Academic Data Science Centers in the United States

A Study of 20 Universities

December 2018

Prepared for:

Joshua Greenberg, PhD
Alfred P. Sloan Foundation

Chris Mentzel
Gordon and Betty Moore Foundation

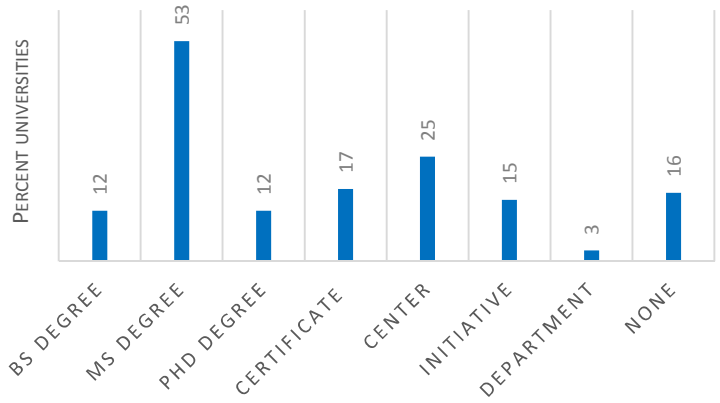
Prepared by:

Luba Katz, PhD
Abt Associates

INTRODUCTION

Data science is an approach to scientific inquiry that uses computational, statistical, mathematical, and domain-specific tools to extract knowledge from a large volume of data. Its potential for discovery is increasingly recognized by the academic community, as illustrated by the growing number of new degree programs, research centers, initiatives, and departments in the United States. A recent review of websites for 116 research universities revealed that more than 80% included at least one data science offering (Exhibit 1).

Exhibit 1: Data science offerings at research universities (N=116)



Source: Moore Foundation, 2017.

This report describes the mission, organization, and activities of data science centers, initiatives, and departments at 20 of these universities (Exhibit 2). The initial set of institutions to include was recommended by the Sloan and Moore Foundations, which funded three of the data science centers through its Moore-Sloan Data Science Environments (MSDSE) program.[1] The sample was expanded with suggestions from the participants. Because of this ad hoc sampling strategy and some level of non-response, this report is neither representative nor inclusive of all efforts in data science, but rather is an attempt to capture a range of models being explored by universities.

For 17 of 20 centers, the information presented is based on three sources: (a) one hour telephone interviews with the leadership of the centers conducted between December 2016 and June 2018, (b) review of the center websites and materials provided to us, and (c) a short survey of participants in the Data Science Leadership Summit held in October 2018. For the remaining three centers, at the New York University, the University of California Berkeley, and the University of Washington, we collected extensive additional data as an external evaluator of the MSDSE program.

In the next section, we describe the organizational models, programs, and activities at 20 universities, illustrated with a few specific examples. Following, we include short profiles of each site, which contain

[1] Moore-Sloan Data Science Environments: http://msdse.org/

additional information.[2] We caution the reader that in our experience data science entities are rapidly evolving, and therefore some data presented in this report may be out of date.

Exhibit 2: Centers included in the report, ordered alphabetically by the university

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

*Note: *Moore-Sloan Data Science Environments.*

[2] All interview respondents were offered an opportunity to correct and update their profiles.

SUMMARY OF FINDINGS ACROSS THE CENTERS

Mission, leadership, organization

We found that 17 of 20 centers were formed within the past five years and the remaining three (at BU, UChicago, and UW) stemmed from or extended pre-existing entities. The creation of the centers was motivated by the growing interest in data science among the faculty and the perceived need to connect and/or boost existing programs. Several respondents recalled an elaborate planning phase, which lasted for several years and involved gathering input from dozens of faculty and administrators. In fact, when asked to share any "lessons learned" about establishing a center, the most commonly mentioned was the importance of engaging the university community in the design process. Several respondents said that navigating the university's political landscape and persuading the faculty that they would benefit from the data science center were their greatest challenges.

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

Review of the center mission statements revealed a shared emphasis on collaborative and interdisciplinary research. Some centers also articulated their commitment to education/workforce development and to societal benefits. Exhibit 3 is a word cloud generated using the mission statements, which picked out the commonly used terms such as "education, research, science, interdisciplinary, methods, and data."

Exhibit 3: Word cloud of mission statements (N=20)

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

All of the centers are led by a Faculty Director (two Co-Directors at Harvard University and the University of Michigan), and 9 of the 20 also include a non-faculty Executive Director (Exhibit 4). Most centers are overseen/guided by faculty executive committees.

Exhibit 4: Center organization and participants

N	Center	Year launched	Space	Based in single departm/ college	Non-faculty managing director	Faculty lines	Data scientists	Postdocs
1	BU DSI	2012/14	✓		✓	✓	✓	✓
2	Caltech CD^3 JPL CDST	2015	✓				✓	
3	Columbia DSI	2012	✓			✓	✓	✓
4	Duke iiD	2013	✓				✓	✓
5	Harvard DSI	2017	✓					✓
6	JHU IDIES	2012	✓			✓	✓	
7	MIT IDSS	2015	✓		✓	✓	✓	✓
8	MSU CMSE	2015	✓	✓	✓	✓		
9	NYU CDS*	2013	✓			✓	✓	✓
10	NW DSI	2015			✓	✓		✓
11	OSU TDAI	2015	✓		✓	✓		
12	Stanford DSI	2014			✓			
13	UC Berkeley BIDS*	2013	✓		✓		✓	✓
14	UChicago CDAC	2000/18	✓					
15	UMass CDS	2015		✓		✓	✓	✓
16	UMichigan MIDAS	2015	✓		✓			✓
17	UNCC DSI	2012	✓			✓	✓	
18	URochester GIDS	2013	✓			✓	✓	
19	UVA DSI	2013	✓			✓		
20	UW eScience*	2008	✓		✓	✓	✓	✓

**Moore-Sloan Data Science Environments.*

In a testament to the interdisciplinary nature of data science, virtually all entities are administratively based outside of any one department or school. The two exceptions are the University of Massachusetts (U Mass) and the Michigan State University (MSU). The Center for Data Science at U Mass is housed within the College of Information and Computer Science. This choice was intentionally made by the leadership of the center in order to retain control over student training within the College and to simplify/accelerate decision-making.

MSU launched a new department, called Computational Mathematics, Science, and Engineering (CMSE), which is administered jointly by the College of Natural Sciences and the College of Engineering. We were told that the planning committee conducted some research and concluded that data science centers at peer institutions were short-lived. Consequently, the choice was made to start a new department as a more permanent solution. Our respondent acknowledged that significant costs to create and maintain a department made widespread support for it difficult to secure. In addition, a department could be perceived as more insulated than a center, potentially limiting faculty engagement. But in his view, the benefits of a department outweighed these limitations.

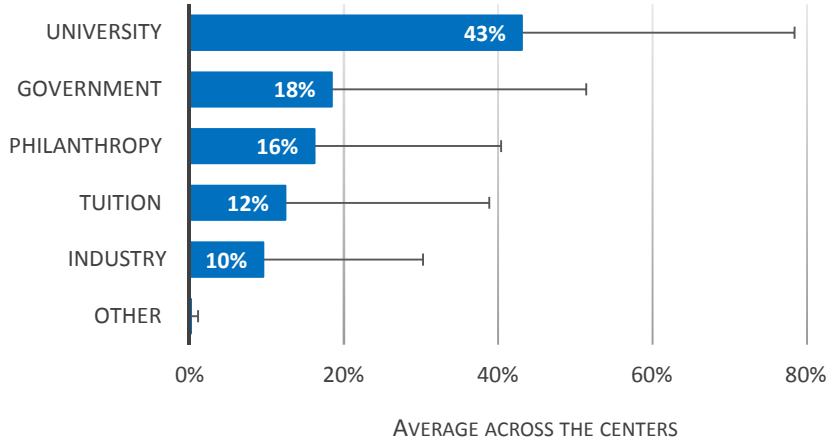
Space and funding sources

All but three centers had secured dedicated space (Exhibit 4), which ranged in size from a few meeting rooms and/or faculty offices to large portions of a building. The spaces were described as open, multi-purpose, configurable, vibrant, and collaborative. In an interesting example at UVA, faculty, staff, and students reside together in an open space, in keeping with the original "academical village" concept of Thomas Jefferson who founded the university. In several cases, the spaces were newly built or renovated with large contributions from private donors or local governments. Several center directors were looking to move to accommodate their growing communities. The centers at Stanford, U Mass, and JPL are "virtual," which does not impede their function.

To collect more systematic data on how the centers are funded, we supplemented qualitative interview data with a survey of data science leaders who attended a recent summit. We found that the centers piece together funding from multiple sources to support their operations and programs (Exhibit 5). Across the board, the university provides the bulk of the funding, at 43% of the total. The rest is divided between the government, philanthropic organizations, tuition, and industry. We note large standard deviations for each funding source which indicate significant variability across the centers.

Exhibit 5. Funding sources to support the centers (N=28)

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



Note: Survey and interview respondents are partially overlapping groups.

We found in the interviews and by reviewing websites that in some cases the investments by university were quite large: $125 million at OSU (which included the hiring of approximately 70 faculty), $20.5 million at U Michigan, $50 million at U Rochester, and $25 million at JHU. In most cases, universities provided funding for the initial period, on the order of 5 years, after which the centers were expected to become self-sufficient. However, U Mass and UW had committed to support their data science centers for the foreseeable future. The leadership of several centers estimated that the annual budget to fund space, staff, equipment, programs, and events was in the range of $1–2 million.

Participants

Faculty

Thirteen of 20 universities allocated faculty lines to the centers (Exhibit 4). Because these entities are not empowered to grant tenure, the faculty are appointed jointly with departments, and their duties divided. Many centers also include numerous affiliated faculty without appointments.

Some universities developed special hiring/promotion policies for joint faculty. For example, at MIT both the department and the IDSS center participate in tenure review, but the department has a stronger vote because it can fully "absorb" the faculty who wish to leave the center, while the reverse is not the case. Similarly, CMSE at MSU, while technically a department, tried to replicate the model used by national labs by recruiting faculty with the skills to develop tools for solving difficult scientific problems. All 27 faculty in the department have joint appointments (with 12 different departments), which are either a 70/30 or 30/70 split, to ensure that each faculty member has a primary "home."

UMass' CDS decided against joint appointments which it views as potentially disadvantageous, especially for junior faculty, and instead uses cluster hiring to promote collaboration. Finally, some centers created non-tenure tracked faculty positions, which do not require departmental affiliation. For example, DSI at UVA hires "general" faculty, who teach students and conduct research, but are not eligible for tenure. (The center also has joint tenure-track faculty).

Data scientists

Of the 20 centers, 12 created data scientist/engineer positions (Exhibit 4). In some cases, these staff devote most of their time to consulting services (e.g., at UNCC), but more typically they participate in collaborative research projects (e.g., at MIT, JHU, JPL, and UW). Several respondents mentioned that while many labs struggle to obtain computational support they need, data scientist positions are difficult to create at a university. A respondent from BU was able to persuade his administration to pilot several software engineer positions, with the understanding that these staff would quickly transition to grant support. The pilot was an immediate success: within a year, nine software engineers were hired and the number continues to grow.

As an evaluator of the MSDSE program, we are particularly familiar with the data scientist track created by eScience at UW (one of the grantees in the program). These staff are the engine of the center, leading most activities and programs, while simultaneously maintaining their own research agenda. eScience put in place several mechanisms to recruit and retain data scientists, which include relatively high salaries and a PI status. Data scientists can also recover a portion of the funding they bring in as a stipend.

Postdocs

Roughly half of the centers launched postdoctoral fellowship programs (Exhibit 4), which were described by some leaders as very competitive. At Harvard, postdocs are based in the department of their primary mentor, but have access to the center's space and participate in monthly lunches with the co-directors. DSI at NW established the Data Science Scholars program to diversify the domain-focused research portfolio of recent PhD graduates and to build their reputation as leaders in data science. The scholars have joint appointments with the Northwestern Institute on Complex Systems and at least one other research center on campus that matches their expertise. IDSS postdocs at MIT "belong" to the center and are expected to support its mission by working collaboratively with faculty across multiple schools. At NYU's CDS the fellows program is viewed as one of its greatest successes. Selected through a

competitive national search, the fellows are independent scholars who operate at the level of assistant professors. Not encumbered by teaching or administrative service, the fellows flourish in the collaborative environment of the center and have been highly successful on the job market.

Research activities

The data science centers in our sample directly support three types of research programs: small seed grants, larger team projects, and student research experiences (Exhibit 6).

Small seed grants and larger team projects

The majority of the centers offer internal grants to nucleate projects which may lead to follow-up funding. Some of these funding programs pair domain scientists with methodologists, others require that faculty represent intellectually distinct disciplines or have not worked together in the past, yet others target junior scholars. These programs are typically open to all faculty and disburse funding through simple, but competitive application process. Some calls for proposals have industry co-sponsors, who review the submissions and follow up with the applicants whose ideas are of interest to them. In most cases, the grants are in the range of $25,000–100,000, and can support a graduate student or postdoc for a short period of time (Exhibit 6).

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

Student research experiences

Half of the data science centers support student research experiences. For example, *Data+* at Duke's iiD is a 10-week summer program that offers undergraduates the opportunity to explore data-intensive problems from nonprofit and corporate clients. The students form several small teams which work in a communal environment. In 2017 the program sponsored 25 projects involving 75 students, who were chosen from 300 applicants. MIDAS at U Michigan supported four data science student groups with a combined membership of more than 400 students and 50 faculty members. These groups have completed 14 public service projects across Southeast Michigan. eScience at UW runs two incubator programs annually, which bring together data scientists and domain scientists. The summer session (called Data Science for Social Good) supports projects with a potential for societal impact, while the winter session (called the Incubator) focuses on high-risk/high-reward projects in any field.[3] Two of the centers launched programs for high school students. OSU's TDAI runs a free data science summer camp for girls attending Columbus high schools, where participants gain experience using software tools and presenting their work. A similar program is offered by MIDAS to economically disadvantaged high school students in southeast Michigan. Finally, many centers incorporated data science projects as components in courses or degree programs.

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

Exhibit 6: Internal funding programs

N	Center	Small seed grants and larger team projects	Student research experiences
1	BU DSI	✓	✓
2	Caltech CD³ and JPL CDST	✓	✓
3	Columbia DSI	✓	✓
4	Duke iiD		✓
5	Harvard DSI	✓	
6	JHU IDIES	✓	
7	MIT IDSS	✓	
8	MSU CMSE		
9	NYU CDS*	✓	✓
10	NW DSI	✓	
11	OSU TDAI	✓	✓
12	Stanford DSI		
13	UC Berkeley BIDS*	✓	✓
14	UChicago CDAC		
15	UMass CDS		
16	UMichigan MIDAS	✓	✓
17	UNCC DSI	✓	
18	URochester GIDS	✓	
19	UVA DSI		
20	UW eScience*		✓

**Moore-Sloan Data Science Environments.*

Note: The table only includes the programs funded by the centers.

