Data Science Leadership Summit
Summary Report

Jeannette M. Wing, Vandana P. Janeja, Tyler Kloefkorn, and Lucy C. Erickson
7 September 2018

Executive Summary

Data science is a burgeoning field. As a result of recent technological advances, widespread and accelerated uptake of these technologies by many sectors, and increasing workforce demands, many data science initiatives across universities and colleges in the US and beyond are sprouting up at a rapid pace. The Data Science Leadership Summit, hosted on March 26, 2018 by the Data Science Institute at Columbia University, was the first convening of leaders of these initiatives. The Summit was co-funded by the National Science Foundation, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation.

The goals of the Summit were:

- To initiate the formation of an academic community for data science;
- To share best practices among academic leaders who face similar challenges and opportunities; and
- To take collective responsibility in the broader effort to prepare next-generation data scientists to contribute in the best interests of society.

This meeting was intended to be the inaugural meeting of a regular series. Also, given the existence of prior workshops and reports on data science, the intention was to minimize repeating what had been said before, but at the same time, set for all attendees a common level of understanding of the state of data science in academia. One of the major outcomes of this meeting is the realization that many of the universities represented at the Summit are just now working through challenges in establishing a data science effort on the participants' respective campuses; at the same time, participants recognized the tremendous opportunity to help shape the field of data science and respond to the overwhelming excitement for data science in academia and industry.

Sixty-five participants from 29 public and private universities and three funding organizations in the US attended. All academic participants are leaders of data science institutes, centers, or initiatives on their respective campuses and/or leaders of projects funded by NSF, the Alfred P. Sloan Foundation, and the Gordon and Betty Moore Foundation. The goal was to share insights and practices among some of the pioneering initiatives in data science. Admittedly, however, this inaugural summit overrepresented highly-ranked universities and overrepresented the computer and information science community. Thus, the group collectively voiced the need for

future meetings to be inclusive of the broader data science landscape, in terms of both breadth of universities and colleges and breadth of academic disciplines. Appendix A lists all participants and Appendix B contains the agenda.

To set context, Section 1 of this report provides a description of the data life cycle, explains the multidisciplinary nature of data science, and provides a targeted roadmap to Section 2 for specific audiences. Section 2 contains a set of key observations and recommendations, summarized below.

Recommendation #1: The academic data science community should continue to hold regular, e.g., annual, Data Science Leadership Summits, building on the momentum started by the March 2018 Summit. Subsequent meetings should be more inclusive of all colleges and universities with data science initiatives, to represent the diversity of higher education institutions in the US. Recognizing the multi- and trans-disciplinary nature of data science, it should also be more inclusive of the diversity of disciplines that underlie the methods of data science and that benefit from its application. Additionally, consideration should be given on whether to include representation from industry at future meetings.

Recommendation #2: The academic data science community should pursue other efforts that would be beneficial to building the community: hold a regular data science research/education workshop or conference; establish a transdisciplinary journal; support activities such as a shared communication channel (e.g., mailing list), request for information, faculty job announcements, etc.; and support data sharing across universities. In order to avoid unnecessary proliferation of efforts, workshops, conferences, and journals should partner or be coordinated with ongoing efforts by professional societies, e.g., Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE), American Statistical Association (ASA), and Society for Industrial and Applied Mathematics (SIAM). Moreover, any new efforts should distinguish themselves from existing ones.

Recommendation #3: The academic community requires coordinated sharing and management of publicly available datasets. Funding agencies, in collaboration with the community, should incentivize responsible data sharing and access.

Recommendation #4: Summit participants should provide a taxonomy for the community and university administrators that identify the design dimensions for supporting one or more data science entities on campus.

Recommendation #5: The academic data science community, working with an agency or professional organization, should create a survey instrument to track numbers (e.g., enrollment, funding, degrees awarded, etc.) for data science. The agency or professional organization should administer the survey periodically, e.g., annually.

Recommendation #6: Given the increase in the number of professional master's degree programs in data science, and industry demand for their graduates, the academic data science community, working with the National Academies of Science, Engineering, and Medicine, industry and professional societies, should come up with a set of minimal standard requirements for a professional master's degree in data science.

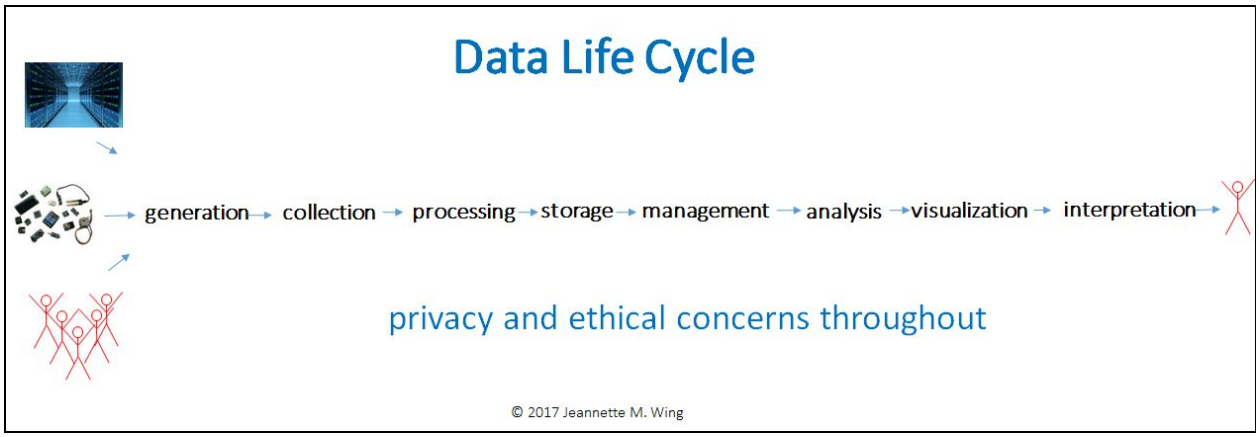
Recommendation #7: The data science community, working across academia, government, and industry, should define a code of ethics for data science. For enforcing this code, these stakeholders should also define Institutional Review Board (IRB) criteria and processes specific for data. This "IRB for Data" should include guidelines for the use of industry data by academics. These definitional efforts should leverage existing community efforts, including studies on data science by the National Academies of Sciences, Engineering, and Medicine.

Recommendation #8: The academic data science community should integrate ethics training in its research and education programs. Such training should recognize new ethical issues that arise with the collection and use of data about people and their behavior, and their implications on society.

Recommendation #9: Academia and industry should have a dialogue to explore new ways to bring data scientists to the data held by industry and to allow academics to test their models and analyses on industry data.

1. Introduction

1.1. The Data Life Cycle



To put data science in context, it helps to consider the entire *data life cycle* (see figure above).[1]

[1] The picture and prose are extracted from Wing's blog post The Data Life Cycle (January 2018).

The data life cycle starts with the *generation* of data, e.g., from people, organizations, populations, sensors, devices, and scientific instruments. After generation comes *collection*. Not all data generated is collected, perhaps out of choice because we do not need or want to, or for practical reasons because the data streams in faster than it is possible to process. After collection comes *processing*. Here we mean everything from data cleaning, data wrangling, and data formatting to data compression, for efficient storage, and data encryption, for secure storage. After processing comes *storage*. Here the bits are laid down in memory. After storage comes *management*. We are careful to store our data in ways both to optimize expected access patterns and to provide as much generality as possible.

Now comes *analysis*. When most people think of data science, they mean data analysis, i.e., computational and statistical techniques for analyzing data for some purpose. The analysis techniques include algorithms and methods that underlie data mining, machine learning, modeling, and statistical inference, be they to gain knowledge or insights, build classifiers and predictors, or infer causality.

Beyond analysis, data *visualization* helps present results in a clear and simple way a human can readily understand. Often it is not enough just to show a pie chart or bar graph. *Interpretation* provides the human reader an explanation of what the picture means. We tell a story explaining the picture's context, point, implications, and possible ramifications.

Finally, we have the end user. The user could be a scientist, who through data, makes a new discovery. The user could be a policymaker who needs to make a decision about a local community's future. The user could be in medicine, treating a patient; in finance, investing client money; in law, regulating processes and organizations; or in business, making processes more efficient and reliable to serve customers better.

This diagram omits arrows that show the many feedback loops in the data life cycle. Inevitably, after we present some observations to the user based on data we generated, the user asks new questions and these questions require collecting more data or doing more analysis.

Underlining this diagram is the importance of using data responsibly—at each phase in the cycle. We must remember to consider ethical and privacy concerns throughout, from privacy-preserving collection of data about individuals to ethical decisions that humans or machines will need to make based on automated data analysis.

Data science is the study of extracting value from data. "Value" is subject to the interpretation by the end user and "extracting" represents the work done in all phases of the data life cycle.

1.2. Multidisciplinary Nature of Data Science

Data science is inherently multidisciplinary in two ways: depth and breadth.

Depth. First, the technical foundations of data science draw on computer science and statistics, but are also informed by other areas of study, such as biostatistics, digital signal processing, mathematics, and optimization.

Summit participants debated whether data science is a new field, emerging from the convergence of existing fields, or the evolution of an existing field. Those who see it as a new, emerging field see data science as drawing on methods from many existing fields, e.g., computer science, mathematics, operations research, and statistics. Others see data science as simply an evolution of statistics, e.g., anticipated as early as 1962 by John Tukey [Tukey 1962, Donoho 2017], or an evolution of computer science, e.g., as probabilistic and statistical reasoning becomes as important as symbolic and logical reasoning in computing. Regardless of whether data science is "new" or not, there was consensus that concepts and techniques from (at least) computer science and statistics are core to data science.

Breadth. Second, data science is used in context, e.g., to explore a data set, to create models, and/or to test hypotheses—in a given domain. Because all domains generate or collect data, all domains have the potential to benefit from the analytical techniques in data science. Thus, one can say data science methods can be applied to all fields, professions, and sectors.

