April 2017

Roundtable on Data Science Post-Secondary Education

Meeting #2 - March 20, 2017

The second Roundtable on Data Science Post-Secondary Education met on March 20, 2017, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, California. Stakeholders from data science training programs, funding agencies, professional societies, foundations, and industry came together to discuss emerging needs and opportunities in data-intensive domains as well as case studies of three innovative data science education programs. This Roundtable Highlights summarizes the presentations and discussions that took place during the meeting. The opinions presented are those of the individual participants and do not necessarily reflect the views of the National Academies or the sponsors. Watch meeting videos or download presentations at **nas.edu/data-science-education-roundtable-2**.

EMERGING NEEDS AND OPPORTUNITITES IN DATA-INTENSIVE DOMAINS

English

Ted Underwood, *University of Illinois, Urbana-Champaign*

Underwood offered that there are both pedagogical opportunities for and challenges to integrating data science into an undergraduate English curriculum. Opportunities include the ability to explore unanswered research questions about significant cultural patterns in works of literature, such as how and why descriptions of different parts of the world have changed over time in fictional texts. Because literary data is abundant and relatively easy to reproduce, incorporating data science methods and tools into the curriculum offers a reliable means to answer such a question. Modeling techniques can even be used to develop a deeper understanding of genre or of the relationship between book sales and content. In response to a question from Nina Mishra, Amazon, about humanities insights gained through machine learning, Underwood noted that supervised classification algorithms can be used to categorize characters from novels, to address questions about how representations of gender have changed over time, and to help scholars to more easily and accurately identify lexical trends in fiction over past centuries.

However, it is rare for undergraduate English majors to have any exposure to quantitative coursework, and many do not understand the value of applying data science methods across disciplines. Digital humanities courses are surfacing on some campuses, but they typically prioritize digital media over computational methods and quantitative reasoning. Even then, many English departments typically hire only one "digital" instructor who offers a single course without much attention to quantitative foundations. As a result, many emerging researchers in the field are teaching themselves, and many current faculty members may be discouraged by the retraining needed to incorporate such content into the curriculum.

Underwood observed that humanities students often begin with computation and then move to statistics, which can make it challenging for students to understand how to *interpret* results. This lack of statistics training makes it especially difficult to interpret high-dimensional data. Assistance is needed to generate a pedagogical pipeline with a redesigned curriculum and more accessible courses. Peter Norvig, Google, suggested that English departments instead rely more on students from information sciences departments to solve data-driven problems. Kathleen McKeown, Columbia University, added that the envisioned pipeline seems unrealistic for English majors and proposed an intermediate path that would allow students to work on data science problems collaboratively across disciplines. Jessica Utts, University of California, Irvine, asked about the level of training that would be required for statisticians to be able to work with text. Underwood suggested that statisticians would need to refine their skills in linguistics and in the formulation of meaningful questions. But because literary students have important insights about and expertise in genre and history, Underwood would prefer to see humanities departments develop their own pedagogical pipelines rather than having data science disciplines mine the humanities.

Patrick Perry, New York University, inquired about the student demand for such coursework, while Antonio Ortega, University of Southern California, asked about the connection between such coursework and improved job prospects. Underwood responded that while employers want new hires to be able to write well, to tell clear stories, *and* to explore social components of data, many students are not yet encouraged to seek out new courses. Some English majors do enter the workforce with programming skills, though for many this knowledge may have been gained from a hobby or from a previous academic program. John Abowd, U.S. Census Bureau, asked how institutions might adapt, and Underwood expressed optimism that English departments will be at the forefront of change.

Astronomy

Joshua Bloom, *University of California, Berkeley*

Bloom credits the increased accessibility of data, computing power, and emerging technologies and methodologies with intensifying the competition for superior inferential capabilities. The most successful researcher will be the individual who knows how to ask the right question and who answers this question better and faster than her colleagues. This, in turn, relies on computational access, inference methods, creation and dissemination of a narrative, and reproducibility. This competitive environment reinforces the need for curricular changes related to data science training, as well as increased collaboration among domain experts and methodologists.