Community engagement

Almost all centers offer seminars, workshops, consulting services, and annual meetings (Exhibit 7). While we did not plan to explore these community-building activities in interviews due to limited time, some respondents described them as highlights of their centers. For example, BU hosts annual Data Science Day to connect domain scientists and methodologists. Each year, the event is organized around several themes, which most recently included such diverse topics as artificial intelligence, cybersecurity and law, and epigenetics. U Mass runs an annual career event, which brings together students and industry partners for presentations and discussions, leading to numerous internship opportunities and job offers. The three centers funded by the Moore and Sloan Foundations take turns hosting annual data summits which combine scientific presentations with discussions of the issues important to this community, such as reproducibility, open science, careers tracks, and ethics.

Exhibit 7: Community-building programs offered by the centers

N	Center	Annual meeting, summit, retreat	Workshops, boot camps	Data science consulting
1	BU DSI	✓	✓	✓
2	Caltech CD^3 and JPL CDST		✓	✓
3	Columbia DSI	✓	✓	✓
4	Duke iiD		✓	✓
5	Harvard DSI			
6	JHU IDIES	✓	✓	✓
7	MIT IDSS	✓	✓	
8	MSU CMSE		✓	
9	NYU CDS*	✓	✓	✓
10	NW DSI		✓	✓
11	OSU TDAI		✓	
12	Stanford DSI	✓	✓	
13	UC Berkeley BIDS*	✓	✓	✓
14	UChicago CDAC			
15	UMass CDS	✓	✓	
16	UMichigan MIDAS	✓	✓	✓
17	UNCC DSI			
18	URochester GIDS	✓	✓	✓
19	UVA DSI	✓	✓	
20	UW eScience*	✓	✓	✓

**Moore-Sloan Data Science Environments.*

Industry partnerships

Eight of the 20 centers (at Columbia, MIT, NYU, Stanford, UChicago, UMass, UMichigan, and URochester) launched industry partnership programs, and several others are funded by industry on an ad hoc basis. Some centers cited substantial contributions: for example, CDS at UMass received $15 million from MassMutual and significant unspecified amounts from IBM, Pratt & Whitney, Google, Oracle, Microsoft, Amazon, and the Chan Zuckerberg Initiative. Stanford's DSI raised approximately $4 million per year through its industry program. This center chose to support its activities almost entirely through corporate contributions because they offer flexibility and larger budgets compared to the government funders, and because the practical nature of the problems of interest to industry resonates with the Stanford community.

Academic programs

As all universities in the sample offer courses and programs related to data science, we asked interviewees to focus only on the examples which are managed by their centers. Our review revealed that master's programs were especially popular, launched by half of the centers (Exhibit 8). Typically, these programs combine courses in quantitative methods and domain sciences. For example, master's students

at UNCC can mix and match courses to earn degrees in crime analytics, anthropology analytics, or health analytics, and a similar approach is planned for the undergraduate and PhD tracks. Another interesting model is the online micro-master's program launched at MIT: the graduates receive a certificate which could be applied toward a master's degree at another university. Four of the entities started a PhD program.

Some leaders mentioned online programs. For example, Caltech and JPL run a joint summer school which uses the data collected in space missions. Initially advertised to fewer than 100 people, the program attracted 30,000 registrants last summer.

Exhibit 8: Educational programs managed by the centers

N	Center	Certificate program	Undergraduate major or concentration	MS program	PhD program
1	BU DSI				
2	Caltech CD^3 and JPL CDST				
3	Columbia DSI	✓		✓	
4	Duke iiD	✓	✓	✓	
5	Harvard DSI				
6	JHU IDIES				
7	MIT IDSS		✓	✓	✓
8	MSU CMSE	✓			✓
9	NYU CDS*			✓	✓
10	NW DSI		✓	✓	
11	OSU TDAI				
12	Stanford DSI				
13	UC Berkeley BIDS*				
14	UChicago CDAC		✓	✓	
15	UMass CDS	✓		✓	
16	UMichigan MIDAS	✓			
17	UNCC DSI	✓		✓	
18	URochester GIDS		✓	✓	
19	UVA DSI			✓	✓
20	UW eScience*				

**Moore-Sloan Data Science Environments.*

Note: The table only includes the programs managed by the centers.

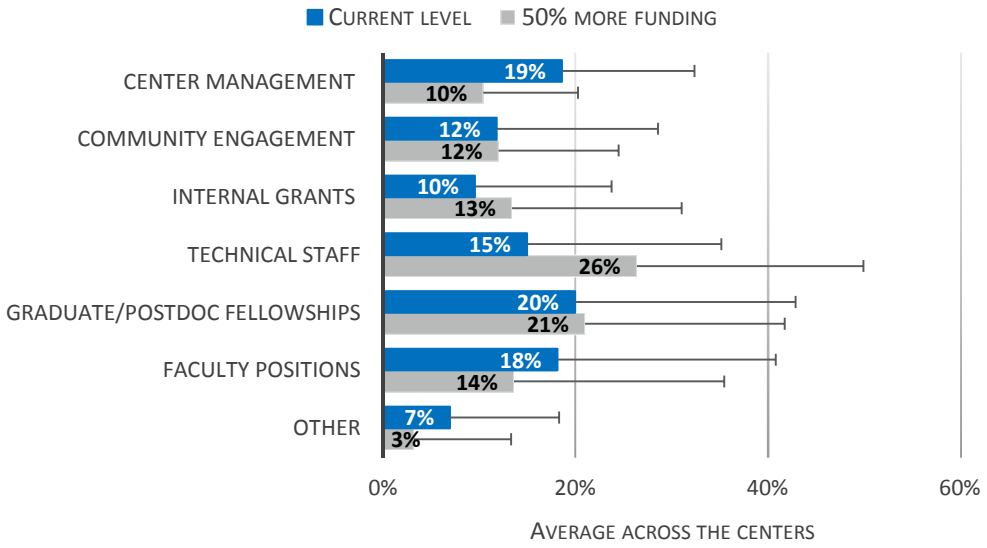
Funding allocation

In the survey of data science leaders, we asked how the total center funding is currently allocated across the various activities described above and how they would invest additional resources. Exhibit 9 shows that the centers currently spend 20% of their budget on average on management, 15% on technical staff (such as data scientists), 20% on graduate students and postdocs, and 18% on faculty. The remaining 20% is roughly divided between community engagement and internal funding. If 50% more funding were

available, the centers would distribute the largest share, 26% on average, to technical staff, almost double the current level for this expenditure category, followed by 21% to fellowships. The remaining half would be divided between management, community engagement, internal grants, and faculty in the allotments of 10-15%.

Exhibit 9: Funding allocation by the centers (N=27 for current and N=25 for 50% more)

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



Conclusions

A survey of 20 data science entities in the United States revealed that they used a combination of similar building blocks to create unique entities that best fit their university contexts. Several foundational elements are common to all or most centers – such as dedicated space and a strong emphasis on collaboration, interdisciplinarity, and community building. At the same time, clear differences also emerged. For example, some centers are primarily focused on research, while others combine research with training and/or formal education. In some cases data scientists/engineers are "the center of gravity," while in others postdocs or faculty play more prominent roles. The nature and extent of industry ties also varies. While each data science center will chart its own course, we hope that our study offers some insights into the range of possibilities for supporting this emerging field.

CENTER PROFILES

Data Science Initiative at Boston University

http://www.bu.edu/datascience/

Mission

"Boston University's Data Science Initiative (DSI) seeks both to leverage BU's existing strengths and further expand its capacity to compete and lead in the Big Data revolution. At the core of DSI is an effort to recruit some of the world's finest interdisciplinary faculty with proven track records in data science and strong potential for long-term impact at BU and beyond."

History and organization

The DSI, established in 2014, is housed within the Rafik B. Hariri Institute for Computing and Computational Science & Engineering (Hariri Institute). At the time, Boston University (BU) already had a strong presence in data science and the objective of DSI was to connect existing faculty and programs, and to identify and fill any gaps with new hires. In the first year of DSI, the Provost convened a search committee that was charged with recruiting new faculty. However, this model did not work as well as expected, and was replaced with the approach where hiring is initiated by departments that nominate candidates to the Provost, who in turn evaluates them with support from faculty advisors drawn from the DSI faculty. If approved by the Provost, the candidate is invited for multiple visits. The benefit of this process to departments is two-fold: they get "free" faculty selected from the candidates of their choice and the candidates are vetted by the DSI faculty who have expertise in a broad range of topics. This approach has been in place for two years and is viewed as effective – it has enabled the recruitment of three mid-career/senior faculty to BU.

In addition to faculty lines, the Provost also supports data science faculty fellows, who are nominated from within BU. They are selected through a systematic and thorough process, which includes obtaining outside letters of recommendation. The fellows are given additional resources and have access to the DSI platform.

The Hariri Institute also created software engineer positions that are viewed as "hugely successful." The idea for this position emerged from Hariri Institute staff observations that software development was highly sought after on campus, and that the laboratories with this need hired contractors or undergraduates, often with mixed results. Consequently, the Hariri Institute's Director persuaded BU to establish several lines for software engineers, who were hired into an entity called SAIL (Software & Application Innovation Lab). These staff (who tend to be PhD-level scientists) are offered competitive salaries by industry standards, and also enjoy intellectual freedom, collaborative opportunities, and access to students who they can mentor. The expectation for these positions was that while requiring a large initial investment from the university, it was temporary, as these researchers would quickly become valuable to the community and would be paid for through grants. Importantly, the tools they develop will be maintained because the university had invested in them, which will be a long-term asset to the community. The idea was an immediate success – within a year nine engineers were hired and the number is growing. Because of this positive experience and the growing need for software development, BU is considering incorporating data scientists as part of the research support infrastructure at BU (there are some challenges in establishing such positions, however).

The Hariri Institute has some dedicated space, but it is too small and all activities have to be scaled down to it. We were told that in April 2018, the University Board of Trustees gave the green light for a major/iconic 17-story 35,000 square feet "Computing & Data Sciences" building at the heart of the campus.[4] The new building will have the top seven floors dedicated to the Hariri Institute and DSI will claim most of the expansion space. However, it will be another three years until that building is completed, and until then the university has developed a plan to carve out some temporary space for DSI in the current building that will allow it to grow.

Academic programs

While BU offers many programs and courses that are related to data science, these offerings are not managed by the Hariri Institute/DSI, which do not get any revenue from them (even though they play a role in coordinating them).

Research and training programs

DSI hosts seminars, colloquia (weekly), and distinguished lecture series (four–five yearly). Its signature event is called the BU Data Science Day, which is run as a symposium, but has the primary goal of connecting methodologists and domain scientists (most of the participants are faculty and postdocs). Each year, the event is organized around several themes, which most recently included artificial intelligence, cybersecurity and law, epigenetics, and unexpected outcomes of big data. The inaugural BU Data Science Day held in 2015 attracted 60–70 people, the number grew to 300 in 2016, and continues to grow each year. According to our respondent, no other event at BU brings so many faculty together.

DSI also offers internal seed funding (up to $50,000 per project with a total annual budget of $250,000–300,000), which is allocated twice yearly through a simple application process. The funding can support a student or postdoc, and priority is given to faculty who have not worked together before. It is hoped that the seed funding will lead to follow-up grant applications. Some of the calls for seed grants have industry co-sponsors. The companies have the opportunity to review the proposals and, if interested, can follow-up with the applicants.

The Provost also supports four postdoctoral fellows and graduate fellowships each year. The latter is a nomination-driven program open to current and incoming PhD students, who receive supplements to their stipends that can be used to cover travel, equipment, or summer salaries.

Industry partnerships

A little over a year ago, BU announced a five-year partnership with the open source software provider Red Hat. This funding will support research laboratories and fellowships for PhD students, postdocs, and visiting scientists. Also recently, the Hariri Institute launched a new collaboration with the Honda Research Institute. The partnership will support research projects in a variety of areas, including data privacy and consumer personalization, and funding is disbursed through requests for proposals.

The Hariri Institute also runs a program *BU Spark!* supported by a gift from an alumna. The students in the program have the opportunity to work on company-initiated projects as part of a course. This program is very successful with both students and faculty, with more than 100 projects having gone through the pipeline (in some cases, companies continue to fund students after the course ends). The students gain

[4] http://www.bu.edu/today/2018/data-sciences-center/

practical experience and it also benefits faculty by making courses more relevant, interesting, and easier to teach. The projects originate with organizations such as the American Civil Liberties Union.

Lessons learned and future plans

We were told that it is important to not empower researchers in one topical area to lead data science efforts at the university, which could bias the effort. BU intentionally did not go down that path: the DSI Director does not see himself as a hardcore data scientist and many decisions are made by a group of faculty members. DSI is growing and on a path to becoming a standalone institute (perhaps even subsuming the Hariri Institute itself).

Center for Data Driven Discovery at the California Institute of Technology and the Center for Data Science and Technology at the NASA Jet Propulsion Laboratory

http://cd3.caltech.edu/
https://nsta.jpl.nasa.gov/dep-data-science

Mission

"The Center for Data-Driven Discovery …, in strong partnership with JPL, helps the faculty across the entire Institute in developing novel projects in the arena of data-intensive, computationally enabled science and technology."

"The Center of Data Science and Technology at NASA's Jet Propulsion Laboratory coordinates the research, development and operations of data intensive and data-driven science systems, methodologies and technologies across JPL Engineering, Science and Programs establishing a virtual center."

History and organization

The Center for Data Driven Discovery (CD^3) and the Center for Data Science and Technology (CDST) at NASA's Jet Propulsion Lab (JPL) became the Joint Initiative on Data Science and Technology in 2015 (each had been operational since the previous year). The two centers complement each other: CD^3 is more heavily focused on basic research and CDST on engineering and technology development. Furthermore, CDST can provide California Institute of Technology (Caltech) researchers (and other partners) with valuable data collected during NASA missions (called "use cases"), resulting in further synergy. The two centers are located approximately six miles apart and share resources and personnel. CD^3 is led by S. George Djorgovski (Professor of Astronomy at Caltech) and CDST by Daniel Crichton (Program Manager and Computer Scientist). Both centers serve as data science hubs and collaborative spaces for their multidisciplinary communities. CD^3 had been nucleated with funds from the university, and is now supported through external grants; and CDST is supported through external sources and funding from NASA. The majority of the work at CDST (85%) is NASA-based. CD3 is now also a part of the Caltech's Information Science and Technology initiative.[5]

CD^3 is housed in a suite of offices in a now defunct computational research center; some of the participants have offices in this space, while others are based in their home departments on campus. In addition to the Director, the center staff includes five staff scientists and a postdoc (it does not currently have faculty lines, but there is interest in creating them). Since Caltech has only about 300 faculty, CD^3 will remain small.