In their *PNAS* 2017 article, "Science and Data Science," Blei and Smyth emphasize the importance of the domain in data science, where "data scientists and domain experts" collaborate:

> Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention. It emphasizes the value and necessity of approximation and simplification; it values effective communication of the results of a data analysis and of the understanding about the world and data that we glean from it; it prioritizes an understanding of the optimization algorithms and transparently managing the inevitable tradeoff between accuracy and speed; it promotes domain-specific analyses, where data scientists and domain experts work together to balance appropriate assumptions with computationally efficient methods [Blei and Smyth 2017].

In his 2007 presentation to the National Academies' Computer Science and Telecommunications Board, Jim Gray anticipated data science by arguing the centrality of data ("the fourth paradigm") for driving new discovery in science, e.g., astronomy and biology [Hey, Stewart and Tolle 2009]. Ten years later, the community now recognizes that data science is applicable not just to science disciplines, but to *all* disciplines.

Any field that emerges from existing fields will face certain challenges in terms of crossing boundaries: communication and culture, faculty hiring and promotion, faculty service, joint degree programs, and so on. Universities have handled such emerging fields in the past, helping to break down disciplinary boundaries. Computer science is one example. Computational biology is a more recent example.

It is the second multidisciplinary aspect of data science, however, that presents an unusual challenge for most universities. How does one embrace a field that has the potential to transform every other field on campus? In Section 2.2 we explore how different universities are addressing this question in different ways.

The excitement on many campuses today in starting a data science initiative comes not just from the faculty who are pushing the frontiers of this emerging field, not just from the students who know they need to know data science for their future careers, and not just from industry whose demand for data scientists outweighs the supply. The excitement also comes from administrators who recognize that tackling societal challenges inherently requires multiple disciplines to come together and work collaboratively. Many forward-looking administrators today worry about the long-term future of higher education and want to connect academia more to the general public and policymakers. They recognize the need to break down silos and to support multidisciplinary collaborative research and education on campus. Many academic leaders also realize that teams are required to solve complex societal problems, and the academy still resists this emphasis on teams, through, for example, tenure criteria that emphasize an individual's contributions. Data science is a natural vehicle to help with breaking down barriers.

1.3. **Purpose of Summit**

Data science is a burgeoning field. As a result of recent technological advances, widespread and accelerated uptake of these technologies by many sectors, and increasing workforce demands, many data science initiatives across universities and colleges in the US and beyond are sprouting up at a rapid pace. Two years ago, there were only a handful of data science institutes, centers, or initiatives; now there are over 20 in the top public and private universities. The goal of this workshop was to convene the leaders of these campus efforts:

- To initiate the formation of an academic community for data science;
- To share best practices amongst academic leaders facing similar challenges and opportunities; and
- To take collective responsibility in preparing next-generation data scientists to contribute in the best interests of society.

Prior to this Summit, many other related workshops on data science had been held and many related reports had been written. The purpose of this workshop was explicitly not to rehash what had been said or written about before. For example, before the Summit, participants were encouraged to read "Creating Institutional Change in Data Science," a report written by the Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington [Moore-Sloan 2017].

On the other hand, as can be seen from the agenda, to level set the audience, participants heard brief presentations about many of these prior activities. Many of these activities were funded by the NSF, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation, also co-sponsors of this Summit.

To avoid repetition, but also to provide one place for a collection of prior work, Appendix E points to the many resources that serve as background material to the Summit, and thus to this report.

2. Key Observations and Recommendations

We summarize key observations and recommendations along seven topics discussed through the course of the day's events, which we elaborate in subsections below:

Guide to report: We wrote this report with different audiences in mind. Although all our observations and recommendations should be of interest to all, certain subsections may be of special interest to specific audiences:

- Academic community: Sections 2.1, 2.5, 2.6, 2.7
- Academic administrators: Sections 2.2, 2.3, 2.4, 2.5, 2.8, Appendix C, Appendix E
- Industry: Sections 2.6, 2.7
- Government and other funding agencies: Sections 2.1, 2.4, 2.5, 2.7, 2.8
- Professional societies: 2.1, 2.4, 2.5, 2.6, 2.8

2.1. Enthusiasm for Community Building

The primary focus of the Summit was to bring together academic leaders in data science and to gauge interest in building a community. One source of inspiration for this Summit was the

Computing Research Association, founded in 1972, which has built a community for leaders in computer science. Every two years, CRA sponsors an event held in Snowbird, Utah for all chairs of computer science departments in the US and Canada and all directors of industrial research labs in the information technology sector. This Summit was meant to be the first "Snowbird" for data science. Future incarnations of this Summit should be sure to include as equal partners other foundational disciplines of data science, e.g., statistics.

Even before the meeting, invitees expressed strong enthusiasm for the Data Science Leadership Summit. Many others involved in data science on their respective campuses or at the National Science Foundation heard about the Summit and asked to attend. Due to budget and space constraints, we had to limit the number of attendees to ~60, and thus maintained a waitlist that kept growing to the day of the Summit. At the Summit, from the start it was clear that there was a lot of enthusiasm from this community to discuss shared challenges and to provide opportunities for sharing experiences and lessons learned.

The Summit participants included: leaders of centers, institutes, programs, or initiatives in data science at 29 different universities; leaders of national studies, reports, workshops on data science; funders of projects, programs, and academic efforts in data science; and one person running a non-profit organization on data ethics. Specific to funded projects, we had representatives from: all four NSF Big Data Regional Innovation Hubs (Big Data Hubs); eight out of twelve NSF Transdisciplinary Research in Principles of Data Science (TRIPODS) projects; the Translational Data Science workshops; the Data Science Corps workshop; and all three Moore-Sloan Data Science Environments. Many of the leaders attending wore two or more hats, e.g., as a leader of a data science institute and as a leader in a report, or as a leader of an NSF-awarded data science project and as a leader of a workshop.

Each of these stakeholders shared their perspectives in establishing and supporting data science programs and the need for establishing a community. There was unanimous agreement to continue holding periodic, e.g., annual, Data Science Leadership Summits.

Summit attendees from academia (see Appendix A) came from both private and public universities, and from diverse geographic regions within the US. Despite the diversity in the types of universities and of data science programs represented at the meeting, there was an overrepresentation of the highest ranked universities in the country and an overrepresentation of the computer science community. Although this choice allowed participants to collect best practices and lessons learned from the universities and institutes who have been pioneers in data science, we recognize that the academic landscape is much broader. Future summits should include a diversity of participating institutions, and of disciplines given the multi- and trans-disciplinary nature of data science.

The Summit participants also discussed the importance and opportunity to build a community that is diverse across traditionally underrepresented populations in STEM. Data science should embrace diversity and inclusion early on in its formation as a discipline to try with intention to

avoid the lack of diversity found in so many technical fields today. Encouraging and facilitating broader participation should extend to faculty hiring, conferences and journals, with emphasis on mentoring data scientists from diverse demographic backgrounds.

Recommendation #1: The academic data science community should continue to hold regular, e.g., annual, Data Science Leadership Summits, building on the momentum started by the March 2018 Summit. Subsequent meetings should be more inclusive of all colleges and universities with data science initiatives, to represent the diversity of higher education institutions in the US. Recognizing the multi- and trans-disciplinary nature of data science, it should also be more inclusive of the diversity of disciplines that underlie the methods of data science and that benefit from its application. Additionally, consideration should be given on whether to include representation from industry at future meetings.

Decision #1: The four NSF Big Data Regional Innovation Hubs agreed to take the responsibility to organize the next two to three Data Science Leadership Summits on an annual basis. This decision makes organizational sense in that, for the US, these hubs have the reach into the broader academic community and they have the infrastructure to organize such events. Both the Moore and Sloan Foundations expressed interest in supporting subsequent events.

Recommendation #2: The academic data science community should pursue other efforts that would be beneficial to building the community: hold a regular data science research/education workshop or conference; establish a transdisciplinary journal; support activities such as a shared communication channel (e.g., mailing list), request for information, faculty job announcements, etc.; and support data sharing across universities. In order to avoid unnecessary proliferation of efforts, workshops, conferences, and journals should partner or be coordinated with ongoing efforts by professional societies, e.g., Association for Computing Machinery (ACM), American Statistical Association (ASA), Institute of Electrical and Electronics Engineers (IEEE), and Society for Industrial and Applied Mathematics (SIAM). Moreover, any new efforts should distinguish themselves from existing ones.

- An annual data science research conference would meet the demands of a growing data science research community. This conference should welcome both methodology and applications research. Professional societies and funding agencies could support such conferences. Before any new conference is started, the community should determine whether one or more existing related conferences already serves this purpose and thus coordinate with their organizing bodies. Similar remarks hold for conferences on data science education.

- Given that many existing organizations, e.g., ACM, ASA, IEEE, and SIAM, are spawning data science journals, we should collaborate with these efforts to ensure quality and integrity of the field, and to identify gaps that would warrant any new journal.

- Support activities, such as those listed above, should be organized initially by the NSF Big Data Regional Innovation Hubs.

- Support for data sharing across academic institutions and indeed across disciplines would advance both research and education on campuses. University libraries could play an important role to support data sharing within, and even across universities. The NSF Big Data Regional Innovation Hubs could play an important role in organizing data sharing across universities, colleges, and local communities.

Decision #2: There was consensus that it was too early to decide whether to piggy-back on existing organizations, e.g., ACM, ASA, IEEE, SIAM (or some combination thereof) to host these community efforts, or to evolve to grow our own, e.g., as for Neural Information Processing Systems (NIPS). We felt that after two to three subsequent Data Science Leadership Summits and after two or three more years, as the field evolves, we will have a better sense of what would be best for the community and field.