The discovery of the Higgs boson in 2012 and the direct detection of gravitational waves in 2016 demonstrate the value of combining domain expertise with methodological expertise to solve a data-driven problem arising from a large-scale physics experiment. In both instances, the use of novel hardware, computational infrastructure, and statistical methods

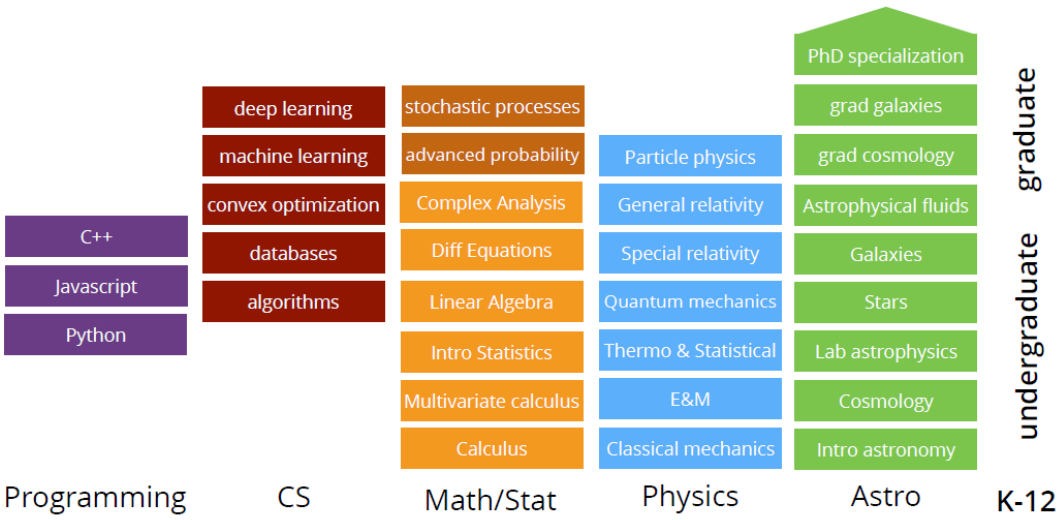

FIGURE 1 Expecting post-secondary students to become domain experts in astronomy while developing computer science, statistics, and programming skills presents a challenge. SOURCE: Joshua Bloom, University of California, Berkeley, presentation to the Roundtable.

was complemented by a team of diverse researchers asking the right questions and interpreting the results carefully. Such partnerships allow for high-impact discoveries and residual inventions. The team at the University of California, Berkeley, built and deployed a robust, real-time supervised machine learning framework, as well as a probabilistic source classification catalog on public archives with a novel active learning approach.

Bloom acknowledged that students in the physical science domains need to be trained to use new tools in order to make novel inferences and discoveries. However, there is too much content to cover in the data-driven domain education stack (Figure 1) to develop true expertise. Further discussion is needed to revise the curriculum in a way that will best serve students. For example, Jeffrey Ullman, Stanford University, suggested that computer science methods be introduced in high school instead of in college.

Bloom expressed concern that the continual release of novel methods and tools has made it increasingly difficult for people to keep pace with the training required for expertise in both domain knowledge and methodological skills. A successful approach to 21st century education could include training a person to develop *either* deep domain knowledge *or* methodological skills (Γ-shaped) instead of attempting to train a person to develop *both* deep domain knowledge *and* methodological skills (∏-shaped). The Γ-shaped people could then be encouraged to collaborate in multi-skill teams. To incentivize participation and discovery, it is crucial that novelty and rewards exist for all parties in such interdisciplinary teams.

History

Matthew Connelly (via teleconference), *Columbia University*

Connelly explained that the social sciences are structured differently from the physical sciences, and this may impact how data science topics are taught: In the social sciences, books are typically the preferred research products, teaching loads are often heavier, courses are usually taught in seminar settings instead of in labs, Ph.D. students can navigate more easily between programs and advisors, and coauthorship is rare. This last standard, in particular, limits opportunities for collaboration between social scientists and individuals who specialize in data science methods, which ultimately hinders the production of impactful research on large-scale problems. Such collaboration would be especially useful given that a large amount of historical data is often archived incompletely, and previously used qualitative methods may not be best suited to address certain contemporary research questions.

