (BA and BS) and minor programs, which will extend students' learning in computational and inferential depth, integrated domain emphases, and human contexts and ethics. We hope to make these available by the end of spring 2018 and expect the demand will be large. The program also provides intensive data science pedagogical training for faculty across campus and assists them in creating data science "modules" to embed in their own courses.

At Berkeley, we have taken the view that data science has strong elements of analogy to computer science emerging as a new discipline as well as computational science emerging as a form of practice. Our local ecosystem may lead us to put a somewhat stronger emphasis than the other MSDSE partners on the conceptual foundations of data science in addition to its applications to science. We see that there is an important place filled by a stand-alone bachelor's degree as well as pervasive integration (e.g., in the form of a minor), and we expect that both modes will quickly become among the largest academic programs at Berkeley.

NYU: CDS has focused on creating a professional Master of Science in Data Science. For most students, this program consists of four semesters of instruction and includes five required courses, a capstone research project, and a series of electives. The required courses are: an introduction to the discipline of data science, mathematical statistics, machine learning, big data (focusing on the algorithmic and systems aspects of operating on data at massive scale), and inference. The electives most often taken are those offered by our faculty who are jointly appointed with other departments and include social science, neuroscience, and computer science courses. Ethics is currently included in the introductory course, but given its centrality to the practice of data science, a stand-alone semester-long ethics requirement is being considered.

For the 2017 intake of students, we introduced a set of "tracks" of instruction, which are essentially specializations. These tracks (currently in big data, natural language processing, mathematics and data, and physics) guide the choice of elective courses. (There is a general track for those who do not want to specialize or want to craft their own specialization.) Since introducing tracks does not require high-level approval, their great advantage is their flexibility: we can create tracks quickly in response to demand, and it is relatively straightforward for students to switch between tracks. We will add a biology track in 2018 and ultimately expect to add more. (The master's track in physics offered an interesting lesson: Many domains are hesitant to embrace data science and are "policing their boundaries." At NYU—in contrast to the University of Washington—a master's in physics with a concentration in data science would not have been acceptable, but a master's in data science with a concentration in physics was easy to push through. Things may change in the future, but in the current climate, there was a huge difference in reactions and attitudes toward the two approaches. An example of the role of institutional context!)

The NYU Master of Science in Data Science, while of the highest academic quality, is aimed at making students maximally employable as data scientists in industry. We pay substantial

attention to commercial and “real-life” applications that will aid students later in their careers. By this metric, we are highly successful: our graduates are widely sought by firms in finance, social media, tech, online retail, the non-profit world, etc.

While we are delighted that our master’s program has proven so popular, this is not without challenges. Simply processing nearly 2,000 applications for approximately 100 slots in a cohort presents administrative difficulties. Our courses are invariably over-subscribed and have long wait-lists as there is heavy demand from other NYU students. Serving the data science education needs of this broader NYU community is absolutely a goal, but one that we have not yet achieved. We do now have the “Introduction to Data Science” course in which undergraduates can enroll with permission and expect to do more in the future. CDS also enrolled an initial cohort of five PhD students in 2017. NYU, in contrast to UW and UC Berkeley, has chosen offer a doctorate in data science.

Encouraging best practices in reuse and open science

Our three DSEs share a strong commitment to promoting best practices in reuse and open science: creating reusable workflows, encouraging the development of open source tools, openly sharing data as appropriate, and making published data science research reproducible. This has been a cornerstone from the beginning of the MSDSE effort, with active working groups on the topic at each institution and active collaboration across institutions. A demonstrated commitment in this direction has been a selection criterion for data scientists, research scientists, postdocs, and data science fellows on each campus. As a result, there is now a highly developed culture of openness and reuse that permeates all activities, including office hours, training and curriculum development, hackathons and summer programs, etc. Reproducibility workshops have been used to encourage broader engagement.

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.

A book of case studies that we have assembled, *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*, is available online and will be published by the University of California Press in 2018.

Ethnography/Data Science Studies

Our DSEs seek to change the process of discovery. Instrumenting this process has been integral to our effort. Data Science Studies (DSS) is our collaborative ethnographic effort to study the emergence of data science as an increasingly central model of discovery in academia. In close partnership, our institute ethnographers and broader DSS teams conduct scholarship that documents the social practices, cultural production, and institution building of the data science community. The teams engage in programs central to our missions, such as the Incubator Program, providing insight and feedback to our leadership as well as generating formal scholarly publications, informal publications (blogs), and pedagogy through workshops and presentations. DSS seminars center around a particular theme and have included: data privacy; transparency, seamful design, and data science; emerging models of data science in academia and their disciplinary tensions; data science and pedagogy; and democratization of data science. These seminars bring together people researching, wrestling with, and reflecting on issues of data science from a variety of perspectives. The DSS working groups have been

invaluable to our institutions as we develop, adjust, and maintain the various programs that support our mission of advancing data-intensive discovery in all fields.