CDST is a vehicle to bring programs, people, and resources at JPL together (JPL has well over 100 researchers who could be called data scientists). The center is a virtual facility, which fits a long-standing model for JPL. It was described to us as a "third dimension" to pull people together who are physically located in different groups within the laboratory. JPL is engaged in a discussion of whether it is becoming necessary to physically co-locate all the data scientists, but in the meantime, CDST runs a working group that manages data science for JPL.

[5] http://www.ist.caltech.edu/

Research and training programs

CDST supports an internal pilot research program to explore how to apply data science to different areas at JPL. Approximately 12 pilots are selected per year. Proposals are reviewed by a selection committee that considers to what extent the projects will articulate the value of data science and promote a data science culture at JPL. The projects are funded for just under a year. The goal of the program is to test some capability and provide an "off ramp" to a more stable funding source. We were told that the program has been very valuable for JPL, as it has enabled researchers to take the time or the risk to test new ideas. It has also raised awareness of data science and created relationships and momentum.

CD^3 and CDST run a joint online summer school in data science, which takes advantage of JPL mission data and is extremely popular. The program was initially advertised to less than 100 people through social media and by the last count, 30,000 had enrolled. It covered different topics pertinent to data science (such as machine learning, statistics, and visualization) and included assignments and laboratories. Professor Djorgovski brought the course to the centers and has been teaching it on campus as well.

CDST also participates in a joint grant from NASA with UC Riverside to build a data science program. Finally, JPL sponsors internships and trains students in data science, with an average of 20 undergraduate and 10 graduate students.

Industry and other partnerships

CDST partners with industry to enhance technology developed at JPL and reapply it to robotic space exploration. The center also maintains academic partnerships with approximately 10 universities in addition to Caltech (including Harvard University, Stanford University, Dartmouth College, and the University of Texas), as well as with national laboratories. The huge advantage of JPL to the scientific community is the availability of real data, which drives many partnerships. CD^3 scientists are also participating in a number of multi-institutional collaborations.

Lessons learned and future plans

CD^3: if there is funding from prestigious places like the Gordon and Betty Moore and the Alfred P. Sloan foundations, it makes the university administration more willing to invest additional resources. Therefore, it is critical that these types of organizations continue to support data science. The plan is to run CD^3 as a demonstration for a few years, but ultimately to start a larger center using donor funds.

CDST had been a tremendous success at JPL: five years ago data science was not on the radar of the leadership but it is now one of the largest thrust areas. Pilot projects were described as especially valuable for building the data science community at JPL and for connecting with external partners. Collaboration with CD^3 will grow beyond the current level of approximately 10 joint projects. The two centers are also exploring ways to provide additional infrastructure services in machine learning, search analytics, visualization, and other data science areas.

Data Science Institute at Columbia University

https://datascience.columbia.edu/

Mission

"The Data Science Institute at Columbia has a three-part mission that encapsulates the great promise this new field has to improve the quality of life for all. Our mission is:

- To advance the state-of-the-art in data science;
- To transform all fields, professions, and sectors through the application of data science;
- To ensure the responsible use of data to benefit society."

History and organization

In 2011, Columbia University (Columbia) responded to a call from New York Mayor Michael Bloomberg to create an applied science and engineering campus in New York City (NYC). Columbia responded with a proposal to start an institute in data science, which would include its own faculty and educational activities, and would engage with industry. As the proposal winner, Columbia received access to NYC-owned land and up to $100 million in city capital, which had to be spent on new infrastructure. The university provided additional funds to hire faculty and get the institute off the ground.

The Data Science Institute (DSI), which was established in 2012, focuses on "training the next generation of data scientists and developing innovative technology to serve society." In the first five years, it was led by Director Professor Kathleen McKeown and Associate Director Professor Patricia Culligan. In May 2017, the university announced the appointment of Professor Jeannette Wing, previously a Corporate Vice President of Microsoft Research, as the new Director.

With Director Wing's arrival, DSI became a university-level institute, and is no longer housed within the School of Engineering. This change was characterized as a major organizational shift, and is meant to convey that the institute serves the entire university. DSI can hire research-track faculty, a "soft money" position. Thirty faculty, who have tenure-track appointments with a home department, were hired or designated as data science faculty to meet Mayor Bloomberg's grant obligations. Approximately 300 faculty from 12 schools across Columbia are affiliated with DSI.

DSI receives support through several sources: revenues from the master's program, indirect cost returns on grants, donor funding for the directorship, and a start-up "acceleration fund" from Columbia to the Director. Columbia currently has three research scientists, who advance data science and help faculty and students across campus apply data science. It has openings for additional research scientists.

With the funds received from the City of New York, Columbia renovated 40,000 square feet of space for DSI. The space includes faculty offices, small working rooms, larger conference spaces for meetings and lectures, and open areas. Feedback from the community about the space revealed that it facilitated interdisciplinary interactions and was valued and enjoyed by faculty and students.

Degree programs

DSI offers a competitive master's program in data science (with an acceptance rate of 10% and placement of graduates at 100%), which in January 2017 enrolled 150 students. For the entering class of fall 2018, applications increased by 42% from the previous year. DSI also offers a certification program intended for working individuals interested in improving their data science skills, with approximately 35 students;

and online courses, with 25,000 students. In addition to its master's level programs, data science is supported at Columbia by an undergraduate major in data science, jointly run by the Computer Science and Statistics departments.

In collaboration with Columbia's center on entrepreneurship, DSI also supports a curriculum development program called Collaboratory@Columbia, which is focused on developing computational and data management skills. Over 25 new courses were co-developed and co-taught at Columbia, representing interdisciplinary collaborations on campus in education.

Research and training programs

DSI launched a postdoctoral program and recruited the first cohort in the fall of 2018 (the program received over 100 applicants for 5 positions). DSI also runs a seed grant program that has two aims: to promote interdisciplinary collaboration on campus, typically through pairing domain scientists with methodologists; and to bring larger external follow-up funding. The seed grants are awarded to risky projects that may be in too early of a stage to get funded from traditional government funding sources [e.g., the National Science Foundation (NSF)]. We were told that DSI received a phenomenal response to the call for proposals and it was difficult to select a handful of awards. The level of support is approximately $100,000 per year for up to two years.

DSI also sponsors bootcamps, hackathons, and local outreach. Through its acceleration funds, it started a new DSI Scholars program to support research by undergraduates and master's students. Finally, the Director hosted the first Data Science Leadership summit (funded by NSF, the Alfred P. Sloan Foundation, and the Gordon and Betty Moore Foundation), with 65 participants. Leaders of data science centers/institutes/initiatives in 30 different U.S. universities exchanged experiences and lessons learned.

Industry and other partnerships

DSI has an industrial affiliates program, which attracted big players including Bloomberg, Microsoft, Google, Adobe, as well as the finance, media, and pharmaceutical industries, and international companies, such as Alibaba and Baidu. At any point in time, there are about 25 companies who are DSI industry affiliates. Participants pay a membership fee and get access to students and faculty for joint research projects and recruitment. DSI hosts various events and career fairs to help industry make connections with the Columbia community. The industry program is very popular.

DSI has spawned a number of startups through the work of faculty and students, many of which are based in NYC. It also has a strong relationship with the NYC Economic Development Corporation (an organization that supports the startup community) and with the local government. Finally, DSI holds an annual Data Science Day on campus, which had 700 registrants in 2018. The day-long event showcases data science research projects and capstone projects by DSI students, and has an industry keynote speaker. In 2018, Professor Wing and President Bollinger had a fireside chat on data and democracy. Attendees come from academia, industry, and government.

Lessons learned and future plans

In the first few years, DSI had to resolve several structural challenges. One of the greatest was faculty hiring. DSI needed to work with departments whose faculty were hired after a lengthy search strategy, search committee, and decision-making process, which took several years. Another challenge was related to the distribution of NYC support for faculty slots across the schools. DSI leadership learned that extensive and continuous communication with all stakeholders was very important, and that it was

necessary to find the right communication strategy for each school. Finally, an internal review revealed that one of the needs for the faculty was adequate time for course preparation and teaching. In response, DSI affiliated faculty were provided with "release time."

Looking to the future, DSI promotes the mission of "Data for Good," succinctly capturing its three-part mission statement: (1) advance the state-of-the-art in data science; (2) transform all fields, professions, and sectors through the application of data science; and (3) ensure the responsible use of data to benefit society. In its transition to being university-level, the goal is to promote multidisciplinary collaborations in research and education across the university.

Information Initiative at Duke University

https://bigdata.duke.edu/

Mission

"The Information Initiative at Duke (iiD) is an interdisciplinary program designed to increase 'big data' computational research and expand opportunities for student engagement in this rapidly growing field. Launched as an initiative of Duke University, iiD is partnered with the Duke University Quantitative Initiative, which promote cross-pollination of ideas throughout Duke's programs and research projects, and works to increase the number of quantitative faculty in all disciplines on Duke campus."

History and organization

iiD was launched in 2013 and is led by Professor Robert Calderbank (Computer Science, Electrical Engineering, and Mathematics). The initiative was designed with support from the Vice Provost and with input from more than 30 faculty from the schools of Arts and Sciences, Public Policy, Business, and Medicine; and the Library. The leadership and management team, in addition to Professor Calderbank, includes an Associate Director, an Administrative Coordinator, and a Program Coordinator.

The initiative is partially funded with a $10 million gift to the Pratt School of Engineering, which convinced the administration that it is possible to raise money for data science. iiD subsequently received matching support from the university. Its total operating budget is approximately $1 million per year. The "center of gravity" of iiD is its undergraduate students.

iiD is located on the third floor of Gross Hall in 20,000 square feet of space that features laboratories, community rooms, classrooms, and offices. An atrium is shared with the Social Sciences Research Institute (SSRI) and is meant to facilitate cross-disciplinary collaboration. About 12 faculty members have their offices in the space, and 85 graduate students and postdocs regularly work there. The initiative does not have its own faculty lines; all participating faculty are hired into and promoted through departments. The decision about faculty affiliation was based on the intent of iiD to add value to participating departments and avoid competing with them for faculty lines.

Degree programs

In fall 2018, Duke admitted the first cohort of students in its Master's in Interdisciplinary Data Science (MIDS) program. Jointly hosted by iiD and SSRI, the program offers training in quantitative sciences, exposure to a variety of disciplines, and experience in interdisciplinary team-based data science inspired by Data+, a 10-week summer research program. Initial program plans were to recruit 20 students, but this number is expected to grow to include students from increasingly diverse academic backgrounds. Tuition earned through the program will go to the instructors and fund visiting faculty.

Research and training programs

The flagship program at iiD is Data+, a summer research experience that offers undergraduates an opportunity to explore data-intensive problems from actual nonprofit and corporate clients. Participating students form several small project teams (two–four students per team) to work on their own problems in a communal environment. Data+ helps students acquire both data science skills and client management skills, which will be invaluable for future employment and are difficult to acquire at a university. In 2017 the program sponsored 25 projects involving 75 students. The program aims to attract talented

sophomores and train them for internship at leading technology companies in the following year. Participants receive a stipend.

Data+ also served as a professional development experience for graduate students and postdocs, who serve as team mentors. This is particularly helpful for students in the humanities, who are able to gain both quantitative skills and social exposure, which are often missing in these fields. The program is very popular and typically receives about 300 applications for 70 slots. It is administered by a staff member on a half-time basis, but the iiD Director is also involved in fund-raising, identifying problems, and other steps. We were told that managing the program is labor-intensive with preparations taking most of the year. Before the program officially starts, student participants complete Institutional Review Board (IRB) training and the iiD staff meet with mentors to develop work plans (this is not standard practice at universities, but is routine in industry). iiD maintains close partnership with the Information Technology (IT) department to host a protected research network, so that students' work, which involves outside partnerships and data, can be monitored. One of the main responsibilities of the iiD Director is to raise funds for the program: of the $450,000 needed, approximately $100,000 comes from iiD, $150,000 from local companies, $100,000 from philanthropic resources, and $50,000 from university departments and *Bass Connections* (a university-wide large educational and training initiative). We were told that contributions from university departments are the "true magic" of Data+, as it makes everyone feel that they are taking ownership of the program.

Another program run by iiD is Data Expeditions, which incorporates exploratory data analysis as a component of undergraduate courses. Pairs of graduate students work with course instructors to formulate research problems for undergraduates, who use datasets to address them. For this service, graduate students (called Pathfinder Fellows) receive a travel grant of $1,500. Data Expeditions is a collaborative project: iiD provides the resources; SSRI maintains the datasets; and representatives from affiliated departments provide students, academic oversight, and direction.

iiD also offers several faculty-mentored research opportunities: *Bass Connections*, a combination of courses and project work that focuses on society and culture; Pratt Fellows, a competitive program for undergraduate students in engineering to perform research projects; and postdoctoral fellowships, one–three year positions funded through grants from the Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Research Office.

Finally, 10% of the iiD budget over four years was invested in the autism project, which is likely to lead to a new research partnership with Apple. The center participants are also working on developing mathematical methods to draw congressional districts, and a digital exhibition where visitors can visualize the altarpiece when it was first painted. These types of projects typically get matching funds from other sources.

Industry partnerships

iiD does not have a formal industry affiliates program, which the leadership of the center is in the process of organizing. Nevertheless, industry contributes $100,000 to $150,000 to iiD annually.

Lessons learned and future plans

The most important lesson conveyed in the interview was the attractiveness of problem-driven projects and team activities to students.

Data Science Initiative at Harvard University

https://datascience.harvard.edu/

Mission

"The Harvard Data Science Initiative will unite efforts across the university, foster collaboration in both research and teaching, and catalyze research that will benefit our society and economy. It will be home to a research platform to accelerate the pace of discovery. It will strengthen the fabric of connections among departments to create an integrated data science community, all to empower research progress and education across the University."

History and organization

In 2014 and early 2015, Harvard University leadership came to recognize that there was a significant opportunity to both support existing activities in data science at Harvard and to open new lines of inquiry, create collaborations, and ultimately have a much larger impact by establishing the data science initiative. Consequently, the Provost asked the Vice Provost for Research to convene a small group of faculty to develop a proposal. The Harvard Data Science Initiative (HDSI) was launched in February 2017 and Professors Francesca Dominici [Biostatistics, Harvard T.H. Can School of Public Health, (HSPH)] and David Parkes [Computer Science, John A. Paulson School of Engineering and Applied Sciences (SEAS)] were appointed Co-Directors. The co-directorship model was chosen to ensure representation from two geographically separated Harvard campuses, and to underscore a commitment to both methodologies and applications. The leadership team also includes an Executive Director, Elizabeth Langdon-Gray. HDSI operates through a Steering Committee composed of eight members (representing a range of fields), which provides advice to the Co-Directors. It also uses a larger planning committee composed of faculty from many schools at the university.