By definition, data is fundamental to data science. For the academic community to be at the cutting edge of data science, data sharing could expedite scientific progress. Academics can benefit not just from sharing datasets with each other, but also from coordinating the sharing of and managing the availability of public datasets.

Recommendation #3: The academic community requires coordinated sharing and management of publicly available datasets. Funding agencies, in collaboration with the community, should incentivize responsible data sharing and access.

Data sharing with industry is as important, but has its own unique challenges (Section 2.7, Recommendation #9).

To address the many challenges inherent to community building, funding agencies represented at the Summit shared their enthusiasm in looking to support future reports, workshops, and conferences for data science leaders, researchers, and educators. Themes for future conversations should include: data code of ethics; data use principles, including findability, accessibility, interoperability, reusability, and reproducibility; data science curriculum in higher education; new models of academia-industry engagement and collaboration; infrastructure to support data sharing across institutions; and balancing data science methodology and applications.

2.2. One Model Does Not Fit All

The opening question of the Summit “How does data science fit into a university?” led to a lively and engaging discussion. The answers reflect the ongoing broader community discussions about whether data science is a new field or not, what fields (e.g., in addition to computer science and statistics) feed its foundations, what fields can benefit from the application of data

science, and perhaps most importantly, how the field will evolve—what will data science look like in 10-20 years?

This opening question led to many participants responding, “Here’s what we do at my university.” One person said data science is within the university’s College of Information and Computer Sciences; one person said data science is part of the newly renamed Statistics and Data Science department; and another said data science defines a new Division of Data Sciences. Some reported that data science is a free-standing entity or a joint effort drawing from multiple departments or schools on campus. Some universities have multiple entities (institutes, departments, colleges, etc.) that support different aspects of data science, e.g., research or education. Depending on the university, these entities report to a dean, multiple deans, the provost, or even the president. How an entity got started and how it will be sustained also vary across universities and influence its structure.

There were four main types of models discussed: (a) creating a brand new academic unit, e.g., a School of Data Science; (b) repurposing an existing entity, e.g,. adding data science to the statistics department; (c) creating a new entity (institute/center/initiative) that is not tied to any one or more academic unit; and (d) creating a new entity that is joint with multiple academic units across campus. Some argued that even if data science draws on computer science, statistics, and other fields as its foundations, there is still value to having a separate entity, e.g., an institute, that draws on these foundational disciplines, but transcends disciplinary boundaries. Some expressed concern that creating or repurposing a new academic disciplinary unit could continue to reinforce disciplinary silos.

Regardless of model, the immediate questions a university faces include:

- Which faculty are part of the new or extended entity? How is membership decided?
- What is the governance structure of the data science entity? To whom does the entity report?
- How are new faculty lines in data science allotted and distributed (e.g., is there a split with other departments)? Who pays for these lines along with startup costs?
- What is the role of the data science entity in—education, research and, service to the university—and what should the balance be? Here, service means both people and computational infrastructure.
- Research/technical staff can play a critical role to serve disciplines across campus, bringing their data science expertise to domain experts. Is there funding to support such staff? Are there career development plans and pathways for such staff? How can universities attract such staff, given that they are in high demand by industry as well?
- Can the data science entity hire its own tenure-track faculty? Research faculty and technical staff? Can it run its own academic programs?
- If faculty have an academic home in one department and membership in the data science entity, how does one incentivize faculty to contribute, e.g., teaching and service, to the data science entity, and reward them for their contributions? How is faculty

recruiting, hiring, mentoring, promotion, and evaluation done especially if faculty are from across disciplines?
- What is the financial model for supporting the data science entity? Does the data science entity get any tuition or indirect cost recovery (on grants)?
- What is the contribution that existing schools make toward supporting the data science entity?
- What is the long term sustainability plan for the data science entity?

From the variety of ways in which participants described their structures, participants realized early on that no one model fits all universities. Each university has its own traditions, culture, funding models, and politics. Participants felt there was no need (at least in the course of one day) to come to consensus on what is best for all universities, especially since the field is still evolving. Rather, Summit participants can best help the community, and in particular, university administrators, by providing a list of questions, such as those above, that each university would have to face and that many of us have had to face or are facing, along with responses where possible, to how such issues are being addressed currently. Given that so many universities are just now planning some kind of data science effort, this Summit's contribution, as documented in this report, would be useful, practical and timely.

Recommendation #4: Summit participants should provide a taxonomy for the community and university administrators that identify the design dimensions for supporting one or more data science entities on campus.

2.3. Taxonomy and Informal Survey

Given the enthusiastic momentum felt at the Summit and a sense of urgency to understand the landscape of the taxonomy of data science models, the organizers adjusted the agenda to have one of the breakout sessions focus on designing a *taxonomy* of data science academic models.[2] Summit participants agreed to create a survey and in the weeks following the Summit administer it on themselves in order to populate the design space suggested by the taxonomy. Appendix C contains details about how we administered the informal survey, the survey questions, a taxonomy of models, and the survey results.

Universities and colleges interested in pursuing data science initiatives can use the taxonomy and survey results for an understanding of the models in use at other institutions. Each university or college should weigh the pros and cons of each model based on its own local structure, culture, needs, and goals.

During the breakout session for a taxonomy of data science models in higher education, participants identified the following high-level taxonomy dimensions for data science entities or

[2] One consequence of this adjustment was that the two breakouts on master's and doctoral education were combined to a single breakout on graduate education.

units: structure, mission, ownership/leadership, resources, educational programs, faculty engagement strategies, university engagement, computing/staff support, stakeholders, and metrics of success.

Subsequent to the Summit, a handful of Summit participants built a survey to capture these high-level taxonomy dimensions. We built the survey, in part, inspired by Katz's survey, sponsored by the Moore and Sloan Foundations [Katz 2018], which covers 20 schools with data science institutes. Katz's survey was conducted with one-on-one interviews with the 20 participants over the course of two years, and thus gives a more detailed presentation and analysis than we were able to do through our more informal survey conducted over five weeks.

Thus, some caveats: Our survey is not meant to be comprehensive. Rather, it represents an initial step to capture some of the relevant features and dimensions of data science entities of the universities represented at the Summit. Moreover, because the survey requested free-response answers to a majority of questions, the resulting data are qualitative. We translated the narrative text from the survey to a summary table (Appendix C.3) and sent this table to the survey respondents to confirm their responses. We caution the reader not to over-generalize from the tabular results, given the small number and limited types of universities surveyed.

Insights from Informal Survey

The informal survey results substantiate all remarks and recommendations of this report. Hence, we do repeat those results. Please see Appendix C.4 for our summary of the survey results. Below are some additional insights gained from the survey.

The majority of the respondents emphasized the need for institutional support and the importance of having the right type of people to advance the broad mission of the entity. There were several responses that addressed the "drivers" of an entity, and most, but not all preferred that it be driven by faculty.

The respondents identified several types of people who are important to the success of a data science entity, including engaged senior faculty who provide the executive support, junior faculty with the drive to support a vibrant agenda for research and education, highly skilled staff data scientists, and strong administrative support staff.

Echoing the Summit discussions, there was a general sense from the survey that data science is a vibrant area of growth. One entity highlighted commitment to a selective program that is interdisciplinary in nature and produces highly sought-after graduates. Another respondent noted that there are several demands for growth in teaching and research opportunities; however, strong staff and personnel support is needed to facilitate these opportunities and grow the entity. Yet another entity cautioned against the perils of becoming primarily a service

provider, because almost all students are looking for exposure to data science, and the demand for hands-on data science training and support is high on campuses.

2.4. Need for Quantitative Data

As the field of data science is still in its formative years, there is clearly a need for comprehensive, quantitative data about the field such as: the number and types of degrees; numbers of degrees awarded at the undergraduate, master's, and doctoral levels; types and numbers of jobs and career opportunities for graduates; number of faculty in data science or with joint appointments, etc. Tracking these numbers now will allow us to monitor trends to prepare for the future. These data can help evaluate the current state of data science and opportunities in workforce development.

The participants agreed that there is a need for a periodic survey such as the Taulbee Survey for Computer Science (https://cra.org/resources/taulbee-survey/). It would be important to set this survey up now, before we lose track of early numbers and thus, initial trends. Funding agencies could help support and sustain this effort.

One of the key challenges in collecting this quantitative data will be defining what lies inside the purview of data science. The broadest definitions of data science will subsume entire disciplines, whereas narrow definitions will exclude research that should be captured. A community effort can help address these challenges.

Recommendation #5: The academic data science community, working with an agency or professional organization, should create a survey instrument to track numbers (e.g., enrollment, funding, degrees awarded, etc.) for data science. The agency or professional organization should administer the survey periodically, e.g., annually.

The results of our informal survey reinforced this recommendation. They also highlighted the need to design any future survey to take into careful consideration that data science is multidisciplinary and data science entities engage with multiple other entities on campuses.

We share below some additional insights gained from our survey that can inform future survey instruments:

- Structured questionnaires scale but lack needed granularity. Our survey was a structured questionnaire, which scales (unlike an interview-based survey such as Katz's [Katz 2018]), but does not easily support high degrees and levels of branching or unanticipated exploration of responses. Since data science entities span a multitude of dimensions in their design, capturing nuances and granularities, can be difficult. Future surveys will need to decide on the type of survey instrument and anticipate the need to fine tune the survey over time. This fine-tuning by adding granularity is similar to how the Taulbee survey has evolved for computer science, e.g., early on, in distinguishing

between computer engineering and computer science, to more recently, in distinguishing between public and private universities, by department size and geographic location.

- Seemingly simple questions can be hard to answer. For example, asking a computer science department "How many faculty are members in your department?" is much easier to answer than "How many faculty are members in your data science entity?" From our survey, we discerned two distinct subtleties if one were to try to count faculty: First, what does it mean to be a "member" of a data science entity? Moreover, most entities have different levels of membership, e.g., regular and affiliate. Second, does the entity have "ownership" of the faculty member—unilaterally, jointly, or not at all? One could be a faculty member of a data science institute but the institute might have no control over that faculty member's hiring, promotion, space, responsibilities, etc.