To discover and better understand historical events, historians could create topic models from event data sets. To identify patterns or anomalies in texts that may affect government policy, historians could rely on machine learning approaches. Because the data wrangling involved in such work is labor-intensive, Ph.D. students in the social sciences may need an additional 1 to 2 years of training to be able to master the analytical skills and computational methods required. A new sub-field of computational social sciences is slowly emerging, but there are still relatively few people who are capable of "doing it all." In response to a follow-up question from Perry, Connelly explained that the solution to this problem is *not* to simply throw a historical problem at a methodologist. Instead, *real* collaboration between domain experts and methodologists is the best way to achieve meaningful results from data and truly "do" history.

OPEN DISCUSSION

Collaboration and Communication

Emily Fox, University of Washington, and Connelly reiterated that institutions should encourage collaboration across disciplines rather than demand that students become experts in both a domain and data science methodologies. Bloom asked if there is a canon for data science similar to the canon in English literature; are there certain tools or methods that students should recognize without needing to develop expert-level knowledge? Connelly noted that because students will seek out training wherever they can find it, institutions should strive to make it easier for them to obtain the right skills.

Alok Choudhary, Northwestern University, advised that collaboration be genuine; both sides should contribute evenly in order to solve a problem. Mark Krzysko, U.S. Department of Defense, and Connelly suggested that faculty view effective collaboration and communication as explicit skill sets that need to be taught and developed. James Frew, University of California, Santa Barbara, agreed with Krzysko and Connelly that collaboration is a skill that must be taught but acknowledged that true collaboration can be complicated when there are institutional and disciplinary barriers to overcome. Krzysko also suggested that stakeholders ground themselves in the reality of building curriculum and opportunity for the talent they *have* instead of for the talent they *wish* they had.

Data Literacy and Course Design

Victoria Stodden, University of Illinois, Urbana-Champaign, has observed a growing demand from graduate students for a Ph.D. in data science, and she asked whether all science *is* data science. Bill Howe, University of Washington, suggested that the primary reason for the popularity of data science on college campuses over the past 20 years is the availability of large, noisy data. Eric Kolaczyk, Boston University, added

that sampling and design processes have also changed over the past two decades, further adding to the appeal of data science. And Nicholas Horton, Amherst College, later commented that data science tools are now much simpler and cheaper for a wider variety of users to manipulate.

Ullman worried about prescribing a specific data science program to first-year students who have not yet selected a major and would benefit from a broader introduction to the field. Fox noted the many challenges that already exist in trying to teach data science methods to students who think quantitatively, not to mention the challenges that will arise when trying to teach those same techniques to non-quantitative students. She reinforced the importance of ensuring that students understand what tools do, instead of simply how to use them. Howe suggested a lightweight organization of particular topics delivered by the domains as a potentially successful curricular model. Choudhary suggested re-evaluating general education curricula: could foundational concepts of data science be integrated into general mathematics and science courses instead of creating new, separate courses? Underwood agreed that there are implications for the future of the general education curriculum, which traditionally has as its mission to equip students with diverse skills and tools. Bloom acknowledged the importance of training students to be ready for careers possibly unrelated to their college majors. He advised that training not solely be vocational; rather, core concepts need to be emphasized as well. In this case, data literacy may be a more fruitful goal than simple science literacy. Mark Tygert, Facebook Artificial Intelligence Research, reminded participants to consider which skills or aptitudes are needed by industry; for example, since 95 percent of data science requires data wrangling, this is an area in which students need formal training. In response to a question from a webcast participant regarding on-ramps to data science for humanities students, Underwood noted that data visualization, and the ability to communicate the results of such an approach, is an important skill for humanities students to develop.