Harvard has a number of faculty initiatives that span different schools and departments. HDSI is different from other initiatives in that it has an administrative home in the Provost's office. This decision was made to highlight the commitment of the university leadership to bring data science to different disciplines across the administrative structures. Furthermore, data science is an emerging discipline that does not easily fit within any one department or school.

The initial funding provided by the Provost was used to seed postdoctoral fellowship and research grant programs. The initiative does not currently have faculty lines; these are a long-term goal and the model being considered is joint appointments between HDSI and university departments.

One of the missions of the data science initiative is to create a career track for professional data scientists. These positions will likely be supported through a mix of sources including grants, university funding, and philanthropic resources. These career tracks are a high priority for HDSI leadership. HDSI is also considering a possibility of hosting data scientists from industry on a part-time basis to help build the community around data science.

At present, HDSI has a small space in Harvard Square (approximately 1,500 square feet), but is hoping to eventually move to larger quarters, likely in Allston.

Degree programs

While HDSI does not serve as a home for degree programs, the planning process for the initiative may have had an effect of spawning several new data science programs at the university. These include a

master's in Biomedical Informatics (launched in fall 2017 at the Medical School), a master's in Health Data Science (launched in fall 2017 at the School of Public Health), and a master's in Data Science (launched in fall 2018 in the Faculty of Arts and Sciences/School of Engineering and Applied Sciences). All master's degree programs are 18 months in duration. Harvard has also launched an online certificate program in business analytics, designed for executives (jointly offered by the Business School and the Faculty of Arts and Sciences/School of Engineering and Applied Sciences). The university also has a growing undergraduate curriculum in data science. HDSI is playing a role in coordinating these programs.

Research programs

A postdoctoral fellows program was established; HDSI received applications in January 2017 and brought in the first 8 fellows in September 2017 (4 more started in September 2018). This is keeping with the program's goal of approximately 12 fellows. The fellows were selected from a range of disciplines, including psychology, statistics, and biomedical informatics. Each fellow has to identify at least two faculty mentors from different disciplines, with the hope that they will form an intellectual glue between these researchers. Postdocs, who are appointed for two-year terms, are based in the department of their primary mentor, but also have access to the shared space. Postdocs participate in monthly lunches with the HDSI Co-Directors and also meet monthly with each other to discuss ideas and form collaborations. This program is funded by the Provost and by philanthropic resources.

HDSI has also launched an internal seed grant program, which funded five small grants in 2017 and seven grants in 2018 ($50,000 each in the most recent year). In 2018, HDSI solicited proposals for planning grants that will be foundational to larger research programs within the research themes. The program does not have an explicit goal to pair the methodologist with domain scientists, but rather to fund the most exciting research, with an emphasis on junior faculty.

HDSI also hosts a monthly "45/45" seminar series, comprised of two back-to-back 45-minute talks, one methodological and one application-based. The seminar is on a tour to existing seminar series around the university.

Industry partnerships

HDSI has mapped out the broad contours of a corporate membership program and is starting the process of talking to potential partners, with the goal of launching in late 2018. The program will offer faculty the opportunity to learn about challenges coming from industry and opportunities for collaboration, and will offer corporate partners a chance to learn more about research being conducted at Harvard and sponsor research, as well as student recruitment opportunities.

Lessons learned and future plans

To be successful, the activities of the initiative have to be led by faculty. HDSI would like to grow and become an institute.

Institute for Data Intensive Engineering and Science at Johns Hopkins University

http://idies.jhu.edu/

Mission

"The Institute for Data Intensive Engineering and Science (IDIES) will foster education and research in the development and application of data intensive technologies to problems of national interest. The institute will provide faculty, researchers and students with the structure and resources needed to accomplish these goals."

History and organization

IDIES was established in 2012 and is directed by Professor Alex Szalay (Department of Physics and Astronomy). It has an annual budget of $1 million, half of which is contributed by the Johns Hopkins University (JHU) President and the rest by deans from participating schools (which include Arts and Sciences, Engineering, Public Health, Medicine, and Business) and the Library. It took IDIES five years to build trust with the schools of Medicine and Public Health, but now they are fully engaged and will triple their contributions to the institute by the end of 2018.

The President's initial commitment was for five years which is coming to an end this year, but he is willing to continue support for another three years if the institute can propose a sustainability plan. IDIES also has additional funding for 6 faculty lines, which comes from a gift by Michael Bloomberg (who donated 50 Senior-Endowed Chairs to JHU) and 14 from the university, for a total investment of $25 million. At JHU. faculty lines are based at departments and the institute needed to find departmental partners interesting in hiring in the areas related to big data. At present, more than 100 faculty are associated with the IDIES.

IDIES has been very successful in raising federal funding (with a success rate of 47% compared to 27% for JHU and 15% nationally). During the first three years of its existence, the Deans and the President invested $3 million in total, but the overhead raised from federal funding was double that amount. The institute's Director is trying to negotiate with the university's administration a mechanism for indirect (overhead) cost recovery to enable the growth of IDIES.

IDIES has a staff of about 20, which includes 6 Programmers (many of whom are science PhDs), 6 System Administrators, Grant Administrators, a Public Relations Specialist, a Science Writer, and a Web Designer. These staff represent a shared resource; 50–70% of their salaries are supported through the institute's budget and the rest is covered through federal grants. IDIES also operates all of the high-performance computing equipment at the university.

IDIES has some meeting space on campus, but is looking for a larger joint space. Most permanent staff are located in the Physics Department, where some of the events occur.

Research and training programs

Every year IDIES awards four–six inter- or multi-disciplinary seed grants, which are funded at approximately $25,000 and are typically one year in duration. The proposals are one-page in length and preference is given to junior faculty. The funding can be used to cover a graduate student, and additional resources can be made available to the grantees free-of-charge in the form of Programmers and System/Database Administrators. The goal of these grants and other collaborative projects is to teach researchers from different fields how to use big data techniques (rather than do this work for them). Once

the participants become comfortable with these methods, IDIES staff remain available for consultations, but spend most of their efforts on new projects.

In addition, IDIES faculty collaborate on several large projects at the JHU, including in genomics and materials science. Finally, the institute is providing valuable community service by building relational databases from large data files and the interface to access these data. This enables users to be able to interact with but not have to download the data, which require powerful computers. This approach is becoming very successful and IDIES is expanding this work from tornados and cosmology data to oceans (more than 100 papers have been written based on the data).

Industry and other partnerships

IDIES is looking to develop relationships with industry and have connections to national laboratories and the City of Baltimore. IDIES offers datasets and other educational aids to K–12 students, teachers, and the general public, such as visualizations of millions of stars and measurements of soil properties.

Lessons learned and future plans

It takes considerable person-to-person communication to establish trust and it is important to have advocates for data science.

Institute for Data, Systems, and Society at the Massachusetts Institute of Technology

https://idss.mit.edu/

Mission

"The mission of IDSS is to advance education and research in state-of-the-art analytical methods in information and decision systems, statistics and data science, and the social sciences, and to apply these methods to address complex societal challenges in a diverse set of areas such as finance, energy systems, urbanization, social networks, and health."

History and organization

The Institute for Data, Systems, and Society (IDSS) was established for two related missions. The first was to create a home for statistics, a discipline that was scattered across MIT. The second was to form a bridge between engineering and social sciences in order to address complex societal problems that cut across many disciplines.

IDSS, which spans all five schools at MIT, was launched in July 2015 under the leadership of Professor Munther Dahleh (Department of Electrical Engineering and Computer Science). A committee of 40 faculty participated in planning the institute, and 38 of these faculty members agreed to the proposal. The Provost provided funding for IDSS for the first five years; after this time, it is expected to become self-sustaining. The institute has already been successful in securing additional funding through endowments and industrial contributions.

Faculty at IDSS have joint 50/50 appointments with appropriate departments. Tenure decisions are made collaboratively, although departments have a more decisive voice because they can fully "absorb" faculty who want to leave the institute, while IDSS cannot (no departures have occurred to date). In total, IDSS was allocated 16 faculty lines representing all schools at MIT; so far 7 faculty have been hired at junior and senior levels. Institute participants also include post-docs and research scientists who are funded through grants.

IDSS occupies 40,000 square feet of space, which has offices for faculty and students. Half of that space is occupied by the Laboratory for Information and Decision Systems, the oldest laboratory at MIT and a core unit within IDSS. The rest of the space is occupied by the newly formed center for Statistics and Data Science (SDSC), a center focused on integrating faculty efforts in research and education in the broader fields of statistics and data science. SDSC, under the leadership of Professor Devavrat Shah, has successfully launched a minor in statistics and data science, an online micro-master's in statistics and data science, and an interdisciplinary PhD program in statistics. The space provides collaborative room for faculty and students to interact with and collaborate in research and teaching. The space also includes a visualization laboratory that is intended to grow to serve both researchers and students in the various programs offered by SDSC.

Degree programs

IDSS offers a doctoral program in social and engineering systems (focused on addressing important societal problems) and hosts a master's program in technology policy. IDSS also offers a multidisciplinary PhD in statistics, a unique program that is done in collaboration with departments to

allow their students to qualify and conduct research in statistics. The first student will graduate from this program in 2018 in mathematics.

An online micro-master's program started in September 2018. Students who complete this program earn a certificate and will be able to take additional courses toward a full master's degree at any university. Finally, IDSS offers a professional online 7-week course in data science, which receives 1,000 registrations per session.

Research and training programs

A portion of the Provost's budget is used to support two-year postdoctoral positions. Unlike a typical postdoc, these researchers "belong" to IDSS rather than an individual faculty, and are expected to support the institute's mission by working collaboratively with faculty across multiple schools. These positions are advertised annually and the applicants are carefully chosen by IDSS. The program has been very successful.

The Provost's budget is also used to host visiting faculty. The goal of this program is to bring to MIT leading scholars in their fields who are not willing to relocate. The program has been especially successful in statistics.

Finally, IDSS funds seed grants, which are typically joint projects between data scientists/engineers and social scientists. These mini-grants can be used to support a student or postdoc, some faculty time, and incidentals such as travel. Such grants has promoted collaborations among political scientists, economists, and computer scientists.

Industry partnerships

The institute has an industry partnership program. A membership fee-based model common at MIT was initially considered, but IDSS was able to create a more ambitious program, where each company commits $300,000 per year for three years. Three participants are currently enrolled in the program (Booz Allen Hamilton, Thompson Reuters, and WorldQuant), and the fees represent a substantial income for IDSS. The funding supports a PhD program, which offers graduate students an opportunity to work on company-related projects that interest them. We were told that it is not always easy to blend open-ended PhD projects with commercial projects that tend to have specific objectives, but nevertheless the program has been very successful so far.

Lessons learned and future plans

We were told that it is difficult to establish a cross-campus institute in a university environment where departments function independently. Our respondent recalled challenges integrating faculty, finding common ground with departments, and explaining the goals and advantages of the institute to faculty and department leadership. In retrospect, hiring faculty could have been made more flexible. The highly structured process put in place when IDSS was launched is appropriate for 1–2 faculty, but too labor-intensive for 16.

In the near future, IDSS plans to further develop its academic programs, raise additional funding to cover all participants and activities, hire the remaining faculty, and launch the data laboratory.

Department of Computational Mathematics, Science and Engineering at Michigan State University

https://cmse.msu.edu/

Mission

"CMSE, jointly administered by the College of Natural Science and the College of Engineering, will enable application-driven computational modeling ('pull'), while also exposing disciplinary computational scientists to advanced tools and techniques ('push'), which will ignite new transformational connections in research and education."

History and organization

The Department of Computational Mathematics, Science and Engineering (CMSE) was officially established in July 2015 following several years of planning and negotiation. The planning committee carefully considered a center- vs department-based model, and ultimately chose the latter because the review of data science centers around the country revealed that center-based models typically disappear after five years, especially at institutions with limited resources. In contrast, a new department would provide a permanent anchor around which a community of faculty and students could be created.

The main limitations of a department-based model are the substantial start-up costs and an estimated $1 million yearly budget needed to maintain the organization, which includes salaries for the Chair and administrative staff. It was difficult to gain widespread support for this substantial investment when considering that the funds could instead be allocated to new faculty lines in existing departments. It is also important to consider "the cost of the social context": faculty may perceive a center as more open than a department, and thus feel more comfortable engaging in the center model. Finally, faculty at established departments may feel threatened when a new department is created. Consequently, the planners made it a priority to try and understand and engage as many people as they could at all levels so that they were aware of other interests and tendencies at the university.

At Michigan State University (MSU), faculty and administration at all levels participated in the planning and negotiation process. When it became clear that the university was interested in the proposal, a committee of 17 faculty was recruited from across the university. This committee spent four months looking at peer universities and debating the center vs department alternative. Once the decision was made to launch a department, the members met with almost 100 people on campus to understand who might be involved, threatened, disengaged, or engaged; and ensured that everyone understood what was occurring. At this point, the committee developed a proposal to justify the need for a department, which was posted openly once the proposal went forward to the university.[6] Despite this extensive stakeholder engagement process, once the proposal began to advance through the administrative process, additional high-level negotiations were necessary to bring everyone on board.

CMSE is located in the former library of the university's engineering complex. The space has been recently renovated to include communal areas in the center and 30 private offices around the perimeter. It

[6] https://acadgov.msu.edu/sites/default/files/content/Accessible/CSMEDeans%27%20cover%20letterand Proposal.pdf

can accommodate 120 graduate students and the entire administrative staff. The space is very attractive and equipped with state-of-the-art technology.

Degree programs

CMSE currently only admits students to the PhD program. The students are required to take four basic courses and can elect more advanced courses in data science or scientific computing. It also offers dual PhD degrees in these two areas to students from other departments. CMSE currently has 28 of its own PhDs and 30 dual PhD students. CMSE is committed to using modern pedagogy in its teaching with most of the course material being taught using a flipped classroom model and engaging students using Jupyter Notebooks.

The department is preparing to launch two terminal masters: in data science, jointly with computer science and statistics; and in scientific computing on its own. It also offers an undergraduate minor (focused on modeling) and two graduate certificate programs. The minor already incudes data modeling courses that include heavy data science components, and CMSE has a range of courses such as data visualization and communication. Finally, the department launched an early version of a freshman undergraduate introductory course in data science in fall 2018, which will ultimately grow into an undergraduate degree program. This undergraduate degree program is being developed jointly with CMSE's partner departments, Statistics and Computer Science, and CMSE is very enthusiastic about this collaboration.

Research and training programs

The CMSE department was formed around the vision of bringing researchers together to solve problems and blurring boundaries between disciplines. This model for CMSE is based on the types of collaborative environments seen at national laboratories rather than traditional academic departments. All 27 faculty have joint appointments, which are based at 12 other departments. Only seven faculty so far have been recruited from within MSU; the rest were external hires. Of the external hires, 1 senior faculty was hired without tenure from a national laboratory and is currently going through the promotion process, and the remaining 19 are assistant professors. The department is bringing together people with a broad range of expertise to develop tools for solving challenging scientific problems. So far, the department had made cluster hires in biology, materials, and other areas. Approximately half of the faculty are in data science and half in scientific computing. All CMSE faculty appointments are 70/30 or 30/70 split, and each faculty has a primary department ("one boss"). The university has a long tradition of joint appointments and no concerns were expressed about challenges related to tenure.