- Impact of a data science entity on campus relies on having sufficient resources and space. The spectrum between ownership and access makes it challenging to assess accurately the resources available to a data science entity: space, administrative staff, technical staff, computational infrastructure, etc.. Resources might be solely owned, shared, or simply accessible for use. Future surveys might want to ask these kinds of specific, distinct questions: Do you have dedicated space [resources]? Do you own the space [resources]? Do you share the space [resources]?

- While education and research are key functions for data science entities, so is service. Future surveys might want to ask more explicitly about each of the service activities provided by an entity: bootcamps, hack-a-thons, training programs, providing (but not teaching) course modules, cloud computing support, etc.

- Understanding the intellectual drivers of data science is important. The phrase "intellectual driver" used in our survey, however, was interpreted in multiple ways. Future surveys should be clearer about separable issues: drivers of data science as a field versus drivers (e.g., active participating departments/schools) of a data science entity.

- Multiple data science efforts exist on most campuses, where collaborations are sometimes informal and not measured. Future surveys should explore ways to measure collaboration between efforts. This type of information should be gathered from a respondent with a broad view data science activities on a given campus (e.g., vice provost/president of research, dean).

2.5. Education

Because data science is a multidisciplinary field in two ways (Section 1.2), how to train next-generation data scientists at all levels is a challenge. As for any multidisciplinary field, one cannot expect a student to take the union of the course requirements for each discipline, so

something has to go; moreover, some existing courses or programs may need to be tailored for the new field and the background of the students.

Data science raises another interesting challenge: how to bring domain knowledge into a data science educational program, and how much. As stated in Section 1.2, because domain context is important for data scientists, how does one bring in other domains in teaching data science?

Since two National Academies efforts, a report on undergraduate education and a roundtable on graduate education, already study these and other questions in more detail, we provide here an outline of the issues for each educational level: undergraduate, master's, and doctoral. These observations should be viewed as our current understanding of data science education today. As the field evolves, so will the content and nature of these programs.

Undergraduate Education*:* Summit participants voiced the demand on campus and by industry for data science majors and more generally a level of data science knowledge that all undergraduates should have regardless of major. Universities are exploring various models, often in combination:

- Create a data science course suitable for freshman and/or non-majors that would provide all undergraduates a minimum skill set in data science. Berkeley's data8 and Columbia's Data: Past, Present, and Future are two examples. Determining a ``minimum skill set'' may depend on the student body; professional organizations that set standards could play in important role. Many agreed that students should, at a minimum, be exposed to data ethics (see Section 2.6), and ideally taught programming and probability and statistics.

- Create specific upper-level data science courses for students without the backgrounds, e.g., Machine Learning for non-CS majors. Having sufficient teaching resources to cover additional courses may be a challenge for some departments.

- Create a data science major through some combination of existing computer science, statistics, and domain-specific electives. Issues for the participating departments would be what courses to include, how not to overload students, and how to administer the major.

- Create, as an alternative to creating a new independent major, a specialization in an existing computer science or statistics major.

- Create a data science track, added to an existing domain-specific major, e.g., biology or economics.

- Create a data science minor.

- Create data science courses for domain experts. Such courses can be done jointly between the entity overseeing data science education and the department for a given domain.

- Create joint or dual majors between data science and some other discipline X. Just as there are such degree programs between computer science and X, we are likely to see growing interest in data science and other disciplines.

The National Academies report "Envisioning the Data Science Discipline: The Undergraduate Perspective" provides a more detailed discussion and set of recommendations on undergraduate education in data science [National Academies 2017]. Similarly the report from the 2016 Park City Mathematics Institute (PCMI) and the Institute for Advanced Study at Princeton [Statistics 2017] provide curriculum guidelines. Given the existence of these reports, Summit participants did not feel the need to make additional recommendations.

Participants felt that universities should explore and experiment with different offerings and models; at subsequent Data Science Leadership Summit meetings, participants can share best practices. Moreover, the high demand for data science raises an immediate challenge on campuses of how to scale up efforts, when most courses and programs, as described above, are themselves just getting started.

Graduate Education*:* There are at least four kinds of graduate programs to consider: a professional master's certificate non-degree program, a professional (terminal) master's program, a research master's program (as a possible step toward a PhD), and a doctoral program.

A professional master's certificate non-degree program would primarily serve a community of practitioners, e.g., professionals working in local industry, who want to learn the basics in data science, e.g., machine learning and statistics. Courses offered through such a program also serve academics in other disciplines. The program could be designed so that the courses would overlap with the professional master's program requirements so a student could choose to go on for a degree.

For the professional master's program, there was no consensus on what every terminal master's student should know. Yet, according to the Institute for Advanced Analytics at the North Carolina State University, there are already over 200 master's programs. Several of these are offered at professional studies schools. With such a vast variety and number of degree programs, now is the time for the academic community to come to consensus on standards. The National Academies report on undergraduate education could serve as a starting point for a set of minimal expectations as well as a model for coming to consensus by the community.

For a research master's program, often the design of a doctoral program implicitly defines the course requirements for a research master's program. The challenge is that there are only a few (less than one handful) doctoral programs in data science, despite the growing number of master's students who want to go on for a doctoral program. So, for example, if they are in a professional master's program in data science, they will need to be given research opportunities. One way to create such opportunities would be through summer internships; another would be through project/capstone/experiential courses.

Akin to the discussion on undergraduate majors, there are many possible models to consider in designing a doctoral program. Which model is best will depend on institutional structures. Should there be an independent Ph.D. in data science, a joint Ph.D. program (e.g., built from computer science and statistics), specializations or tracks built from existing Ph.D. programs (e.g., awarding a certificate in data science), or some combination thereof? Another design consideration is how to support the multidisciplinary nature of data science: Should Ph.D. students in data science have co-advisors (one in data science and one in a domain)? Should there be two tracks in the Ph.D. program, one on advancing methods of data science and one on advancing the domain through data science?

One current challenge for a research master's or Ph.D. program is whether potential employers will prefer a data scientist or a domain expert with data science skills. Data science may be different enough from other fields that both kinds of graduates would be in high demand. This challenge highlights the need for collecting quantitative data as discussed in Section 2.4 (Recommendation #5).

Commonalities Across Programs: For all three levels of education, there was consensus that ethics training for our students should be required. We devote Section 2.6 to a discussion of ethics in data science since it cuts across both research and education.

One common theme across educational programs was teamwork. Most master's students will take their data science skills and work in a company with sector-specific data and on teams with different technical and non-technical skills. Research master's and doctoral students, by the inherent multidisciplinary nature of data science, can benefit from working with domain experts. Opportunities for teamwork are practicums and internships.

Another common theme across programs is the need to teach reproducibility in data science. Participants raised the issue of reproducibility of findings from data and reproducibility of algorithms, through development of tools. Some argued that reproducibility should be embedded in the core curriculum elements. This argument was motivated, not just from a educational perspective, but also from an ethical perspective of reusing data and reproducing the research with the right set of constraints and careful considerations.

Related to ethics, teamwork, and reproducibility is providing opportunities for students to work on real-world data. Possible ways to provide this opportunity include:

- Capstone courses: Here, an industry partner brings to a team of students a real-world data set and the students answer real-world questions about it. The challenge is to ensure sufficient mentoring by the industry partner and sufficient oversight by a faculty member, who needs to provide technical expertise. Industry expectations need to be set appropriately; they should not expect a product.

- Campus data: Faculty can offer projects (for all levels of students, including undergraduates) that involve interesting data sets for students to analyze. Lightweight administration is needed to match students with the appropriate skill set to projects. The recently announced NSF funding for the Open Storage Network project can provide a strong foundation for sharing of data and facilitating local campus efforts.

- Community data: Students could self-organize into not-for-credit groups and do a data science project as a voluntary community service activity.

Finally, data science attracts unconventional applicants. Many applicants may not have a solid computer science background or sufficient mathematics to jump right into a course on machine learning. Some may have an advanced degree (even a Ph.D.) in a different field and want to learn data science. Also, given the newness of the field, the academic community has the opportunity to attract applicants from populations, e.g., female, underrepresented minorities, and underserved groups, that have traditionally been a challenge for STEM disciplines.

Course and program prerequisites should consider the diversity of students who come from many different backgrounds, including non-technical ones. Graduate programs may also need to offer students (e.g., before enrolling) boot camps, self-study or remedial courses to make sure entering students satisfy the appropriate prerequisites. The diversity of applicants who are attracted to study data science is considered a plus, because it brings breadth of perspectives and expertise to the field.

Because there is already a National Academies roundtable on graduate education, there were no specific recommendations on curricular issues for graduate programs. It would be good if the National Academies roundtable discussions or relevant professional societies led to a concrete recommendation on minimal requirements for professional master's programs, as so many have proliferated in the past few years.

Recommendation #6: Given the increase in the number of professional master's degree programs in data science, and industry demand for their graduates, the academic data science community, working with the National Academies of Science, Engineering, and Medicine, industry and professional societies, should come up with a set of minimal standard requirements for a professional master's degree in data science.

There is also an urgency for this recommendation, given how broadly industry interprets what data scientists do, from data cleaning to advancing deep learning.

Other Educational Outreach Programs: Many faculty, postdocs, and staff across campuses are eager to learn data science methods. Faculty could spend sabbaticals learning data science and then take their newly learned skills back to their discipline. Data science programs should be prepared to provide bootcamps and training workshops to colleagues on campus. These educational activities can also target local high schools and industry.