Charles Isbell, Georgia Institute of Technology, cautioned against conflating two separate issues: a data science degree and an education in data science. Chris Mentzel, Gordon and Betty Moore Foundation, suggested the Roundtable continue to explore the boundary between data science as a discipline and data science as a paradigm. McKeown noted that because the differences among disciplines and their approaches to research and teaching are so striking, it is unlikely that a one-size-fits-all model for teaching data science would be effective. Cathryn Carson, University of California, Berkeley, noted the importance of looking at the past trajectories of disciplines but suggested dedicating more effort to looking forward and trying to build new programs. McKeown acknowledged that it may be beneficial to have various experiments at different schools that do *not* converge. Abowd suggested that a discipline-based data science department may be needed to establish a pathway to diffuse knowledge into other disciplines more easily. Kolaczyk noted that programs grown from within generally have more success than those imposed from without. Stodden suggested that schools be deliberate with their vision for students by using the life cycle of data science as a curricular development tool. Doing so may engage younger students, allow a specialized trajectory, and emphasize the scientific components of data science. Kyle Stirling, Indiana University, noted that innovation in academia is incredibly difficult. And because there are vastly talented students in Master's programs without a shared vocabulary to communicate with one another, he also suggested implementing one-credit-hour on-ramps for students to learn fundamentals.

CASE STUDIES

University of Washington

Bill Howe, *University of Washington*

Howe explained that the University of Washington's eScience Institute was founded over 10 years ago, based on the notion that data-intensive science, enabled by intellectual infrastructure, would eventually be pervasive in industry and academia. The mission of the eScience Institute is to develop a cycle for data science that establishes working groups to bridge the gap between scientific theme areas and data science methodologies (Figure 2).

The University of Washington has a variety of formal data science education programs, including the following:

- A professional certificate in data science,
- A data science massive open online course (MOOC) with Coursera,
- An Information School data science sequence for undergraduate and graduate students,
- A Ph.D. with an advanced data science specialization,
- An undergraduate data science specialization, and
- An interdisciplinary data science Master's degree.

The goals of the MOOC, in particular, are to capitalize on students' interest in data science by exposing them to real problems; to strengthen delivering education at scale; to condense multiple courses into one introductory course; and to highlight the importance of database concepts in the broader data science discussion. This 8-week course includes instruction in the data science landscape, data manipulation at scale, analytics, visualization, and special applications. Approximately 9,000 students have enrolled in the course, although the largest population has been professional software engineers rather than the undergraduates the university had hoped to attract. From this experimental

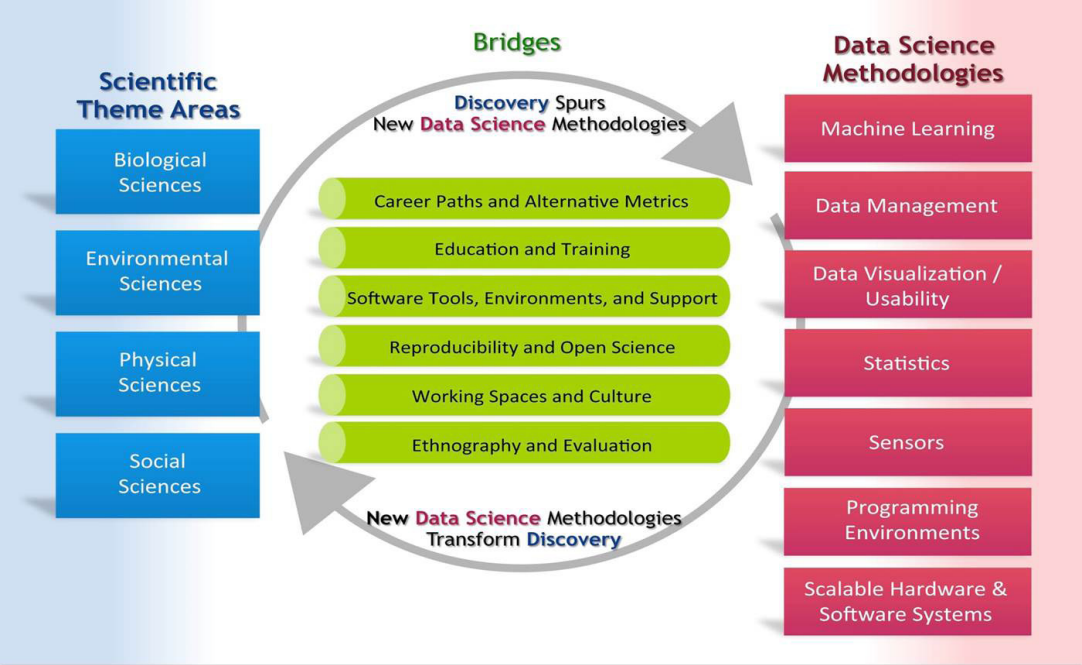

FIGURE 2 The eScience Institute's data science cycle includes education and training as a means to connect domain science inquiries to methodological developments. SOURCE: Produced by Ed Lazowska, University of Washington, and Moore/Sloan Data Science Environments and presented by Bill Howe, University of Washington, to the Roundtable.