The department co-funds some computational/data scientists based at the MSU's High Performance Computing Center. In addition, it hired specialists in curriculum development and other areas to assist other departments with their needs.

Our respondent believed that generous start-up packages and various university-wide internal funding programs make seed grant or similar programs unnecessary for CMSE faculty. The department plans to build stronger connections to industry but is currently focused on establishing its educational programs. CMSE sponsors a semi-annual workshop in foundations of data science and computing. The event brings together representatives from industry, national laboratories, and academia for three days. Finally, the department hosts brown bag lunches and other community-building events centered on building the foundational relationships needed to be successful.

Lessons learned and future plans

Our respondent believed that the key requirement to successfully building a department is to listen to the university community. It is also important to be open to an outside audit, which the department engages in through annual meetings with the external advisory board of industry, and academic and government leaders. This oversight helped the department to quickly correct problems. Finally, the respondent believed that the centers are unstable (80% disappear by the five-year mark), and that only departments can ensure long-term persistence (successful centers will be eventually converted to departments).

The department will continue to grow, with the plan of including 50 joint faculty. After a few more years of junior hires, CMSE will recruit senior faculty to provide gravitas to the department and to start applying for center-level grants.

Center for Data Science at New York University

https://cds.nyu.edu/

Mission

"The Center was established to help advance NYU's goal of creating the country's leading data science training and research facilities, and arming researchers and professionals with tools to harness the power of big data."

History and organization

The Center for Data Science (CDS) was established in 2013. Initially an administrative home to the master's program in data science, CDS significantly expanded its activities and operations when it was awarded a $12.6 million grant to become one of three Moore Sloan Data Science Environments (MSDSE). The MSDSE is nested within CDS and the two entities share space and resources, but each has its own leadership team. In August 2016, the center moved into newly renovated space, composed of two floors, one devoted primarily to "quiet work" and the other to events and collaborative projects.

The university allotted 18.5 faculty half-lines to CDS; as the center cannot grant tenure, all faculty must have a joint appointment at a "home" department. Through extensive negotiation and participation of stakeholders from different units at New York University (NYU), CDS developed a systematic protocol for the hiring of joint faculty, which was viewed as a major accomplishment at the university where academic units function very independently. CDS has hired top experts in various areas of data science, including faculty members whose research crosses disciplines. In addition, over 50 faculty are informally affiliated with CDS.

In addition to faculty, CDS/MSDSE participants include data science fellows, postdocs, and research engineers. The fellows are selected through a highly competitive national search, and are professionally closer to assistant professors than postdocs. These researchers are appointed for one–three year terms and are fully salaried through CDS. Postdocs are hired both from within NYU and externally, and are co-funded with departments. Finally, Research Engineers, who are fully or partially supported by the MSDSE, participate in collaborative projects, develop tools, and help with other data science needs at the university. Fellows, post-docs, and research engineers have served as important bridges between CDS and research groups throughout NYU. They have also engaged in educational activities where the students have greatly benefited from their expertise and experience, which complements what is covered in theoretical courses and gives students a better perspective into real data science problems.

Degree programs

NYU launched its master's program in data science in 2013, and it has become highly successful. In 2017, the program received 1,800 applicants for 150 slots, a 10-fold increase since the first year. All master's students participate in capstone projects focused on real-world problems that can be approached using data science methods. In 2016, CDS announced a PhD program in data science, one of the first such efforts in the United States. The program has attracted substantial interest, having received 386 applications in 2017. Finally, CDC offers a non-degree program for professionals who wish to acquire or strengthen their data science skills.

Research and training programs

For two years, the MSDSE ran a seed grant program designed to bring together data scientists and domain scientists to work on collaborative projects. Each team received up to $25,000 to cover a graduate student or postdoc for one semester. The program was highly successful and the funded projects produced novel research that led to publications, open-source systems, and several grant proposals.

In 2018, the program was replaced with a Summer Research Initiative, which supports master's students working with NYU faculty for 250 hours. In the first round of competition, 35 faculty submitted proposals and 8 projects were funded.

In addition, CDS hosts a number of seminar series, including Text as Data (focused on natural language processing), Math and Data (topics on the intersection of applied mathematics, statistics, and machine learning), and Moore-Sloan Data Science Lunches (an informal gathering of affiliated faculty and researchers to discuss data science topics of interest).

In collaboration with the other two MSDSEs (at the University of California Berkeley and the University of Washington), NYU helped organize the very popular Astro Hack Week, Geo Hack Week, and Neuro Hack Week, which led to white papers on how to run these types of events.

Industry partnerships

CDS offers a three-tiered industry partnership program: "leaders circle" with gifts of $500,000 and more, "honors circle" with gifts from $100,000 to $500,000, and "fellows circle" with gifts up to $100,000. Regardless of the level of contribution, all partners are eligible to recruit students for internships, attend career fairs, and get involved in research projects. Companies in higher tiers are also invited to guest lectures, visits, and smaller community events. The partnership program links to a diversity initiative, through which members can support groups underrepresented in data science by funding scholarships, research projects, and travel. Current CDS partners include DeepMind, CapitalOne, Samsung Research, RiskEcon, and Lunit.

Other contributions

Like other MSDSEs, CDS participants contributed to the development of popular, open-source software tools, such as scikit-learn and Julia. In partnership with the NYU Library, CDS has also been actively involved in promoting reproducible science practices on campus by offering training sessions, consultations, and lectures on how to obtain and manage large datasets. Staff at CDS continue to develop open-source software to support computational reproducibility, including Reprozip and ReproServer. These tools allow researchers to store the data and software underlying computational experiments in virtual machines for later execution on multiple platforms.

Lessons learned and future plans

According to the leadership of CDS, the best way to support data science is to invest in people. The Fellow program, which offers complete intellectual freedom, is seen as a particularly successful model for supporting academic data scientists: most of the alumni obtained tenure-track jobs, while a few chose data science positions in industry. Another observation made by CDS was the importance of space. The center flourished once it moved from temporary space that was universally disliked into new, beautiful quarters. The new space is heavily occupied by students, postdocs, and faculty representing different disciplines, which has sparked unusual interdisciplinary connections.

Data Science Initiative at Northwestern University

https://datascience.northwestern.edu/

Mission

"Northwestern University's Data Science Initiative aims to bring together scholars and students from across the institution to enhance learning, development and application of data science methods and to promote the utilization of data science approaches in teaching, research and entrepreneurship."

History and organization

The beginnings of the Data Science Initiative (DSI) date back to 2013 with the Lawrence B. Dumas Domain Dinner sponsored by the Provost's office; the overall purpose of these dinners is to stimulate faculty interactions across departments and disciplines, and to highlight Northwestern University's distinctive interdisciplinarity. Professor Luis Amaral from Northwestern's McCormick School of Engineering and Applied Science used the event as an opportunity to discuss the availability and use of new data sources across fields. The dinner was very well-received and prompted the organization of a faculty workshop in 2014 to discuss how the university could be more proactive toward data science activities. The success of this follow-up event – 100 faculty from across the university attended – confirmed the interest in formalizing the DSI. To continue the momentum, Professor Amaral and a colleague organized a programming boot camp in 2015, which was attended by 400 students and faculty from 10 schools at the university, including Arts & Sciences, Engineering, Management, and Medicine. In summer 2015, the Office of Research committed $3.8 million for a period of three years to support the DSI. Since its inception, the DSI has also received support from several university donors. In particular, a major donor has contributed $500,000 per year for operating expenses.

The DSI is governed by a 13-member Steering Committee chaired by Professor Amaral. The mission of the Steering Committee, which includes representatives of most schools and several interdisciplinary centers at the university, is to gather information on the aims of units across Northwestern regarding data science, and to define and implement a strategy for how to spend DSI funds.

To enhance true collaboration and interdisciplinarity, the DSI leveraged space across units and fostered cross-utilization. For instance, Data Science Scholars were appointed to a minimum of two Research Centers and given work space in each center. Another example is the faculty research networking luncheons, which occur at the center for executive education within the Kellogg School of Management because this facility is conducive for gatherings of this type.

It was decided early on that the DSI would not have direct control over faculty lines. Faculty lines at Northwestern are under departmental controls, and the Provost and Vice President for Research did not want to "impose hiring from above." Instead, the DSI has coordinated with the $150 million CS+X initiative, which established 20 new faculty lines, half of which are joint appointments between Computer Science and other departments. The DSI has also been working with departments to help them grow organically in the data science areas most relevant to their interests by advising faculty search committees.

Degree programs

The Department of Statistics is creating a minor in data science that will formalize a course sequence that has been under development for the past two years. The minor (open to all Northwestern undergraduate

students) will require the completion of a three-course sequence in data science methods, a course on data visualization, plus an introductory statistics course and one elective.

Northwestern created one of the first Master of Science in analytics in the country, which runs out of the School of Engineering. This school has also created a Master of Science in artificial intelligence, and started enrolling students in fall 2018.

Northwestern was awarded one of the first National Science Foundation Research Traineeship (NRT) aimed at data science. The Integrated Data-Driven Discovery in Earth and Astrophysical Sciences (IDEAS) program is a multifaceted program designed for master's and PhD students in a variety of departments. At its core, the program offers a range of coursework in data science, but it goes far beyond that. As part of the program, students can also participate in summer school activities focused on data visualizations and computer programming, engage in science communication workshops, participate in the development of a citizen science project related to their research focus, and earn internship credits.

Research and training programs

The DSI supports four programs and activities: the Data Science Fellows program, the Data Science Scholars program, the faculty research networking luncheons, and an internal funding program.

The goal of the Data Science Fellows program is to assist departments across the university in recruiting to Northwestern outstanding graduate student applicants interested in big data and analytics. Thirty doctoral programs at Northwestern are invited annually to nominate between one and four students to the program, depending on the program's size. The fellowship includes an additional stipend beyond the typical graduate student stipend; as well as a research allowance to cover computer purchases, travel, or other costs. The program makes an effort to attract students from different departments to create an interdisciplinary community.

The aim of the Data Science Scholars program is to expand the domain-focused research portfolio of junior researchers and to establish their reputations as leaders in data science. The scholars, who are recent PhD recipients, are offered joint appointments for two-year terms (with a possibility of a third year) affiliated with the Northwestern Institute on Complex Systems and at least one more Research Center on campus that matches their area of expertise. Participants receive a full stipend and a research budget. The program receives applications from computer science, political science, astronomy, physics, and other fields, and the candidates are selected by the Steering Committee.

The faculty research networking luncheons support five thematic groups that meet monthly to identify important current and emerging research areas impacted by data science. During each lunch, two participants deliver short presentations to facilitate discussion. The groups focus on computational social sciences, quantitative biology, and sustainability; and comprise approximately 90 faculty representing 31 academic departments.

The DSI also runs an internal grant program to support new multidisciplinary collaborations in data science. The program provides funds for graduate students or postdocs working on a problem of interest to their mentors, representing different areas, and aims to transfer knowledge and expertise. The DSI has run four competitions, which funded 46 proposals at the level of $20,000 to $50,000 per award.

Finally, the center has established a close partnership with Research Computing Services (RCS), a group within Northwestern's central information technology (IT), with the goal of aiding in the offering of

training opportunities. RCS offers hands-on workshops that are open to all members of the university, and provides consulting on research projects to both students and the faculty.

Industry partnerships

The Kellogg School of Management and the Medill School of Journalism, Media, Integrated Marketing Communications have relationships with companies such as Facebook, Google, International Business Machines Corp. (IBM), IDEO, and State Farm. The DSI Steering Committee communicates with faculty who lead those efforts but has not yet taken action to launch a formal partnership program.

Lessons learned and future plans

The main programs supported by the DSI have been extremely successful. The Data Science Fellows program emerged as a powerful incentive to bring talented graduate students to Northwestern. The Data Science Scholars catalyzed programs that enhanced the experience of other members of the university and fostered new collaborations. The research networking luncheons are bringing faculty together across disciplines, and have already succeeded in catalyzing new collaborations and promoting the submission of research proposals to funders. The research funding program supported many new collaborations and the training of tens of graduate students.

The biggest lesson learned, however, has been the need to customize programs to different areas of scholarships. A program structure that will work just fine for physicists may not work as desired for sociologists. Over time, DSI has learned the value of gathering information from different groups so that the program design is attentive to their individual needs.

Looking toward the future, the biggest impact of the DSI will come from the numerous new connections that now join scholars across the university.

Translational Data Science Institute at the Ohio State University

https://tdai.osu.edu/

Mission

"The Ohio State University's Translational Data Analytics Institute brings together Ohio State faculty, students and industry and community partners to create data science and analytics solutions for global problems, develop a diverse and inclusive workforce, and advance scholarship. Together, we are using big data for good."

History and organization

In 2012 Ohio State University (OSU) committed to investing in three areas that could contribute to solving major societal problems: (1) food production and security; (2) areas of health and wellness such as infectious disease; and (3) energy and the environment, particularly in the areas of materials and sustainability. As all these problems could be more effectively addressed by applying data science techniques, the decision was made to invest in what ultimately became the Translational Data Analytics Institute (TDAI) three years later. At the time of the interview in December 2017, the institute was co-led by Interim Faculty Director Professor Raghu Machiraju, a professor of computer science and engineering, and bioinformatics and pathology; and Managing Director David Mongeau, who was recruited from the private sector to bring an industry perspective. This dual leadership model, designed to report to the Office of Research, was described to us as very effective in the design and build phases of TDAI.

OSU committed $125 million to create TDAI, which included funding to hire approximately 70 faculty, of which 53 had been recruited at the time of the interview. These faculty hold joint appointments and are co-funded by departments on a 50/50 basis. TDAI was also planning to hire research scientists.

In June 2018, TDAI moved into a new home made possible by a $40 million investment from the State of Ohio, with 21,000 square feet of space for collaborative research, teaching, community-building activities, and special events. This "living lab" environment includes software, hardware, and visualization laboratories; opportunities to test and leverage building-wide technologies for gathering and analyzing data; and reconfigurable furniture and technology for team projects of varying sizes.

Degree programs

After conducting a university-wide inventory of data science and analytics academic programming, TDAI is developing a professional science master's degree in translational data analytics that is expected to launch in 2019 to train mid-career professionals in the latest data-intensive methods. To understand the needs of the marketplace, the institute consulted with a dozen businesses, including Cardinal Health, Hewlett-Packard Enterprise, IBM, Nationwide Insurance, Ford Motor Company, and others. The degree program is offered with the Advanced Center for Arts and Design on the OSU campus and will incorporate design thinking into methods data science and analytics. Traditional graduate programs are also being planned and all of this programming is leveraging OSU's very successful undergraduate program in data analytics, which is currently in its fourth year.