2.6. Ethics Training for Data Scientists

Daily headlines motivate the need to ensure that data scientists are given training in ethics. Participants felt ethics training is paramount for data scientists, especially those who will be working with data about people and with automated techniques that can have consequences on people's lives, e.g., self-driving cars, medical treatments, and society, e.g., recidivism and fair housing.

Teaching ethics should include teaching general principles, e.g, a code of ethics, but also cover case studies, e.g., ethical failures due to biased data, privacy failures due to joining ``anonymized" datasets. Given how much data about people are collected by industry, understanding how industry needs to balance ethical concerns with business value, would also be important to cover in teaching ethics.

Summit participants heard from Natalie Harris's Community Driven Principles for Ethical Data Sharing (Appendix D). The National Academies report on undergraduate education contains in Section 5 a "Data Science Oath" analogous to medicine's Hippocratic Oath. Both are good starting points for defining a code of ethics for data scientists.

Although there was consensus that teaching ethics to data scientists is imperative, there is still uncertainty on how best to do so. Should ethics be a stand-alone course, should it be woven into all the courses, and/or should it be done in the context of working with a real data set (e.g., in a project or capstone course)? Participants are trying different models and should learn from each other over time as to what works and what does not. Appendix D points to a growing list of courses on ethics and technology.

Ethics training represents an opportunity to involve the social sciences and humanities, which offer frameworks to help think about ethical questions, as well as schools (e.g., Business, Law, Journalism, Medicine) across campus that teach their students ethics as part of their professional training. For example, ethics courses could be co-taught between faculty in computer science/statistics and relevant faculty in these other fields; similarly, joint hires could be targeted to those who could teach ethics. Such a task can be done collaboratively across institutions and deployed in the same way through online courses and material.

Recommendation #7: The data science community, working across academia, government, and industry, should define a code of ethics for data science. For enforcing this code, these stakeholders should also define Institutional Review Board (IRB) criteria and processes specific for data. This "IRB for Data" should include guidelines for the use of industry data by academics. These definitional efforts should leverage existing community efforts, including studies on data science by the National Academies of Sciences, Engineering, and Medicine, and resources listed in Appendix E.

Recommendation #8: The academic data science community should integrate ethics training in its research and education programs. Such training should recognize new ethical issues that arise with the collection and use of data about people and their behavior, and their implications on society.

2.7. Working with Industry

Research: The most immediate concern in the context of working with industry is access to data. Many of the stupendous advances in data science, AI, and machine learning are due to large amounts of data used to train and test machine learning algorithms. The success of deep learning is due to the large amounts of data collected—by industry—enabling tasks such as speech recognition, image recognition, and even playing Go, to perform at or beyond human capability. In some areas of AI and machine learning, industry is ahead of academia because industry has the data.

The current climate makes sharing data by industry with academics especially challenging. The harvesting of Facebook data by Cambridge Analytica, and regulations such as the EU Global Data Protection Regulation (GDPR) may make industry wary of sharing data with academics.

However, it is in industry's best interests to ensure the academic research enterprise is healthy—to continue to provide new ideas and new talent for the long term future.

Industry additionally has the expertise of professional data scientists to process and analyze real-world data. Academia can benefit from learning from practicing data scientists. These professionals could work with students and faculty, exposing academics to industry-scale computational tools and providing real-world problems to work on.

The acceleration of advances by industry in data science, AI, and machine learning today feels different from advances in technology of the past. Industry has two advantages over academia: big data and big compute (including massive GPU clusters that academia cannot afford). Academics can already access big compute, i.e., cloud computing (including GPU clusters), though there are some financial and logistical obstacles [NSF Academic Cloud 2018]. Since data are a valuable commodity to industry and to preserve customer privacy, it is much more difficult for academics to gain access to data collected by industry. Industry has supported such access through visiting faculty programs and student internships. A more direct dialogue

between academia and industry is warranted, where new models of industry-academia partnerships need to be explored. One such example is the Social Data Initiative, which makes use of a trusted third party, to give academics access to Facebook data. New models might make more routine what are now viewed as exceptions, e.g., a faculty member holding simultaneous part-time positions with both a company and a university beyond the usual two-year leave of absence granted by most universities.

Recommendation #9: Academia and industry should have a dialogue to explore new ways to bring data scientists to the data held by industry and to allow academics to test their models and analyses on industry data.

Education: Just as the academic community is grappling with data science education at all levels (Section 2.5), industry is facing the challenge of what makes a "data scientist." The title in industry could range from someone who does data cleaning to someone who has a PhD in machine learning. To address some of these differences, many companies have both "applied data scientist" and "data scientist" titles. There might be others. Now would be a great opportunity for academia and industry to get together and come to some agreement on what skills a data scientist graduating from data science program should have and what skills are expected by someone with that title or similar titles in industry. One way to achieve common ground is through program standards (Recommendation #6). Without some set of standards, the academic community helps neither the field nor industry.

Ethics: Please see Section 2.6.

Compute Infrastructure: One presentation at the Summit summarized an NSF workshop [NSF Academic Cloud 2018] that promoted the idea of an *academic cloud,* which would be a vehicle to provide computational power not affordable by any one academic institution and to support data sharing across universities. The purpose of the NSF workshop, held in January 2018, was to activate academia, government, and industry (in particular, commercial cloud providers) to discuss the unique needs of academia and the current obstacles that make academia hesitate to move its research and education to the cloud. Subsequently, the workshop produced recommendations for a wider adoption of cloud by academic which included addressing issues such as no indirect cost on cloud computing resources.

One benefit of having shared data sets in the cloud is for scientific reproducibility. Experiments run in the cloud on a data set from one institution can be rerun easily from another institution. Another benefit of an academic cloud is to provide a mechanism for industry to give the academic community authorized access to industry data sets stored in the cloud; data analysis could be done on these data sets remotely with the appropriate protection.

While there was no explicit recommendation on this topic, the academic data science community should monitor the development of the academic cloud and work with industry and

government to support its realization. The community should leverage the academic cloud effort in sharing data, models, and algorithms across institutions and across disciplines.

2.8. Other Landscape Studies

The Summit included presentations from past data science workshops and funding programs. The briefings covered:

- Workshops/Conferences: Open Knowledge Network (July 2016, February 2017, October 2017); Translational Data Science (June 2017 and November 2017); Data Science Corps Conference (December 2017); Enabling Computer and Information Science and Engineering Research and Education in the Cloud (aka “Academic Cloud”) Workshop (January 2018)
- Reports/Roundtables: National Academies Study on Envisioning the Data Science Discipline: the Undergraduate Perspective; National Academies Roundtable on Data Science Post-Secondary Education; National Academies
- Funding programs: NSF TRIPODS; NSF: Big Data Regional Innovation Hubs and Spokes; Moore-Sloan Data Science Environments

The workshop and conference reports and academy studies are available for people to learn about topics like: what is data science, what is data science good for, what goes into an undergraduate data science program, how does data science get applied to problems in industry and other domains?

Appendix E contains a list of these and other resources (links and citations), including Katz’s survey of 20 data science institutes [Katz 2018], providing a snapshot of the state-of-the-field in data science. It is a good starting point to learn from the community what is being done, especially for university administrators who need to understand the importance of and excitement about data science.

Finally, this Data Science Leadership Summit was focused on only US academics. On August 20, 2018, the Alan Turing Institute and Imperial College of London, in conjunction with the twenty-fourth international conference on Knowledge Discovery and Data Mining, held a similar meeting called “The Future of Data Science and the Role of a Data Science Institute” for 30 directors of data science institutes worldwide, with representatives from Canada, China, the EU, Singapore, US, and UK. A summary report is in progress. The academic data science community in the US should consider in the future whether US efforts should be done jointly with international efforts.

References

[Blei and Smyth 2017] David Blei and Padhraic Smyth, "Science and Data Science," *Proceedings of the National Academies of Sciences*, vol. 114, no. 33, June 2017, pp. 8689-8692.

[Donoho 2017] David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

[Hey, Stewart and Tolle 2009] Hey, Tony, Stewart Tansley, and Kristin M. Tolle. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Redmond, WA: Microsoft research, 2009.

[Katz 2018] Luba Katz, "Landscape review of academic data science centers in the United States," Sloan and Moore Foundations, in preparation, 2018.

[Moore-Sloan 2017] "Creating Institutional Change in Data Science," Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington, 2017.

[National Academies 2017] Envisioning the Data Science Discipline: The Undergraduate Perspective, National Academies, Washington, D.C., December 6-7, 2017.

[NSF Academic Cloud 2018] Magdalena Balazinska, David Culler, Jennifer Rexford, and Jeannette M. Wing, "Enabling Computer and Information Science and Engineering Research and Education in the Cloud," NSF Workshop Report, January 8-9, 2018, Alexandria, VA, ACM Digital Library, June 2018.