MOOC, six themes emerged for the undergraduate data science curriculum: programming, data management, statistics, machine learning, visualization, and societal implications of data science. Though each domain might approach these themes in different ways, all have the capacity to satisfy these requirements.

Howe noted that the University of Washington is also introducing two large-scale courses available to all first-year students: (1) Introduction to Data Science Methods and (2) Data Science and Society. Concurrently, it plans to develop learning modules, increase advising support, and begin a topic review process for these courses. Ultimately, the university hopes to teach students to construct convincing arguments and to learn to manipulate large, noisy, heterogeneous data sets. Students will hone this skill by working on real problems, though they are not expected to become experts at the conclusion of either course.

Fox mentioned that because there are many different versions of data science classes offered at the university, course sequencing can become problematic. Howe noted that there are interdepartmental working groups in place to try to resolve such an issue so that students are enrolling in the appropriate prerequisites for more advanced courses. Stodden asked about the university's 5-year plan, as well as what other institutions can learn from its programs. Howe said the university would like to think more about workforce training as well as course topic refinement. In response to a question from Kolaczyk about institutional challenges, Howe acknowledged that the university is generally open and collaborative and has supported innovation in this area. However, streamlining the processes and developing an education working group could be beneficial.

Columbia University

Kathleen McKeown, *Columbia University*

McKeown described how a task force of deans and a data science directorate came together at Columbia University 1 year ago to discuss how to overcome the institutional barriers (e.g., differences in tuition and faculty load requirements across schools) that hinder the development of team-taught courses. As a result, the Columbia Collaboratory was formed, enabling funding for data scientists to partner with discipline specialists to team-teach classes across schools within the university. In the most recent round of funding, 18 requests for course proposals were submitted, and the following 4 were accepted:

- "Points Unknown: New Frameworks for Investigation and Creative Expression Through Mapping" (School of Journalism and School of Architecture, Planning, and Preservation), which reinforces the notion that data both defines and is part of city infrastructure;
- "Programming, Technology, and Analytics Curriculum for Columbia Business School" (School of Business and School of Engineering), which provides industry-specific data-intensive electives;
- "Computational Literacy for Public Policy" (School of International and Public Affairs and School of Engineering), which highlights the value of computational literacy for policy makers; and
- "Analysis to Action: Harnessing Big Data for Action in Public Health" (School of Public Health), which prepares students to translate data to non-scientific audiences.

These courses differ in their approaches and in how much programming students will do, given the specific needs of the individual disciplines, though there is a common emphasis on the value of communication. In addition to these four courses, additional pilot courses have been funded by the Collaboratory, such as "Data: Past, Present, and Future." That undergraduate course is taught by a historian and an applied mathematician, and it contains a core of knowledge that every citizen should have in order to understand data's role in society over the next century. This course contains two tracks, the technical and the humanist, which offer students a variety of assignments and applications to their majors.

In response to a question from Stodden, McKeown noted that student interest in team-taught courses is strong, though some of the funded courses have not yet operated (they will begin in Fall 2017). Underwood asked if there is a mechanism in place to ensure that such collaboration continues across schools, and McKeown confirmed that the deans of each school have already committed to working with the Collaboratory for a number of years.