Research and training programs

TDAI is building four–five "communities of practice" around Sensing and Smart/Connected Cities, Foundations of Data Science, Quantitative Life Sciences, and Complexity and Human Behavior. Each is led by a faculty member whose time TDAI reimburses in exchange for the responsibility of convening a

robust community of collaborators, defining and facilitating the community's research agenda, and identifying industry partners and major funding agencies to support this research.

TDAI also launched a seed grant program to foster transdisciplinary connections between faculty and students, and to create additional funding opportunities. Proposals must include at least two intellectually distinct disciplines from different colleges. As of 2018, TDAI has awarded more than seven such projects.

In summer 2018, TDAI hosted a free data science summer camp (with transportation and food provided) for girls in grades 8–10, with preference given to students from Columbus city schools. Participants worked in teams to develop hands-on experience with software tools and presented their work.

Industry partnerships

TDAI identified eight industry firms with which to establish large master agreements; and was nearing the completion of two large proposals, one in the technology space and one in workforce development. These agreements include components such as advisory roles for the partners, workspace for co-locating project teams at TDAI, student internships and other engagement opportunities, and in-kind gifts of technology.

Lessons learned and future plans

TDAI is endeavoring to ensure sufficient operational funding beyond the five-year start-up budget provided by the university; and to cultivate a sense of connectedness, purpose, and collaboration among a diverse population of new and existing faculty throughout a vast, highly decentralized university community.

Data Science Initiative at Stanford University

https://sdsi.stanford.edu/

Update

Stanford University launched a new and ambitious initiative in 2018 to create a major university-wide data science organization focused on advancing and connecting data science research to address the most pressing societal and scientific challenges. This new initiative was taking shape just as this study was wrapping up and this report was going to press.

Mission

"The Stanford Data Science Initiative aims to make Stanford a data enabled university. The Initiative advances data science methods and tools, and weaves them into the fabric of the university, to effectively respond to our most pressing societal and scientific challenges."

History and organization

The Stanford Data Science Initiative (SDSI) is one of several data science-related organizations at Stanford, along with the Institute for Computational & Mathematical Engineering (ICME), the Stanford Artificial Intelligence Laboratory (SAIL), the Data Analytics for What's Next (DAWN) program for enterprise-scale machine learning, and the Center for Artificial Intelligence (AI) Safety.

SDSI was created in 2014, late relative to other universities, because Stanford had established itself as a leader in data science decades ago, and the faculty/administration were satisfied with the university's position and reputation. However, it became clear in the past few years that more complex research problems required larger, interdisciplinary teams and more funding; and that it was becoming necessary to actively organize faculty around new, scientific challenges. As these discussions began to take shape, Stanford quickly realized that it needed someone who would own the design and administration of the program and Professor Hector Garcia-Molina agreed to be the Director (SDSI is currently led by Co-Directors Professor Jure Leskovec and Professor Euan Ashley). Under Professor Garcia-Molina's leadership, a working group with members from across the university developed the vision and organizational structure for SDSI. It emerged early on in the planning process that neither the faculty nor the administration were interested in creating a new department or school for data science, and the decision was made to establish an interdisciplinary initiative outside of the existing academic structure.

From the beginning, the planners were looking to support the initiative through corporate contributions, as they offered greater spending flexibility and potentially higher funding amounts than would be available from the university or the federal government. In addition, the Stanford community is very focused on real-life applications and making an impact on society, and therefore a program with strong connections to industry was seen as a good fit for this culture.

While many participating faculty members are still funded through grants, the bulk of SDSI support comes from industry membership fees. Two membership levels are available: regular at $100,000 per year and founding at $500,000 per year. The partners are encouraged to develop long-term, multi-year relationships with SDSI. Currently SDSI raises about $4 million in corporate support per year through its partnership with approximately 20 companies, including Accenture, American Family Insurance, Google, Hitachi, Microsoft, Toshiba, and Intel. The fees paid by companies are used to support research projects,

salaries for a small management team, and several activities such as seminars. Companies may provide funding above the membership fee to support research by SDSI faculty.

The SDSI is a "virtual" and "dynamic" organization, which has no dedicated physical space or faculty lines. The faculty are free to rotate in and out of the initiative depending on their interest in engaging with each other and with industry partners. The operating costs for the initiative are very low, and the bulk of corporate contributions is spent on research.

Degree programs

The critical element of the SDSI approach is to identify projects with intellectual and scholarly merit that are of interest to faculty and students (and which will not be perceived as work for hire), but which might offer practical solutions to societal and industry problems. The SDSI leadership team uses their knowledge of the research interests, career aspirations, and funding needs of a broad range of Stanford faculty to engage with the companies to help them identify problems that might be addressed through collaboration with Stanford. SDSI also issues internal (i.e., university) requests for proposals to bring in additional participants.

SDSI sponsors two types of projects: flagship projects, which include 6-8 faculty groups working collaboratively over a period of about five years; and small projects involving 1-2 groups. The current flagship projects are in precision medicine, social media and social networks, the "internet of things," and machine learning. Each faculty can obtain between $50,000 and $200,000 per year.

Other activities

Another component of SDSI is Data Commons – a repository of data and software tools for research and education. The goal of Data Commons is to collect and curate data, improve the accessibility of data, develop and share data processing tools, and create a community of developers and users of data. Finally, SDSI holds weekly seminars, retreats, workshops, and conferences to bring together faculty, students, and corporate funders.

Lessons learned and future plans

Each industry partner can allocate a portion of their fee to the research project(s) of their choice (this set-aside funding is called "tokens"). The goal of this approach is to signal the companies that they have control of their funding, and also to facilitate dialogue with faculty. However, rather than choosing to use tokens to support flagship projects, which was SDSI's expectation, the companies tend to invest in smaller unrelated projects, sometimes led by faculty not associated with the initiative. These somewhat random choices had the effect of fragmenting the SDSI portfolio of work. To focus the research, SDSI is working on identifying themes and types of challenges that tend to recur across the partners. In 2017–2018, the university's long-range planning process identified data science as an important area for further growth and investment.

Berkeley Institute for Data Science at the University of California Berkeley

https://bids.berkeley.edu/

Mission

"BIDS is a central hub of research and education at UC Berkeley designed to facilitate and nurture data-intensive science."

History and organization

The Berkeley Institute for Data Science (BIDS) was established in late 2013 with a $12.5 million grant from the Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation, as well as contribution from the University of California, Berkeley. In its first five years, BIDS has come to play a major role in bringing data science to the forefront of cross-disciplinary academic research and education at Berkeley.

BIDS is led by a faculty Director and a non-faculty Executive Director and reports to the Vice Chancellor for Research Office. An Executive Committee (EC), which includes four faculty members and other academic and professional staff, provides primary governance for the institute. In addition, over 50 other faculty members are affiliated as Senior Fellows and form a strong intellectual core for BIDS. They advise the EC in defining and guiding BIDS' efforts across campus and beyond. To plan and implement these efforts, BIDS draws upon Data Science Fellows, research staff scientists and engineers, and a professional staff. BIDS staff originally organized around six working groups: 1) career paths and alternative metrics, 2) education and training, 3) software tools and environments, 4) reproducibility and open science, 5) working spaces and culture, and 6) data science studies. These themes reflected perceived needs of the data science community and were share with two co-grantees, UW and NYU. As of 2018, some groups remain active, while other structures have emerged to meet discovered needs in areas such as image, graph, and text processing across research domains.

Since 2103, 55 BIDS Data Science Fellows have advanced data science methods and inquiry, expanded and built new open source software tools, propagated best practices, and trained others. The fellows are postdocs and graduate students who take 50% appointments with BIDS for two years, while continuing to pursue diverse interdisciplinary projects in the life, social, physical and computing sciences and in the humanities. They contribute to a strong "water cooler effect," learning new approaches and best practices from each other. For example, fellows have recognized how a modeling approach from quantitative biology is useful in social sciences, or approaches for managing spatial data in climate science is useful for neuroscience. Fellows have gone on to corporate, government, and not-for-profit careers and to tenure-track positions at universities such as Caltech, Harvard and Tufts. Research staff scientists and engineers and professional staff hold PhDs and other advanced degrees. They collaborate with Senior Fellows on research projects, implement open source software solutions for education and research programs, and lead their own research and development projects. Professional staff also account for organizational needs, such as communications, event programming, and human resources.

From its location in the University Library at the heart of the Berkeley campus, BIDS serves as a convening space for all constituents in the data science community, from undergraduates to nationally recognized faculty experts to non-university partners. The BIDS space welcomes over 1,500 unique visitors and hosts nearly 100 lectures, panels, and workshops annually. For example, BIDS and the Division of Data Sciences co-host the Berkeley Distinguished Lectures in Data Science there, a lecture

series featuring faculty doing visionary research that illustrates the character of the ongoing data, computational, inferential revolution. BIDS also hosts outreach events in the space, such as Black Girls Code and a recent state-wide California Water Data Hackathon.

Degree programs

While BIDS does not currently offer or manage degree programs, it plays an important role in data science education at Berkeley. Notably, BIDS fellows and staff developed the infrastructure supporting foundational courses in Berkeley's data science degree programs, such as Foundations of Data Science (Data 8) and Principles and Techniques of Data Science (Data 100). Moreover, Senior Fellows are PIs on major NSF training grants and national workshops regarding the pedagogy and practice of data science.

Research and training programs

In addition to its open source software projects to enable data science research broadly, BIDS funds interdisciplinary research project teams led by faculty members in specific fields of inquiry, including hydrology, global biodiversity, and physical systems. BIDS projects target high-impact, interdisciplinary data science research involving multiple campus departments or entities.

BIDS also hosts many experiential learning programs and training for students. Its data science skills training, which augments degree programs, includes Data and Software Carpentry workshops, intensive 2-day sessions to learn basic research computing tools and techniques, such as design, version control, testing, and automation; projects that allow undergraduate students to work with Data Science Fellows to gain research experience; and weekly peer-learning sessions, for example, in Python and R programming, visualization, data management, and version control.

Industry partnerships

BIDS benefits from support from two corporate partnerships and intends to cultivate additional partners who share common or complementary goals around data science.

Other contributions

BIDS is strongly committed to developing and disseminating tools and practices to enable open science and reproducible research. Notable evidence of this commitment is "The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences," as well as its open source software contributions to Binder, Jupyter notebooks, Numpy, scikit-image and rOpenSci.

Lessons learned and future plans

BIDS is assuming an increasingly important role in the data science landscape at Berkeley, considered integral to the university's emerging strategic plan. The influence of BIDS also extends beyond the university, as suggested by the following:

- Global dissemination of open source software borne or extensively developed at the institute – including Jupyter Notebooks, scikit-image and rOpenSci
- Influencing acceptance of new criteria of merit for tenure-track faculty positions
- Launch of a cross-University of California system discussion to establish new career positions and tracks to incent data scientists to remain in academia.

Computation Institute and Center for Data and Applied Computing at the University of Chicago

https://www.ci.uchicago.edu/

Mission

"The Computation Institute (CI) was established in 2000 as a joint initiative between the University of Chicago and Argonne National Laboratory to advance science through innovative computational approaches."

History and organization

The Center for Data and Applied Computing (CDAC), which launched in the summer of 2018, is replacing the Computation Institute. CDAC will have a dual role: as an administrative mechanism for bringing scientists from national laboratories into the University of Chicago (UChicago); and as a hub for resources, space, and collaborations for computationally focused interdisciplinary projects on campus.

CDAC will be based in the renovated John Crerar Library, which is scheduled to open in fall 2018. The center will work closely with the Department of Computer Science (and co-locate with it), while bridging other departments and divisions within the university. Funding for CDAC will come from the Provost's office, augmented with federal grants, foundation support, and other philanthropic resources. A group of faculty from across the university will serve as the governing board. CDAC will be led by a full-time Executive Director and a part-time Faculty Director. It will not have its own faculty (at least initially), because the model of joint appointments at the CI is seen as not very effective. Faculty based at various departments at the university will have affiliations with CDAC.

Degree programs

The CI participated in several joint master's degree programs including computational analysis and public policy, business analytics, and computational social science, one or more of which will likely continue to be offered by CDAC. The center also plans to manage an undergraduate minor in data science and to offer a data science certificate program for doctoral students modeled after the University of Washington.

Research and training programs

The CI, CDAC's predecessor, served as an umbrella for several projects. These include the Center for Urban Data, which works closely with the City of Chicago and involves people from public policy, economics, and computer science who analyze large datasets. Another project is a collaboration with the medical school, called the Genomic Data Common, which is slated to be one of the largest repository for human cancer data. There is also a project called the Knowledge Lab, joint with social sciences, which is focused on text analytics of literature to learn about the production and evolution of science. Finally, there is the Data Science for Social Good (which has been copied by other universities), an educational project for graduate students to work on real-life problems. Some of these programs, particularly those interdisciplinary in nature, will transition to CDAC. Others will move to a corresponding department.

CDAC will support three types of programs. These will include low-cost activities, such as workshops, to bring together researchers to brainstorm ideas. In addition, CDAC will run seed grant competitions, which will support postdocs or students to do initial work on new ideas. Finally, there will be a mechanism to support larger projects that bring in external funding, but CDAC will need staff support, space, and other resources. The projects to fund will be chosen by a faculty committee drawn from across the university.

The center will also provide opportunities for researchers from national labs to collaborate with the UChicago faculty on data science projects.

Industry partnerships

The CI had an active entrepreneurship program to incubate startups stemming from the work of the UChicago faculty. CDAC is planning to have an industrial sponsorship program similar to UC Berkeley.

Lessons learned and future plans

It was informative to learn that joint appointments between the CI and departments, while bringing in talented faculty, were not effective in creating a community. These joint faculty ended up fully transitioning into the departments.

Center for Data Science at the University of Massachusetts Amherst

https://ds.cs.umass.edu/

Mission

"The Center for Data Science at the University of Massachusetts College of Information and Computer Sciences is creating new technology to manage and gain insight from "big data" while also educating tomorrow's data scientists."

History and organization

The Center for Data Science (CDS) is administratively based in and is responsible to the College of Information and Computer Sciences. The decision to place it within one university unit was intentional, based on the desire to produce students with in-depth computer science training, as well as the desire for nimble action enabled by avoiding multi-college consent on all decision-making. The University of Massachusetts Amherst (UMass) is explicitly taking a "diverse" rather than "one-size-fits-all" approach to data science – there are distinct but coordinated data science efforts in Computer Science, Statistics, Engineering, Social Sciences, Public Health, and the School of Management, where each can develop a program best tuned to the needs of each field. However, at the same time CDS serves as the data science central hub and coordinating entity on campus. Two-hundred-and-fifty faculty members from across five colleges are affiliated with the CDS. The center is directed by Professor Andrew McCallum, an ACM fellow and past President of the International Machine Leaning Society. The mission of the center is "to facilitate the highest quality data science education, research and industrial collaboration." It has partnerships with a dozen centers and laboratories at the university.