[Statistics 2017] Richard D. De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, Ping Ye,Curriculum Guidelines for Undergraduate Programs in Data Science, Annual Review of Statistics and Its Application 2017 4:1, 15-30

[Tukey 1962]]Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no. 1, 1–67. doi:10.1214/aoms/1177704711. https://projecteuclid.org/euclid.aoms/1177704711

Appendices

Appendix A: Summit Participants

Name	Affiliation	Secondary Affiliation
Ahalt, Stan	University of North Carolina Chapel Hill	South Big Data Hub
Aluru, Srinivas	Georgia Institute of Technology	South Big Data Hub
Athey, Brian	University of Michigan	Midwest Big Data Hub
Balazinska, Magdalena	University of Washington	
Baru, Chaitan	National Science Foundation	
Bastón, René	Columbia University	Northeast Big Data Hub
Bonneau, Richard	New York University	
Cheikes, Brant	University of Massachusetts Amherst	
Dahleh, Munther	Massachusetts Institute of Technology	
Dey, Tamal	The Ohio State University	
Dhar, Vasant	New York University	
Dominici, Francesca	Harvard University	
Eglash, Steve	Stanford University	
Erickson, Lucy	AAAS S&T Policy Fellow at the National Science Foundation	
Florida, Robert	Columbia University	
Franklin, Michael	The University of Chicago	

Freire, Juliana	New York University	
Greenberg, Josh	Alfred P. Sloan Foundation	
Gropp, William	University of Illinois, Urbana-Champaign	Midwest Big Data Hub
Grossman, Robert	The University of Chicago	
Harris, Natalie	Harris Data Consulting	
Hero, Alfred	University of Michigan	
Hoffman, Michelle	Columbia University	
Indyk, Piotr	Massachusetts Institute of Technology	NSF TRIPODS
Iyengar, Garud	Columbia University	
Janeja, Vandana	University of Maryland, Baltimore County	AAAS S&T Policy Fellow at the National Science Foundation
Kannan, Nandini	National Science Foundation	NSF TRIPODS
Kautz, Henry	University of Rochester	
Kelly, Anthony	National Science Foundation	
Kloefkorn, Tyler	AAAS S&T Policy Fellow at the National Science Foundation	
Lazowska, Ed	University of Washington	West Big Data Hub
Lee, Meredith	University of California, Berkeley	West Big Data Hub
Machiraju, Raghu	The Ohio State University	
Mahoney, Michael	University of California, Berkeley	NSF TRIPODS
McCallum, Andrew	University of Massachusetts Amherst	

McKeown, Kathy	Columbia University	Northeast Big Data Hub
Mentzel, Chris	Gordon & Betty Moore Foundation	
Mongeau, David	The Ohio State University	
Muñoz-Avila, Héctor	Lehigh University	
Nicolae, Dan	The University of Chicago	
Norman, Michael	University of California, San Diego	West Big Data Hub
Odegard, Jan	Rice University	
Orabona, Francesco	Stony Brook University	NSF TRIPODS
Papakonstantinou, Yannis	University of California, San Diego	
Parker, Micaela	Moore-Sloan Data Science Environments	
Rajan, Hridesh	Iowa State University	Midwest Big Data Hub
Randall, Dana	Georgia Institute of Technology	NSF TRIPODS
Rappa, Michael	North Carolina State University	
Rodriguez, Abel	University of California, Santa Cruz	NSF TRIPODS
Saltz, Jeffrey	Syracuse University	
Scala, Ralph	Columbia University	
Scheinberg, Katya	Lehigh University	NSF TRIPODS
Shah, Devavrat	Massachusetts Institute of Technology	
Smyth, Padhraic	University of California, Irvine	
Spielman, Daniel	Yale University	

Sputz, Sharon	Columbia University	
Stark, Jonathan	Columbia University	
Stone, Sarah	University of Washington	West Big Data Hub
Szalay, Alex	Johns Hopkins University	
Vohra, Rakesh	University of Pennsylvania	
Wing, Jeannette	Columbia University	
Wolfe, Patrick	Purdue University	
Wright, John	Columbia University	NSF TRIPODS
Wright, Stephen	University of Wisconsin-Madison	NSF TRIPODS
Zhou, Harrison	Yale University	

Appendix B: Summit Agenda

Data Science Leadership Summit
Organizer: Jeannette M. Wing

The Interchurch Center
475 Riverside Drive, New York, NY 10115
Columbia University

Monday, March 26, 2018

AGENDA

All sessions, except for the breakouts, will be in The Lounge. Breakout groups will use Rooms C and D and the DSI third-floor conference room in the Interchurch Center.

8:15-8:45 Continental breakfast

8:45-9:00 Introduction to summit: motivation, purpose, goal, outcomes — Jeannette Wing

9:00-9:15 Data Science in Academia — Jeannette Wing

9:15-10:30 Foundations and Applications of Data Science (plenary) — moderator: Devavrat Shah

- 9:15-9:27 Foundations — Stephen Wright
- 9:27-9:40 Applications to other fields — Francesca Dominici
- 9:40-10:30 Discussion

Questions: How can we support inherently interdisciplinary research, if not in our foundations, for sure in our applications? How is faculty hiring (e.g., joint appointments) done? What kind of institutional support is needed? How are data science demands across campus met (e.g., through applied data scientists, post-doc fellows)?

10:30-11:00 Break

11:00-12:00 Responsible Use: Ethics, Privacy, etc. (plenary)

- 11:00-11:15 Code of Ethics for Data Science — Natalie Harris
- 11:15-12:00 Discussion

Questions: What are you doing in teaching students and encouraging research on this subject? What are your institutional challenges? What should we advocate as a data science community?

12:00-1:00 Lunch (partially working) — moderator: Jeannette Wing

12:30-1:00 Report on NSF Workshop on Academic Cloud — Magda Balazinska

1:00-2:30 Education: Content and Levels (combination of plenary and breakouts)

1:00-1:10	National Academies Study on Envisioning the Data Science Discipline: the Undergraduate Perspective	Alfred Hero
1:10-1:20	National Academies Roundtable on Data Science Post-Secondary Education	Kathy McKeown
1:20-1:30	Ph.D. Programs	Vasant Dhar
1:30-2:15	Breakouts: Three parallel sessions	

Questions:
Undergraduate: What should every undergraduate know? What should every undergraduate data science major know?

Masters: What should every (terminal) masters student know to be prepared for industry, not just the technology industry?[3]

Doctoral: What makes sense for courses, dissertation topic, advisor(s)? How do we both advance the field of the data science and support its broad applicability at the PhD level?

2:15-2:45 Report back from breakouts (plenary)

2:45-3:00 Break

3:00-3:45 External Partnerships (plenary) moderator: Steve Eglash

3:00-3:15	Translational Data Science and Data Science Corps	Raghu Machiraju Vandana Janeja
3:15-3:45	Discussion	

Questions: What kinds of engagement do data science units have with industry, local government, foundations, other universities, and K-12? What is working and what does not work?

3:45-4:45 Building and Sustaining Community (plenary) moderator: Jeannette Wing

3:45-3:55	NSF Big Data Innovation Hubs	Stanley Ahalt
3:55-4:05	Moore/Sloan Data Science Environments	Micaela Parker
4:05-4:45	Discussion	

[3] The breakouts were revised to combine the Undergraduate and Masters session into one and have the discussion on taxonomy as the third breakout session.

Questions: Is there something as academic leads for data science in our respective schools that we could be doing together that could be more than we could do alone? More that could be done by existing organizations, e.g., the NSF DataHub initiative? How should we sustain our academic data science community? Are there action items to recommend to NSF, Sloan, and/or Moore?

4:45-5:00 Next Steps, Summary, Final Remarks — Jeannette Wing

5:00-5:30 Travel to reception

5:30 Reception: Northwest Corner Building
550 W. 120th Street, New York, NY 10027
DSI Conference Room, 14th floor

The Data Science Leadership Summit is sponsored by the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, and the National Science Foundation.

Appendix C: Survey and Taxonomy

In this appendix we provide a description of the survey administration process (C.1), survey questions (C.2), taxonomy dimensions with tabular results (C.3), and a qualitative summary of survey results (C.4).

C.1 Survey Administration

The survey was administered in Google Forms. The survey introduction and instructions were given as follows:

> *The objective of this survey is to enumerate the most common issues on which choices must be made in creating a data science entity on a university/college campus. For a complete list of the questions appearing in this survey, please see:*
>
> *https://drive.google.com/file/d/1ArkoeiD_H7MZa-7C5_vCeuxAXiamDwFZ/view?usp=sharing*
>
> *For this survey, entity means school, department, institute, center, or initiative that supports data science.*
>
> *We are collecting responses from university representatives who attended the March 2018 Data Science Leadership Summit held at Columbia University.*
>
> *If you are affiliated with multiple entities that support data science (e.g., a data science division and a data science institute), we kindly request that you submit a response to this form for each entity. Also, we realize that many entities have co-directors; we only need one response per entity, so you can decide among yourselves who should fill out the survey.*
>
> *Survey Contact: NSF AAAS Fellow - Tyler Kloefkorn, tkloefko@nsf.gov / tyler.kloefkorn@gmail.com*

The survey was available from June 20, 2018 to August 8, 2018. We collected 26 responses from 22 institutions of higher education.

We sought survey responses only from Summit participants in higher education. In the cases where a university has multiple primary data science entities, we encouraged each entity to fill out a response. The survey focuses on data science entities, e.g., institutes and centers, but does include a limited number of questions about other major data science efforts on campus.

We repeat our caveats from Section 2.3: Our survey is not meant to be comprehensive. Rather, it represents an initial step to capture some of the relevant features and dimensions of data

science entities of the universities represented at the Summit. Moreover, because the survey requested free-response answers to a majority of questions, resulting data are qualitative. We translated the narrative text from the survey to a summary table (Appendix C.3) and sent this table to the survey respondents to confirm their responses. We caution the reader not to over-generalize from the tabular results, given the small number and limited types of universities surveyed.

C.2 Survey Questions

Survey questions were asked in three formats: multiple-choice, checkboxes, and free-response. Multiple choice-type questions directed respondent to “Choose only one oval”. Checkbox-type questions directed respondents to “Check all that apply.” “Free-response” questions allowed respondents to add a limited amount of prose.

The survey questions, as they appeared to respondents, are listed below.

Section 1:

- Email address
- What is the name of your university?
- What is the name of the data science entity?
- Please provide a link to this entity's webpage.
- Please provide your contact information (name, title, and affiliation).