Diversity, equity, and inclusion

Because data science is a new field in the university, we have the opportunity to establish a new culture that addresses the problems of diversity that afflict many of the fields upon which data science draws as well as well-publicized experiences of exclusion and harassment in tech. For example, our institutes and the meetings that we organize have codes of conduct and have adopted various best practices for promoting diversity, equity, and inclusion in inward- and outward-facing ways.

We have found that efforts on behalf of equity, inclusion, and diversity are most effective when they are rooted in explicit public commitments by institutional leadership and are overtly aligned with core university values. Along with paying attention to demographics and numerical representation, these efforts need to be enacted in deliberate, daily practice of respect for difference. In addition, universities are the educators of new generations of data scientists. They have a major opportunity to shape the field by creating data science programs that embody the values of diversity as well.

Sustainability

Our DSEs were bootstrapped by multiple funding sources—most obviously the Moore and Sloan Foundations but also other foundations; federal agencies (e.g., NSF); and, in some cases, modest institutional funds. We have been conducting a distributed collaborative experiment in supporting, diffusing, and accelerating data-intensive discovery. The results of this experiment—the things we are confident have high impact—are described above. The ongoing cost of sustaining these activities is a few million dollars per year—truly a tiny amount in the context of universities that provide central subsidies of $50–100 million for computing services, $50–100 million for libraries, etc.

As with computing services and libraries, the leverage of our DSEs is enormous: they provide intellectual infrastructure that changes the process of discovery in a broad range of fields, enabling our institutions to remain leaders. We have been working with our computing services organizations, our libraries, our provosts, our vice provosts for research, and our deans to ensure that they appreciate the value of our activities and the natural fit of these activities with their various missions. We are optimistic that they will embrace institutes like ours as essential activities of the institution. Indeed, this is well underway; for example, UC Berkeley has established the Division of Data Sciences.

A short summary of successes, lessons, and challenges

<u>Space and governance</u>

- Space and governance are fundamental elements of a successful data science environment.
- Both must create a perception of neutrality, openness, and willingness to be helpful.
- Space should be central, flexible, open, and welcoming. Space in libraries is often especially appropriate given libraries' central physical and intellectual positioning as well as their natural role in data curation.
- Steady-state governance structures for data science within universities, following initial bursts of activity and enthusiasm, can be challenging. These structures need to accommodate a complex multi-school mission with many "department-like" features

(e.g., teaching, research, outreach, faculty affairs [hiring, tenure, and promotion]) with a scope that inherently spans departments, schools, and colleges.

Careers

- Professional data scientists and research software engineers have been central to our success. Recognition by campus leadership of the broad impact of these individuals on research, education, and outreach has provided a strong argument for the institutional sustainability of our efforts.
- There is a strong interest in data science careers inside academia. We find that in nearly every domain, interest in data science is strong and is growing rapidly.
- We have identified many of the needs and impediments to creating these careers and have had success in creating some new career paths.
- Even more than we might have anticipated, data science career tracks have opened up inside many of the domain sciences. These seem likely to expand as important sources of employment for data scientists within the academy.
- Many of these careers have a hybrid industry/academic flavor. While there is intense competition from industry, roles in academia—in accelerating discovery—have proven to be attractive to many.
- At the faculty level, we have significantly accelerated the recruitment of scholars who simultaneously innovate in data science methodology and lead in some domain science.
- Progress has been slow but steady in convincing promotion and tenure committees to reward contributions like the creation of widely used software and datasets, adhering to best practices in reuse and open science, etc.

Consulting in various forms

- We have developed a spectrum of useful models for providing data science consulting, including drop-in office hours, tutorials and short-courses, hackweeks, incubator and "machine shop" programs, and long-term collaborative research engagements.
- We do not have a clear picture of the relative effectiveness and cost effectiveness of the different models.
- We have not implemented recharge/reimbursement mechanisms for shorter-term consulting engagements; very low barriers to engagement (office hours, incubator and "machine shop" programs, etc.) have proven to be important.
- In some cases, consulting activities have led to the resolution of governance and sustainability issues.

Educational programs

- We find great enthusiasm and demand for data science education at all levels. (Data 8, for example, is the fastest growing course in UC Berkeley's history.) For this reason, any new program needs to have a plan for scalability. Staffing, space, and computing support can be significant challenges depending on institutional policies for supporting rapidly expanding educational initiatives.
- Freshman-level introductory courses are extremely popular. Having multiple departments jointly develop these courses can help spread the teaching load, bring multiple perspectives, and avoid turf wars. (Joint data science educational initiatives help create connections across campus, which can in turn help build research collaborations and lead to joint hires.)
- Upper-division and graduate-level data science specializations within existing programs garner significant buy-in and excitement and are relatively easy to deploy.
- Data science master's programs are extremely popular and can easily be made self-sustaining.