The idea of the center was proposed by Professor McCallum to the Provost and Chancellor, who readily agreed. CDS launched in April 2015 and 300 people attended the inaugural event, including leadership from major technology companies, such as Google, Amazon, Thompson Reuters, and Microsoft; National Science Foundation (NSF) staff; a representative from the State of Massachusetts; venture capitalists; and start-ups. The event was reported in the Boston Globe as a $100-million effort by the university.

The center is supported through several sources. The university is providing $1 million in funding annually for staff, operations, equipment, and events. This level of support is expected to continue indefinitely. In return, the college portion of overhead from the grants awarded to CDS faculty goes to the Provost. This financial arrangement was put in place to avoid negotiation across the schools about sharing overhead and to remove obstacles to faculty collaboration across the university. In addition, the university is providing computer science with six new tenure-track faculty slots (one per year).

Over the past several years the university developed connections to a nearby major insurance company, Massachusetts Mutual Life Insurance (MassMutual), by providing technical guidance in machine learning. MassMutual committed $15 million over 10 years to support the center, which will fund an additional 8 faculty lines, bringing the total number of new data science tenure-track faculty in computer science to 14. Other substantial corporate investments are providing over $10 million in research funding; these partners include IBM, Pratt & Whitney, Google, Oracle, Microsoft, and Amazon, as well as a $5.5 million grant from the Chan Zuckerberg Initiative for a project called Computable Knowledge, which aims to create "intelligent and navigable map of scientific knowledge" using artificial intelligence tools. The center also receives support through its Industry Affiliate Program, further described below. Finally,

the Massachusetts Technology Collaborative (MassTech) donated $5 million for capital expenditures, which has been used to build a computer cluster with over 800 GPUs in support of deep learning research.

By design, the center does not have its own space. However, commercial co-working spaces are within walking distance from the campus, which are used by the community. MassMutual is building a research laboratory also within walking distance, and some space might be made available in that building.

Degree programs

The college and CDS created a Master of Science concentration in data science within its computer science master's program. Our respondent at CDS told us that he was reluctant to create a master's degree named "data science" because of wariness about adding new degree names to the university's accumulation of such names, and also because this name does not necessarily imply a depth of training in computer science that employers value and can expect from a student with a degree in computer science. CDS has an arrangement with MassMutual, whereby undergraduates are hired by the company for two years to work full-time on MassMutual projects of interest to them while they complete the Master of Science concentration. The university also offers a Certificate in Statistical and Computational Data Science (jointly between computer science and statistics), which requires completion of four relevant graduate-level courses.

Research and training programs

The center focuses on data science applications in the areas of education, training, and workforce; health and biomedicine; food, water, and energy (sustainability); and prediction, risk, fairness, accountability, and transparency. It also continues to build capacity in machine learning, parallel and distributed computing, and data science theory.

Every year or two, the center plans to launch a new research initiative to bring together multiple partners. The first such initiative, which began in 2017, is in workforce analytics – focusing on gathering and analyzing data about career paths in STEM fields. While this initiative was launched by CDS and is administratively driven by it, its purpose is to bring together collaborators from across the university. As next themes (for approximately 2019), CDS is considering wearable health monitors, sustainable agriculture, and efficient cloud systems.

Most of the research-related efforts of the center at the time of the interview have focused on hiring new faculty. The center also runs a distinguished lecture series in data science, weekly machine learning seminars, and a weekly data science tea. Finally, CDS-affiliated faculty collaborate with industry on several research projects.

CDS does not have a seed grant-type program, which had been considered, but determined to be unnecessary, as faculty are successful at obtaining external funding.

Industry partnerships

The center has performed extensive outreach to industry (one of the center's Boards is tasked with making industry connections). An industry affiliates program was created in the first year and currently includes approximately 15 members (such as Google, Amazon, IBM, Oracle, Microsoft, and also many smaller companies), each paying an annual fee. The partners are invited to an annual career mixer, a recruiting event held in the fall that involves graduate and advanced undergraduate students. Students present posters to company representatives in the first half of the evening, and discussions with industry

scientists continue afterward at the company tables. The industry partners are also invited to attend a research symposium each spring. According to the interviewee, these events have been successful as a source of internships opportunities and job offers.

Funds from the industry affiliates program are used to support the Data Science Postdoctoral Research Fellowships. Computer science at UMass does not have as many postdocs as other schools, so the program was created to increase their number. Faculty members and candidates apply together to obtain two years of funding.

Lessons learned and future plans

Our respondent believed that, while not necessarily the right choice for all universities, there are significant advantages to avoiding a cross-campus institute model. To support this view, he gave an example of the master's program in data analytics offered by the UMass business school. While that program does teach students data science, the training is very different than a similarly named program in a Computer Science Department. So, rather than trying to force different schools or departments into the same model, it is better to let each effort takes its own shape, while simultaneously coordinating data science education and research across the campus.

Another point made in the interview was that CDS chose not to create joint faculty appointments out of concern that these can be detrimental to faculty, especially at a junior level. This type of arrangement can cause complications during the tenure process, and having a sense of belonging to a department is important to faculty development. Rather than using joint hiring to promote collaboration, CDS implements cluster hiring (brining in several faculty in similar or complementary fields simultaneously). These faculty are affiliated with one department, but are hired with an understanding that they would be part of a bigger culture that encourages collaboration. CDS regularly organizes "faculty soirees" that bring together data-science-related faculty from across campus, as well as Smith, Mount Holyoke, Hampshire, and Amherst colleges.

The center is growing. In spring 2017, it had open searches for four tenure track computer science faculty in data science (of particular interest were candidates who can build bridges to other disciplines). Similarly vigorous hiring was expected for the coming years. (Other departments on campus who are also aggressively hiring in data science include the Social Sciences, Statistics, and the Business School.) In addition, CDS is interested in improving undergraduate education, and UMass is importing UC Berkeley's Data 8 to be taught jointly by Computer Science and Statistics in both the fall and spring of the coming academic year.

Michigan Institute for Data Science at the University of Michigan Ann Arbor

https://midas.umich.edu/

Mission

"The Michigan Institute for Data Science (MIDAS) is the focal point for the new multidisciplinary area of data science at the University of Michigan. This area covers a wide spectrum of scientific pursuits (development of concepts, methods, and technology) for data collection, management, analysis, and interpretation as well as their innovative use to address important problems in science, engineering, business, and other areas."

History and organization

MIDAS was created in July 2015 as part of the University of Michigan (UMichigan) Data Science Initiative. A proposal to establish a university-level data science initiative, prepared by a university-wide faculty committee, was submitted to the Provost in November 2014. Funded by the Provost, the Data Science Initiative established MIDAS as the academic home for data science on campus that also includes coordinated data science consulting and infrastructure services offerings for faculty and students through Consulting for Statistics, Computing & Analytics Research (CSCAR); and Advanced Research Computing – Technology Services (ARC-TS). Together, MIDAS, CSCAR, and ARC-TS provide a coordinated and comprehensive home for data science as part of Advanced Research Computing (ARC) in UMichigan's Office of Research.

UMichigan provided $20.5 million to establish MIDAS, of which $15.5 million was funded centrally through the Provost's office with an additional $5 million provided by participating schools and colleges. The Provost also provided another $11.6 million to enhance the data science consulting services of CSCAR and the data science infrastructure services of ARC-TS. The MIDAS governance structure includes two faculty Co-Directors, a staff Managing Director, a faculty Associate Director, three full-time staff members, an Executive Committee comprised of UMichigan leadership at multiple schools and colleges, a Management Committee, an Education and Training Committee, and an External Advisory Board.

MIDAS recruits data science faculty by partnering with the university's schools and colleges that have hiring authority at the university (seven faculty have been hired). MIDAS contributes to startup funds to augment recruitment packages. Recruited faculty have a tenure-track appointment in one of the university's 19 schools and colleges, with a joint research appointment in MIDAS. Seven faculty recruited to the university, together with another 19 existing faculty, comprise the institute's core faculty members. Each MIDAS core faculty member contributes to the programmatic goals of MIDAS (research, education, industry engagement, and outreach). MIDAS also has 200+ affiliate faculty from the Ann Arbor, Dearborn, and Flint campuses.

MIDAS has a space of 2,200 square feet that is co-located with the Center for the Study of Complex Systems and the Center of Applied and Interdisciplinary Mathematics. The space includes offices, meeting rooms, and open collaboration areas.

Degree programs

MIDAS participates in undergraduate and master's in data science degree programs offered through the Departments of Biostatistics, Statistics, Mathematics, and Computer Science and Engineering; as well as

the School of Information. MIDAS also operates a Graduate Certificate in Data Science program, and offers a joint postdoctoral training program with the Chinese University of Hong Kong-Shenzhen.

Research and training programs

MIDAS has established a data-intensive Challenge Initiatives Program that aligns with university research strengths and priorities in transportation and mobility research, precision health, computational social science, and learning analytics where data science has the potential to have significant impact. Through two rounds of internal competition and funding, MIDAS has funded over $10 million worth of research awards to nine different projects that bring together highly multidisciplinary groups of researchers, and collectively involve 75 faculty members and 79 students and postdocs. These teams, chosen for their potential scientific, educational, and societal impact, are attracting more collaborators within and outside of UMichigan, as well as substantial extramural funding from federal and state agencies, and industry.

In Phase II of the research support, MIDAS launched the Data Science for Music Initiative, awarding a total of $300,000 to four different interdisciplinary research projects that apply data science methods to research at the interface between data science and music theory, composition, performance, and audience participation. These projects involve faculty from the School of Music; the School of Information; the College of Literature, Science, and the Arts; and the College of Engineering. MIDAS will fund additional data science research projects that organically emerge through the interaction with the UMichigan data science community.

To further catalyze data science research and build community, MIDAS organizes working groups based on data science methodology and application themes. These include “Data Integration,” “Mobile Sensor Analytics,” and “Trustworthy Data Science” that are creating new crosscutting research ideas and teams. Finally, MIDAS hosts a large number of events (sometimes jointly with other research units) to build the UMichigan data science community, including research symposia, conferences, and seminar series.

MIDAS also supports four data science student groups focused on “outside the classroom” experiences that include tutorials, competitions, and, projects that teach practical data science skills by solving impactful problems for societal good. With a combined membership of more than 400 students and 50 faculty members, these student groups have completed 14 public service projects across Southeast Michigan that include Flint Lead Level Prediction and Service Line Localization projects, City of Detroit Blight Compliance and Vehicle Maintenance projects, as well as food pantry location optimization.

Finally, MIDAS organizes a data science summer camp for high school students that includes outreach to and scholarships for economically disadvantaged families in southeast Michigan.

Industry and other partnerships

MIDAS works closely with the UMichigan Business Engagement Center to identify industry partners for both sponsored data science research and talent recruitment. Under a general framework, MIDAS crafts collaboration models that fit the goals of each industry partner. MIDAS has enabled joint research projects and large initiatives with industry funding for UMichigan researchers. The level of industry sponsorship for these projects range from a few hundred thousand dollars to millions of dollars. MIDAS also has an expanding corporate affiliates program where companies can engage more deeply with MIDAS on workforce development, student recruiting, and scientific exchanges. Industry representatives have participated as attendees and speakers of MIDAS activities that include working groups, conferences, seminars, and student meetings.

MIDAS faculty and students have collaborated at the state and local levels, including the cities of Detroit and Ann Arbor, to provide data science expertise. MIDAS is a founding member of the NSF-funded Midwest Big Data Hub and provides leadership in several focus areas. MIDAS faculty leaders have participated in data science strategic planning at the national level through the National Academies of Science, Engineering, and Medicine; NSF; and advisory boards at other universities. Internationally, MIDAS has developed formal collaboration relationships with the Chinese University of Hong Kong-Shenzhen and the University College London.

Lessons learned and future plans

MIDAS is in an excellent position to bring focus to a wide spectrum of data science issues and to catalyze innovative multidisciplinary research at the university level. In this vein, MIDAS has recently focused on supporting research in the society-impacting areas of data science that include "data for good," data security, fairness and privacy, and computational reproducibility. These are extremely important research questions for our day, but few companies or federal agencies consider these as strong priorities for funding. It will be important for MIDAS to strike the right balance between investments in data science research with long-term societal impact versus research that might generate more immediate commercial interest. MIDAS has been very successful, yet faces the challenge of sustaining its activities over the long-term. In a landscape of competing interests, sustained commitment by the university that goes beyond initial investment is not guaranteed. Industry and federal agencies cannot be depended upon to fund all of MIDAS's activities in research, education, and outreach. Over the next two years, MIDAS will focus on sustaining its successes through extramural funding, continued internal investment, and philanthropic engagement. Specific future plans include:

- Continue to partner with schools and colleges to improve recruitment and retention of top data science faculty
- Develop new programs that accelerate building data science capacity on campus that benefits a significantly larger portion of the UMichigan faculty
- Partner with key data science teaching faculty across the UMichigan to establish a data science PhD program, and launch a MIDAS-accredited Data Science Certificate Program for working professionals supplemented by tailored professional development workshops
- Build synergy with research units and initiatives on campus through co-sponsorship of research, training, and conferences; as well as informal gatherings that further develop a vibrant data science community
- Increase industry collaboration by expanding the corporate affiliates program and growth in industry-sponsored research and tailored training programs
- Strengthen established local, state, regional, national, and international outreach activities to provide enhanced research; and educational opportunities for faculty and students.

Data Science Initiative at the University of North Carolina Charlotte

https://dsi.uncc.edu/

Mission

"The Data Science Initiative brings academia and industry together to turn data into knowledge, and knowledge into insight to see what's possible in the new digital age."

History and organization

The Data Science Initiative (DSI) started in 2012 with the graduate certificate program in data science, to which a master's in data science and business analytics program was added the following year. The University of North Carolina Charlotte (UNCC) has launched a number of programs in data science since 2000, before the term "big data" became popular, and DSI was created to unite them. In addition, local energy, retail, logistics, and healthcare industries conveyed to the university their need for workers with data science skills.

DSI is the university-level entity directed by Professor Mirsad Hadzikadic (Software and Information Systems), who reports to the Provost of Academic Affairs. Its staff include two data scientists who do not conduct research, but provide services and support to the community, particularly in developing computational infrastructure. DSI has two kinds of faculty: those hired into DSI (called "data science faculty") and affiliated faculty, who were already at the university but have an interest in data science-related areas. While data science faculty are based in departments through which they get tenure, these departments "owe" a certain level of teaching to DSI, which can be distributed across several faculty.