Section 2: Data Science Entity Structure

- When (year only) was the entity established?
- **Who leads the entity (e.g., Chair(s), Dean(s), and/or Director(s))? Please give title(s) only.**
- For the entity's leadership position(s), is there teaching and service release? Mark only one oval.
    - Yes
    - No
    - Maybe
    - Other:
- To whom does the data science entity leadership report (e.g., Dean(s), Provost, and/or President)? Please give title(s) only.
- What are the functions supported by the data science entity? Check all that apply.
    - Education
    - Outreach - government
    - Outreach - industry
    - Research
    - University service - computational infrastructure
    - University service - expertise
    - Other:

- Please identify the top priorities for the entity. Check all that apply.
    - Address societal challenges
    - Advance methods and foundations in data science
    - Advance one or more application areas
    - Connect to industry
    - Educate undergraduate students
    - Educate graduate students
    - Support interdisciplinary, collaborative research across the university
    - Other:
- What are the academic disciplines (e.g., computer science, statistics, domain sciences, etc.) that drive the intellectual agenda of the entity?
- What are the key performance indicators (i.e., metrics for success) for the entity?

Section 3: Entity Resources and Support

- What are the major funding source(s) (i.e., 25% or more of the total budget) for the entity? Check all that apply.
    - Foundation, other non-profit
    - Government (federal, state, and/or local)
    - Indirect cost return (from grants)
    - Industry
    - Private donors
    - Tuition
    - University
    - Other:
- Briefly describe the sustainability plan for the data science entity.
- In general, if an external grant is secured for data science activities within the entity, are Indirect Cost Returns distributed directly to the entity, to various units across campus, or something else?
- If the entity supports teaching, is tuition funding distributed directly to the entity, to various units across campus, or something else?
- Describe the dedicated space for the data science entity. How is it used? Who controls it? Does the entity have its own building?
- Was there a specific allocation of new faculty lines in connection with the establishment of the entity? If so, how many, how are these lines managed, and what requirements are placed upon new hires? How are the faculty lines funded? Are these lines tenure track and/or research faculty? Who makes hiring and tenure/promotion decisions?
- After the establishment of the entity, has there been or will there be more faculty hiring? If so, approximately how many per year, how are these lines managed, and what requirements are placed upon new hires? How are the faculty lines funded? Are these lines tenure track and/or research faculty? Who makes hiring and tenure/promotion decisions?
- Are there faculty appointments joint with the entity? If so, what are the service and teaching obligations to the entity?

- What are the core duties of the technical staff (e.g., applied data scientists and software engineers) and to whom do they report?
- Describe the computing support (both in terms of facilities and expertise) within the data science entity.
- Does the entity have administrative staff support? If so, what are their roles and/or titles?

Section 4: Education

- If applicable, please provide links to the entity's data science curricula for all levels of higher education.

In the remainder of this section, you will find two questions for each level of higher education (undergraduate, masters, and doctoral).

- At the undergraduate level, the data science entity offers: Check all that apply.
    - Introductory courses
    - Upper division courses
    - A data science minor
    - A data science major
    - Other:
- Does the entity provide support (e.g., teaching, curricula, and bootcamps) for other undergraduate programs on campus? If so, please describe the support.
- At the masters level, the data science entity offers: Check all that apply.
    - Part-time or full-time professional program
    - Professional certificate program
    - Minor or certificate for other programs
    - Full-time research degree program
    - Other:
- Does the entity provide support (e.g., teaching, curricula, and bootcamps) for other masters programs on campus? If so, please describe the support.
- At the doctoral level, the data science entity offers: Check all that apply.
    - Minor, track, specialization for other programs
    - Doctoral degree in data science
    - Other:
- Does the entity provide support (e.g., teaching, curricula, and bootcamps) for other doctoral programs on campus? If so, please describe the support.

Section 5: Faculty Engagement Strategies

- What are the strategies that the entity uses to engage faculty across the campus? Check all that apply.
    - Administrative support for starting and sustaining novel programs such as hackathons and workshops
    - Association/network of prestigious leadership faculty on campus
    - Be part of a cross-disciplinary community of data science practitioners
    - Dedicated space
    - Graduate student support

    - Internal research grants
    - Postdoctoral fellows support
    - Voice in the direction of data science on campus
- What are any additional incentives for faculty to participate in activities with the entity?
- Are there criteria for membership in the data science institute? If there are criteria for membership, describe it. Are there tiers (e.g., regular and affiliate)?

Section 6: Entity Activities and Additional Stakeholders

- What are the community building activities offered by the entity? Check all that apply.
    - Annual meeting, summit, or retreat
    - Career fairs
    - Training, tutorials, bootcamps
    - Symposia for general audience
    - Symposia for researchers
    - Weekly or monthly seminars
    - Other:
- Does the entity collaborate with campus libraries? If so, how?
- Does the entity have an industry affiliates program? If so, please describe the program.
- Beyond formal industry affiliate programs, does the entity collaborate with industry? If so, how?
- Is there an external advisory board for the entity? Mark only one oval.
    - Yes
    - No
    - Other:
- Are there any other stakeholders (e.g., campus learning centers, campus IT centers, independent entrepreneurs, etc.) for the data science entity? If so, who?

Section 7: Other Efforts and Entities

- Outside of the entity and to the best of your ability, briefly describe any major data science efforts at your university.
- If applicable, please list any other primary entities for data science at your university and complete the rest of this section.
- Do the entities collaborate? If so, how?
- How do the functions of the entities compare to or complement one another?

Section 8: Final Thoughts

- Is there anything else you would like to add? Include any surprising lessons learned, best practices, or other insights for setting up a data science entity.

C.3 Taxonomy Dimensions with Tabular Results

The table below summarizes the narratives from our informal survey. High-level taxonomy dimensions appears as rows. Data science entities in higher education appear as columns; in

some cases, multiple entities from the same institution responded to the survey. An “x” indicates that a taxonomy dimension is active in the data science entity.

The entries in the table are meant to convey a design space, i.e., the range of ways different campuses support data science. As many participating data science entities are in their formative stages, the entries are meant to be a snapshot of their status as of the writing of this report.

		Columbia University Data Science Institute	Georgia Institute of Technology Institute for Data Engineering and Science	Harvard University Harvard Data Science Initiative	Lehigh University Institute for Data Intelligent Systems and	Massachusetts Institute of Technology Institute for Foundations of Data Science	New York University Center for Data Science	North Carolina State University Institute for Advanced Analytics	Purdue University Integrative Data Science Initiative	Stony Brook University Institute for AI-Driven Discovery and Innovation	Syracuse University Applied Data Science at the iSchool	The Ohio State University Translational Data Analytics Institute	University of California, Berkeley Social Sciences D-Lab	University of California, Berkeley Division of Data Sciences	University of California, Berkeley Foundations of Data Analysis Institute	University of California, Irvine UCI Data Science Initiative	University of California, Irvine UCI Data Science Major	University of California, Santa Cruz Data Science Santa Cruz	University of Chicago Center for Data and Applied Computing	University of Chicago TBD - Data Science Education Initiative	University of Massachusetts Amherst Center for Data Science	University of Michigan Michigan Institute for Data Science	University of Pennsylvania Warren Center for Network and Data Sciences	University of Rochester Goergen Institute for Data Science	University of Washington eScience Institute	University of Wisconsin, Madison Institute for Foundations of Data Science	Yale University Department of Statistics and Data Science
Year established		2012	2016	2017	2018	2017	2012	2007	2018	2018	2017	2015	2013	2017	2018	2014	2015	2015	2018	2018	2015	2015	2013	2013	2008	2017	2017
Participants and Organization	Faculty lines	x						x	x	x		x				x				x	x			x	x		x
	Joint faculty appointments	x					x		x			x		x						x		x		x	x		x
	Technical staff (applied data scientists and software engineers)	x	x				x	x	x		x	x	x	x					x					x	x		x
	Administrative staff	x	x	x	x		x	x	x		x	x	x	x		x	x	x	x		x	x	x	x	x		x
	External advisory board	x			x			x	x									x			x	x		x			
Functions	Education	x		x		x	x	x	x	x	x	x	x	x	x	x	x			x	x	x		x	x	x	x
	Outreach - government	x	x	x	x		x	x	x			x		x					x		x	x		x	x		
	Outreach - industry	x	x	x	x	x	x	x	x	x		x		x		x		x	x		x	x		x	x		
	Research	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x		x	x	x	x	x	x	x
	University service - computational infrastructure		x		x				x		x	x	x								x	x					
	University service - expertise	x	x	x	x		x		x			x	x			x						x			x		x
	Other		x				x		x					x							x	x			x		
Research	Graduate student support	x			x		x	x	x	x	x		x	x				x	x		x				x	x	x
	Industry affiliates program	x	x	x	x		x		x			x		x		p		x			x	x			x		x
	Internal research grants	x	x	x	x	x	x		x	x	x	x							x		x	x		x			
	Postdoctoral fellowships	x		x	x	x	x		x	x				x	x				x		x	x	x		x	x	x
Undergraduate Education	Introductory courses	x							x		x			x						x	x	x			x		x
	Upper division courses								x		x			x						x		x		x	x		x
	A data science minor								x	x	x									x		x					
	A data science major								x					x			x			x		x		x			x
	Other	x	x	x					x			x	x	x						x		x		x	x	x	x
Masters-level Education	Part-time or full-time professional program	x					x	x	x		x	x		x						x	x	x		x	x		
	Professional certificate program	x							x		x										x				x		
	Minor or certificate for other programs									x	x											x					
	Full-time research degree program																					x					
	Other	x	x	x									x	x		x				x		x		x	x		x
Doctoral-level Education	Minor, track, specialization for other programs								x		x			x						x		x		x	x		
	Doctoral degree in data science						x														x						x
	Other	x	x	x	x			x	x	x			x	x		x						x					x
Community Building Activities	Annual meeting, summit, or retreat	x	x	x	x			x	x	x	x	x							x		x	x	x	x	x	x	
	Career fairs	x					x	x			x	x		x							x	x		x	x		
	Symposia for general audience	x		x			x	x	x			x		x		x		x	x			x	x	x			
	Symposia for researchers	x	x	x	x	x	x	x	x	x	x	x	x	x		x		x	x		x	x	x	x	x		
	Training, tutorials, and bootcamps	x	x		x	x	x	x	x				x	x		x			x			x	x		x	x	
	Weekly or monthly seminars	x	x	x	x	x	x		x		x	x		x				x	x		x	x		x	x	x	x
	Other	x	x				x		x													x			x		
Primary Funding Sources	Foundations, other non-profit						x		x					x					x						x		
	Government (federal, state, and/or local)		x		x	x	x			x	x	x			x	x		x			x	x		x	x	x	
	Indirect cost return (from grants)	x			x		x		x		x	x							x								
	Industry	x		x	x		x		x			x		x				x	x		x	x					
	Private donors	x		x				x	x	x				x					x	x			x				
	Tuition	x					x	x	x		x						x			x				x			
	University	x	x	x	x		x		x		x	x	x	x		x			x	x	x	x		x	x		x
	Other								x																		

C.4 Qualitative Summary of Survey Results

Establishment Date

Of those surveyed, the first entity was established in 2008. The majority of entities were established from 2014 to 2017. Some are still in planning stages.