- Computational infrastructure for teaching is a *sine qua non*. Jupyter Notebooks are brilliant for pedagogy. Zero-to-Jupyterhub and Kubernetes allow management and scaling. (Educational initiatives provide a platform for disseminating best practices in software engineering, reuse, and open science.)
- Informal and non-traditional initiatives, such as seminars, poster sessions, software/data workshops and short courses, hackweeks, etc., have proven to be highly popular, and we recommend them highly. Our fellows, professional data scientists, and research software engineers have been particularly helpful with these initiatives.

Reuse and open science

- We have adopted a multi-pronged approach to encouraging best practices in reuse and open science, including tools, incentives, and training.
- Software infrastructure to streamline the creation of reproducible results is emerging; the user base for this infrastructure is growing steadily.
- Mechanisms that recognize/reward research exemplifying best practices (e.g., badging, formal recognition at conferences, etc.) are proliferating

Involving a wide range of people in data-intensive discovery

- We have developed a wide range of approaches to outreach and engagement, all of which have proven to be valuable. These include consulting services with many timescales, traditional and informal/non-traditional educational initiatives, our data science fellows programs, and dual mentoring of graduate students and postdocs (one faculty mentor from a domain science and one from a methodology field).
- Cost-effectiveness and sustainability are issues for a number of these approaches.
- As a new field, data science provides the opportunity to establish a culture of inclusion—beginning in introductory courses and extending all the way up the stack.

Links to specific programs

New York University: http://cds.nyu.edu/

- Contacts for additional information:
  - Rich Bonneau <bonneau@nyu.edu>
  - Juliana Freire <juliana.freire@nyu.edu>
  - Mik Laver <ml127@nyu.edu>
- Lunch Seminars: https://cds.nyu.edu/data-science-lunch-seminar-series/
- Text as Data: https://cds.nyu.edu/text-data-speaker-series/
- Showcases: https://cds.nyu.edu/news-listing/data-science-faculty-seminar-series/
- Newsletter: https://cds.nyu.edu/newsletter/
- Seed Grants: https://cds.nyu.edu/nyu-data-science-seed-grant/
- Master of Science in Data Science: https://cds.nyu.edu/academics/ms-in-data-science/
- PhD in Data Science: https://cds.nyu.edu/academics/phd-in-data-science/

UC Berkeley: http://bids.berkeley.edu/

- Contacts for additional information:
  - Henry Brady <hbrady@berkeley.edu>
  - Cathryn Carson <clcarson@berkeley.edu>
  - David Culler culler@berkeley.edu
  - Jonathan Dugan <jmdugan@berkeley.edu>
  - Marsha Fenner <mwfenner@berkeley.edu>
  - Saul Perlmutter <saul@lbl.gov>

- Lectures and events: https://bids.berkeley.edu/events
- Videos: https://bids.berkeley.edu/resources/videos
- Cross-Domain (XD) collaborative groups:
    - ImageXD: https://bids.berkeley.edu/research/image-xd
    - TextXD: https://bids.berkeley.edu/research/hacker-within
    - GraphXD: https://bids.berkeley.edu/research/graphxd
- Contributions: https://bids.berkeley.edu/contributions
- Machine Shop (Incubator): https://bids.berkeley.edu/research/bids-machine-shop
- The Hacker Within (peer-driven skill sharing): https://bids.berkeley.edu/research/hacker-within
- Data Science Education Program: http://data.berkeley.edu/education
- Foundations of Data Science class (Data 8): http://data8.org/
- Division of Data Sciences: http://data.berkeley.edu/

University of Washington: http://escience.washington.edu/

- Contacts for additional information:
    - Magda Balazinska <mbalazin@uw.edu>
    - Ed Lazowska <lazowska@uw.edu>
    - Micaela Parker and Sarah Stone <exec-director@escience.uw.edu>
- Data Science Studio: http://escience.washington.edu/wrf-data-science-studio/
- Incubator: http://escience.washington.edu/get-involved/incubator-programs/overview/
- Data Science for Social Good: http://escience.washington.edu/dssg/
- Hackweek examples:
    - Astro: http://astrohackweek.org/2014/
    - Neuro: https://neurohackweek.github.io/
    - Geo: https://geohackweek.github.io/
    - Image XD: http://www.imagexd.org/
- Office Hours: http://escience.washington.edu/office-hours/
- Seminars: http://escience.washington.edu/get-involved/
    - Data Science Seminar: http://escience.washington.edu/uw-data-science-seminar/
    - eScience Community Seminar: http://escience.washington.edu/get-involved/escience-community-seminar/
- Working groups: http://escience.washington.edu/working-groups/
- Education: http://escience.washington.edu/education/
    - Undergraduate Data Science Option: http://escience.washington.edu/education/undergraduate/
    - Graduate Data Science Options: http://escience.washington.edu/education/phd/
    - Data Science Master's: https://www.datasciencemasters.uw.edu/
    - Tutorials and Bootcamps: http://escience.washington.edu/education/tutorials-and-bootcamps/
    - MOOCs: http://escience.washington.edu/education/mooc/
    - Professional Certificate: https://www.pce.uw.edu/certificates/data-science