Because DSI is not an academic unit, all educational programs have been based at the graduate school. However, as DSI is interested in expanding its offerings to undergraduate and PhD programs, this arrangement is becoming unsustainable. The university considered various options, including a new institute, college, or department. Ultimately, DSI will be managed by the College of Computing and Informatics and the College of Liberal Arts and Science; and have all responsibility for interdisciplinary university-wide degrees. The vision of the DSI Director is that all faculty at the School of Data Science will have joint appointments between the school and another discipline. They will be officially affiliated with one or the other, their service obligations will be split between the two, and the tenure will be defined as a joint appointment.

DSI currently has small space in one building, enough for 10 offices; it also has several laboratories distributed across the university and is looking for additional space.

Degree programs

Currently, DSI offers master's and graduate certificate programs in data science and business analytics, and in health informatics. When the School of Data Science is established, all master's programs will be divided into two parts: technical courses (techniques) and domain courses. Mixing and matching these two components (a total of about 10 courses) will enable students to obtain degrees that best fit their interests, such as crime analytics, anthropology analytics, or health analytics. DSI is working with domain departments interested in launching these Master of Science programs and a similar model will be used for the upcoming doctorate and undergraduate programs.

Research and training programs

In most years DSI provided seed funding for several research pilots, which needed to involve people from at least two colleges. DSI also established a Data Science for Social Good program modeled after the University of Chicago; and hosted a summit and an open house focused on social good, which were well-attended. (Charlotte is interested in economic mobility because it is not doing well as a city, so DSI and the university are trying to help.)

Industry partnerships

Like most professional science master's programs, Business Analytics and Health Analytics have industry boards that include 15 companies. In addition, the College of Computing and Informatics has a partnership program (called CCI partners), with companies paying a fee to have access to students through career fairs and other activities. DSI is currently using this program, but is also signing up its own partners. In contrast to CCI, which is designed to offer access to students, DSI is planning to include additional elements, such as an "innovation lab," through which companies can invest in research projects. In addition, DSI is competing to become an industry-university cooperative research center funded by NSF (companies pay $50,000 to participate and decide collectively with the university how the funds are spent), and is hopeful of the positive outcome.

Other

DSI developed a System for Observation of Populous and Heterogeneous Information (SOPHI), which enables the storage and use of large volumes of structured and unstructured data in virtually any format.

Lessons learned and future plans

We were told that this type of initiative should be promoted from the university leadership to take root because faculty are inherently discipline-based and have no incentives to "step out" of one area.

The School of Data Science is expected to be in place in fall 2018, and new programs will be offered soon after. The goal of the new school is to redefine how UNCC educates students and to be responsive to the needs of its community. The research mission of the school will also continue to grow in the next three years.

Goergen Institute for Data Science at the University of Rochester

http://www.sas.rochester.edu/dsc/index.html

Mission

"…the Goergen Institute for Data Science is home to interdisciplinary data science research and the interdepartmental data science academic programs."

History and organization

In 2013, the University of Rochester (URochester) announced that it was committing $50 million to expand its activities in data science, which included establishing the Institute for Data Science, subsequently named the Goergen Institute for Data Science (GIDS) after one of the donors; constructing new building; and hiring approximately 20 faculty. The Center of Excellence in Data Science (CoE), which is funded by the New York State Department of Economic Development, is housed within GIDS. The institute was founded and directed until 2018 by Professor Henry Kautz, a well-known artificial intelligence researcher and former president of the Association for the Advancement of Artificial Intelligence. The university is currently searching for a new director.

In 2017 GIDS moved into its new building, Wegmans Hall, built with support from the university and the Wegman Family Foundation. The Hall includes space for collaborative research, conferences, workshops, and public events.

The university is hiring 1–2 new data science faculty per year and 14 positions have been filled to date. A committee representing 5–10 departments identify and recruit interdisciplinary faculty whose work uses computational tools, but who are not necessarily a good fit for a single department. For example, in 2016 a new faculty member was hired who works in digital restoration of ancient manuscripts. Approximately 50 faculty are affiliated with GIDS. Faculty members are hired fully into departments, not GIDS, and are tenured or tenure-track. GIDS has hired a full-time Professor of Instruction and three full-time data scientists to work on industry-academic collaborations.

Degree programs

The university offers an undergraduate major in data science (both BA and BS), which combines required coursework in computer science and statistics with a concentration in business, biology, earth and environmental science, or political science. The students enrolled in these programs have an opportunity to work for a semester with industry partners on capstone projects, which use real-life problems and data. GIDS expected approximately 60 undergraduates to enroll in this major in the 2016–2017 academic year. Undergraduates can also participate in a summer research experience through the NSF-funded program in data science called "Computational Methods for Understanding Music, Media, and Minds."

The university also offers an MS program, which can be completed in one year or a year-and-a-half. The students can either study data science broadly or focus on computational and statistical methods, health and biomedical sciences, or business and social science. The program planned to admit 30–50 students for the 2016–2017 academic year. Starting in 2018, the university began offering an MS degree in data science (on a full scholarship) to a small number of qualified PhD students. Finally, the university has a training grant from NSF (Data-Enabled Science and Engineering) to support data science-focused doctoral training in computer science, and brain and cognitive science.

Research and training programs

GIDS brings together research groups, centers, and programs from across the university and sees itself as an umbrella entity for collaborative research. The initially chosen areas of focus in data science for the university and for GIDS include health analytics; cognitive science and artificial intelligence; and methods, tools, and infrastructure to handle large-scale data. Each domain is represented by several distinguished researchers.

In 2016, GIDS established an internal seed grant program called the Pilot Award Program in Health Data Analytics in response to opportunities and challenges identified during a community retreat. The goal of the program was to support new cross-campus multidisciplinary collaborations on projects that use data for prediction of health outcomes. The awards were $25,000–50,000 for one year. All tenure-track faculty at the university were eligible to apply, and preference was given to projects with a high probability of follow-up funding. The projects could not be an extension of ongoing research. This award program is ongoing and is now in its third year.

Finally, GIDS sponsors various training and social community events such as hackathons, talks, tutorials, and discussions held in the Rochester area (including at local restaurants).

Industry partnerships

The industry partnership efforts are led by the New York State Center of Excellence in Data Science and include activities that support research collaborations, student capstone projects, and internships; and offer access to commercialization expertise. Industry partners are invited to participate in the annual career fair.

In fall 2017, New York State announced that it was investing $22.5 million to create the Rochester Data Science Consortium, a partnership between URochester and industry aimed at fostering regional economic growth. The university side of the consortium has its home at GIDS. The first industry partner in the consortium is the Harris Corporation's Space and Intelligence Systems Division.

Challenges and lessons learned

The biggest challenge was ensuring that all departments were aware of GIDS activities. Because everyone at the university is busy, it is not sufficient to simply post the information on the website or include it in a newsletter. Rather, it is important to use multiple channels through which to actively engage department chairs and faculty.

Data Science Institute at the University of Virginia

https://dsi.virginia.edu/

Mission

"The mission of the Data Science Institute is to achieve excellence in data-driven research and education through solving important problems and providing the workforce of tomorrow."

History and organization

The Data Science Institute (DSI) resulted from the recognition by the university that this was a growing area. The vision of the founding director Professor Donald Brown and the university administration was to create a collaborative, pan-campus initiative interwoven into all 11 schools at the University of Virginia (UVA). DSI was established in 2013 and is funded through an endowment, revenue from the MS program, and some support from the private sector; the institute is self-sufficient. Since 2017, it has been directed by the Stephenson Chair of Data Science, Professor Philip Bourne, from the Department of Biomedical Engineering. In addition to the Director, DSI staff include an event planner and communications, admissions, and career placement specialists.

DSI has its own not tenure-tracked "general" faculty (the institute cannot grant tenure) who do teaching and some research. It has also begun to hire faculty who will have dual appointments with departments (50/50 share). The institute has 4 general faculty and 10 joint faculty, and plans additional hires. DSI and the home department make joint decisions about hiring and promotion, and have an arrangement to share indirect costs on grants and teaching responsibilities. DSI is interested in creating non-traditional tracks for data scientists to accommodate researchers from the private sector, who are particularly well-suited for these jobs, but may not have a record of publications or independent funding. We were told that it is challenging to hire this type of researcher at a university, but that DSI has much support among the administration and it should be possible to create these positions.

DSI has its own space, where faculty and staff reside together (following the original "academical village" concept of Thomas Jefferson who founded the university). The DSI Director shares space with students.

Degree programs

DSI offers several masters-level programs: an 11-month MSDS, which includes capstone projects under the supervision of academic and industry mentors; a 2-year MBA/MSDS jointly with the School of Business; and an 11-month MD/MSDS jointly with the School of Medicine. In partnership with other schools at UVA, DSI also offers a PhD/MSDS program.

Research and training programs

In collaboration with the Office of Vice President for Research, DSI offers Presidential Fellowships in Data Science. The program brings together graduate students from different departments to work on collaborative projects involving data science. The vision for the program is for the students to tackle complex interdisciplinary projects with potential social impact. Fellows work together on designing proposals, which are reviewed by a panel of faculty for clarity, relevance, potential impact, innovation/novelty, and composition of the team. Thirteen students received the fellowship in academic year 2017–2018.

Industry partnerships

At the time of the interview, DSI industry partnerships were limited to capstone projects, but the institute was working with the development office to create additional opportunities. The placement of students in industry is expected to grow DSI's connections in this sector.

Lessons learned and future plans

We were told that understanding the political landscape and convincing the departments that they will benefit from a data science institute is the greatest challenge. Our respondent believed that these types of initiatives should aim to improve student experiences in addition to building research capacity in data science.

DSI is establishing an open data laboratory and an online MS program. With the hiring of additional faculty and an Associate Director of Research, the institute is in the process of expanding its research portfolio, while continuing with its successful educational programs.

eScience Institute at the University of Washington

https://escience.washington.edu/

Mission

"The eScience Institute empowers researchers and students in all fields to answer fundamental questions through the use of large, complex, and noisy data. As the hub of data-intensive discovery on campus, we lead a community of innovators in the techniques, technologies, and best practices of data science and the fields that depend on them."

History and organization

Launched in 2008, the eScience Institute (eScience) was able to significantly expand its staffing and programs when it was awarded a $12.5 million grant by the Alfred P. Sloan and the Gordon and Betty Moore foundations to become one of three Moore Sloan Data Science Environments (MSDSEs). Shortly thereafter, eScience was awarded a $9.3 million grant from the Washington Research Foundation (WRF) and a $2.8 million National Science Foundation (NSF) Integrated Graduate Education and Research Traineeship (IGERT) award. eScience is located in the WRF Data Science Studio, which was renovated in 2014 with funds from the WRF. Its location in a former library space on the top floor of the University of Washington's (UW's) Physics/Astronomy building is consistent with eScience's clear focus on advancing discovery as well as methodology. The open and modular layout of the space was intended to signal inclusivity to the university community and to foster collaboration.

A faculty Director, two Co-Executive Directors, and a Director of Research form the leadership team at eScience. All key decisions related to the Institute are made by an Executive Committee, composed of the leadership team and approximately 10 additional faculty representing a range of departments. The non-faculty participants include administrative staff, data scientists, research scientists, postdocs, and students. Postdocs are recruited externally for a two-year term and have dual mentors chosen at the time of application – a domain scientist and a methodologist. Data scientists and research scientists are similar in qualifications and duties, but the former are typically recruited externally and fully funded by eScience, while the latter come from within UW and are co-funded with other units. Data scientists and research scientists are the engine of eScience, leading most activities and programs, while simultaneously carrying out their own research. eScience puts in place several mechanisms to support these staff. One of these is a "salary buyback program," whereby a portion of the grant funding obtained by data scientists is returned to them as a research stipend. In addition, data scientists can be granted PI status, allowing them the opportunity to direct their own work and apply for independent funding. Finally, their salaries have been adjusted to be more competitive with industry.

eScience activities are organized around several working groups developed during the design phase of the MSDSE. These include (1) Career Paths and Alternative Metrics, (2) Education and Training, (3) Software Tools and Environments, (4) Reproducibility and Open Science, (5) Working Spaces and Culture, and (6) Data Science Studies. Each working group includes several staff who plan and implement its agenda.

Degree programs

eScience does not offer or manage its own degrees or courses. However, eScience's Education and Training Working Group developed the framework and led the creation of new concentrations and degree programs in data science across UW. At both the undergraduate and graduate levels, students can receive

their degrees with a "Data Science Option" noted on their transcript (e.g., Bachelors in Biology, Data Science Options) by completing specific courses in areas such as statistics, machine learning, data management, data visualization, and ethics/privacy. Data Science Options are currently in 15 departments at UW, with many more in progress. A master's in data science is also offered, on both a full- and part-time basis. Two eScience Executive Committee members led the launch of this program and continue to serve on its 3-person Executive Committee.

Research and training programs

eScience offers a wide variety of research and training programs. These include semi-annual, 10-week "incubators," which bring together interdisciplinary teams of data scientists, faculty, and students to work on a joint research project. The Winter Incubator, which has been running since 2014, has supported 29 projects across 21 UW departments/organizations. The summer Data Science for Social Good (DSSG) incubator program draws projects from academia, government, industry, and nonprofits. Each summer the DSSG supports 3-4 teams to focus on problems with potential societal impact and is modeled after the DSSG program created by the University of Chicago. These incubator programs are very popular.

In addition, eScience hosts Software Carpentry workshops – intensive 2-day, hands-on sessions to teach basic computing skills. eScience also offers office hours, where anyone at UW can drop in for consultation with a data scientist. Finally, eScience helps organize Hack Weeks and "xD" (cross-domain) workshops with their MSDSE peers at New York University and the University of California Berkeley. These are participant-driven events, usually with a broad disciplinary or methodology focus, where attendees learn from their peers through lectures, open-ended projects, networking, and discussions. So far, eScience has led the organization of Hack Weeks with a focus on astronomy, neuroscience, geospatial science, and oceanography.

Industry and other partnerships

The DSSG program described above is partially supported through external partnerships, and is on a path to full sustainability. Sponsors of the program include Cascadia Urban Analytics and Microsoft. The Bill & Melinda Gates Foundation has supported research projects at eScience that have emerged from the DSSG program. Another eScience partner is Microsoft Research, which supports the development of "database-as-a-service" technology.

Other contributions

eScience staff and faculty are very focused on the development of data science tools and on the practice of reproducible research. They have produced software tools and datasets in a wide range of fields, including astronomy, physics, biology, oceanography, neuroscience, psychology, mathematics, medical imaging, computer science, ecology, and geoscience. eScience also contributed case studies to a book describing practices in reproducible research with the other two MSDSEs.

Lessons learned and future plans

The dual mentorship model of postdocs was most effective when the two mentors had a prior relationship. eScience also found that postdocs do not spend much time in the space, and was looking for strategies to increase their engagement.

eScience recently secured a financial commitment from the university to support all its staff and programs for the foreseeable future. To be able to continue experimenting with various offerings and career tracks, eScience is looking to supplement this core budget through grants and partnerships.