Functions and Priorities

In response to identifying their entity's functions, the top three functions supported in rank order were reported to be: research, education, and outreach to industry. Several respondents wrote in functions such as organizational development, facilitating cloud access, policy activities, advancing the land grant mission of economic and workforce development, and marketing.

In response to identifying their entity's top priorities, the top priority identified was supporting interdisciplinary, collaborative research across the university. The next top priority was advancing methods and foundations in data science.

In response to, "What are the academic disciplines (e.g., computer science, statistics, domain sciences, etc.) that drive the intellectual agenda of the entity?," the most commonly mentioned disciplines were computer science and statistics. However, because of the different interpretations (see Section 2.4) of "intellectual [drivers]," we received four classes of responses: existing fields, subareas of fields, sectors/problem areas, and broad categories.

Collectively across all survey responses, the existing fields mentioned were: applied mathematics, astronomy, astrophysics, bioinformatics/biomedical informatics, biology, biomedical science, biostatistics, business, chemical engineering, chemistry, civil engineering, computer science, economics, electrical engineering, finance, government, health sciences, industrial engineering, information sciences, law, linguistics, marketing, materials science, mathematics, medicine, neuroscience, operations research, pharmacy, physics, political science, public health, science technology studies, and statistics.

These subareas of fields were listed: econometrics, high performance computing, learning analytics, machine learning, optimization, and theory of computing.

These sectors/problem areas were listed: energy, smart cities, and transportation.

These broad field-based categories were listed: engineering, humanities, life sciences, physical sciences, and the social and behavioral sciences. These even broader categories were listed: application areas, domain sciences, professional schools, and STEM fields.

Participants and Organization

The majority of entities are led by a director, an executive director, or multiple directors (e.g., co-directors, or a director with a deputy). Other entities are reportedly led by a faculty member, a department chair, or a committee of faculty members. Some entities include leadership from a Dean, a Provost, or in one case, the Senior Vice President of Research. The leader of the entity, usually a director, typically reports to a Dean, a Provost, a President, or some combination.

Many entities have reported active or planned faculty lines in connection with the establishment or growth of the entity, indicating a vibrant area of growth. Fewer entities reported joint faculty appointments tied to the entity.

Almost all entities have some administrative support staff. Many of them have complex administrative structures and very few do not have dedicated staff.

The existence of an external advisory board varied across entities; approximately half have an active or a planned external advisory board. Others reported annual ad hoc external reviews, an industrial advisory board, or a cross-sector steering committee.

Education

Relative to other educational activities, there are many professional Masters programs across the entities surveyed. In line with Recommendation #6, this finding emphasizes a need for standardization across the professional Masters programs in data science.

Many data science entities reported that they do not offer formal classes or degrees; instead, they supply materials and expertise to departments and colleges interested in data science.

Research and Community Building

Survey responses indicate a tremendous amount of enthusiasm for community building activities in the data science entities. When asked about the strategies used by the entity to engage faculty across campus in a check-all-that-apply-type question, the most commonly selected option was participating in a cross-disciplinary community of data science practitioners. The next most popular strategy was providing a voice to help shape the directions of data science on campus.

Many of the entities reported offering or organizing community building activities around data science. Organizing activities included symposia for a research audience or general audience. More than half of respondents reported 1) holding an annual meeting, summit or retreat, 2)

conducting trainings, tutorials, and/or bootcamps, or 3) holding weekly or monthly seminars. Some respondents indicated that they organize a career fair.

In response to membership criteria and structure questions, models varied widely, with some having loosely organized structures without official criteria, or minor expectations of service (e.g., serving as a faculty advisor for a small number of students), and others involving more requirements and formal processes to join (e.g., required participation in activities or courses; vetting of each regular and affiliate member by a membership committee or invitation by a director). In general, affiliate faculty appears to be a less formal role, subject to fewer obligations; regular or core faculty, or direct recruits have more formally defined roles and obligations. However, in general, across the entities surveyed, many opportunities were outlined for interested faculty to become involved.

Survey responses indicate a tremendous amount of enthusiasm to work with industry. Many expressed a desire for more formal interactions especially around data and tapping into expertise from industry, in line with Recommendation #9**.**

The existence of an industry affiliates program also varied by entity. Several entities reported having formal programs, with a tiered membership structure whereby companies pay membership fees to get benefits such as access to MS students taking a capstone course. Other entities described partnerships and relationships, but did not mention fees. Of the programs and informal interactions, some of the interaction mechanisms were: career fairs; industry participation in working groups, conferences, and seminars; research collaborations; project sponsorship; and collaborations through student capstone courses. Many of the respondents from entities lacking industry partnerships indicated that they are just now exploring and developing these programs.

In response to a question about other campus efforts, respondents described other major data science efforts, ranging from other institutes, research centers, and individual research efforts to creating new data science majors, departments, professional certificates, to hiring data-science focused faculty in existing departments across campus. Through the response to this question, it is clear that campuses are still grappling to figure out how to support data science campus-wide, but to do so in a way that balances top-down university supported efforts and grass-roots efforts from individual faculty and departments.

Structure and Funding

The majority of the funding for all entities is coming directly from university sources and from federal, state or local governments. In addition some reported funding from industry and tuition. However, many entities reported that the tuition does not directly come to them but is generally routed through other entities on campus. The majority of the entities cited federal funding in their sustainability plans, while very few cited tuition income. Similarly, indirect cost return (on grants) does not go directly to most of the surveyed entities.

From the entities surveyed, there is a spectrum of structures represented: at one extreme, entities reported acting as stand-alone enterprises, with dedicated space, faculty lines and data science majors and minors. At the other extreme, responses showed that a single university might have multiple data science entities on campus, each with a specific purpose. The majority of the entities are somewhere in between, where a single entity functions interdependently with other units on campus, and thus also shares the burdens of service, evaluation and support for faculty lines, etc.

Appendix D: Example Ethics Training Resources for Data Scientists

List of courses on ethics: This links to a shared spreadsheet of 57 (and counting) university courses on ethics and technology, including links to syllabi. The list is curated by Prof. Casey Fiesler at Colorado University.

Community Principles: This links to the Community Principles on Ethical Data Practices (formerly the Community Driven Principles for Ethical Data Sharing), which is a living document for ethics that changes as our understanding of ethics changes and evolves.

Appendix E: Related Landscape Reports

Publications and Reports

Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, and Alexander S. Szalay, "Realizing the Potential of Data Science," *Communications of the ACM,* vol. 61, no. 4, April 2018, pp. 67-72.

Business Higher Education Forum and PwC, Investing in America's data science and analytics talent, April 2017

Richard D. De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, Ping Ye, Curriculum Guidelines for Undergraduate Programs in Data Science, Annual Review of Statistics and Its Application 2017 4:1, 15-30

David van Dyk, Montse Fuentes, Michael I. Jordan, Michael Newton, Bonnie K. Ray, Duncan Temple Lang, Hadley Wickham, ASA Statement on the Role of Statistics in Data Science, October 2015
Luba Katz, "Landscape review of academic data science centers in the United States," Sloan and Moore Foundations, in preparation, 2018.
Envisioning the Data Science Discipline: The Undergraduate Perspective, National Academies, Washington, D.C., December 2017.

"Creating Institutional Change in Data Science," Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington, 2017.

Success, Opportunities, and Challenges for Statistics and Biostatistics in the Data Science Era, A Report of the July 2016 NSF-Sponsored Workshop for Chairs of Departments of Biostatistics and Statistics, American Statistical Association, 2016

National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18374

Workshops

- Workshop on Land Grant Universities and Data Science, NC State University, Raleigh, NC, June 5-6, 2018

- Enabling Computer and Information Science and Engineering Research and Education in the Cloud, Jan 8-9, 2018, Alexandria, VA
- Data Science Corps Workshop, December 7-8, 2017, Washington DC
- Workshop on Translational Data Science: Industry / Academic Confluence (TDS-IAC), Nov 13-14, 2017, Berkeley, CA
- Keeping Data Science Broad: Negotiating the Digital and Data Divide, Oct 30-Nov 1, 2017, Atlanta, GA
- 3rd Workshop on a Open Knowledge Network: enabling the community to build the network, October 4-5, 2017, Bethesda, MD
- NSF Workshop on Translational Data Science, June 26-27, 2017, Chicago, Il
- National Academies Government-University-Industry Research Roundtable on Data Matters: Ethics, Data, and International Research Collaboration in a Changing World (March 2018)
- Bloomberg and Brighthives: Code of Ethics

NSF Programs

- Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA)
- Partnerships between Science and Engineering Fields and the NSF TRIPODS Institutes (TRIPODS + X)
- Transdisciplinary Research in Principles of Data Science (TRIPODS)