University of California, Berkeley

Cathryn Carson, *University of California, Berkeley*

Carson recounted that the University of California, Berkeley, strives to enable all students to "engage capably and critically with data" in response to increased student demand for data science training and increased diversity in faculty expertise. In an effort to achieve this goal, the university offers a foundational data science course, "Data 8," (data8.org) to all students, no matter their educational backgrounds or majors of study. Currently, 700 students across 60 majors are enrolled in the course. This foundational course leverages a browser-based computational platform (Jupyter notebooks), and students learn computational and inferential thinking by working with real data in their societal and ethical contexts. No prerequisites are required to enroll, and the course is cross-listed in the departments of computer science, statistics, and information. This course, taught by an interdisciplinary team of faculty, is offered in tandem with "Connector" courses that link data science concepts directly to students' areas of interest and which are offered by a variety of academic departments. Such courses draw the university closer to the development of an integrative and comprehensive curriculum that better serves students. These course offerings have thus far been possible as a result of the university's bottom-up collaborative and innovative culture. The data science education philosophy at the university is centered on intellectual, organizational, and social values, and it relies on the motto, "Try, Learn, and Scale it Fast."

For students who wish to build on this platform after they have completed "Data 8," faculty have developed a number of other new courses, including the following: "Data Science 100: Principles and Techniques of Data Science" (Figure 3), "Stat 28: Statistical Methods for Data Science," and "Stat 140: Probability for Data Science." As the university continues to expand its offerings, it has begun to scaffold a data science major and minor, both shaped by a collaborative approach.

The University of California, Berkeley, currently offers a short course for faculty to learn more about data science peda-

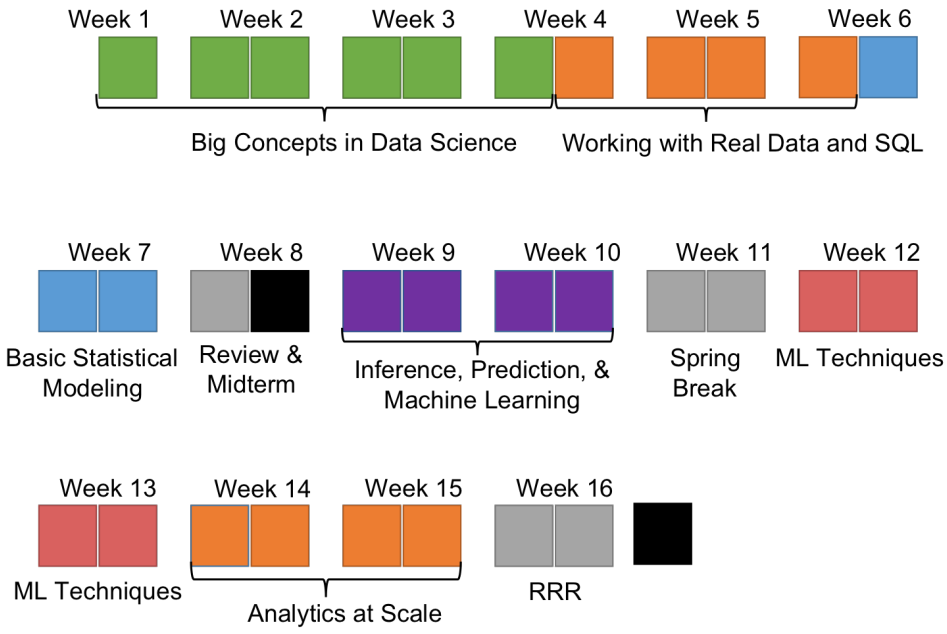

FIGURE 3 The syllabus for a pilot "Data Science 100" course, inspired by the data science life cycle. SOURCE: Produced by Professor Joseph Gonzalez, University of California, Berkeley, and presented by Cathryn Carson, University of California, Berkeley, to the Roundtable.

gogy and practice, and a number of course modules came directly from this work. There is also a student team working on data science education curriculum development, outreach and diversity, and program infrastructure. A central question that continues to be explored is how to collectively meet the needs of the students and faculty from each domain, according to Carson.

McKeown asked about the challenge of presenting information to "Data 8" students who may have diverse experiences and educational backgrounds. She wondered if "Data 8" would eventually need to be offered in a variety of formats and at different levels. Carson said that the university wants to keep the course diverse in terms of students' incoming knowledge but acknowledged that it is likely a challenge that will have to be addressed in the coming years. There is currently a pre-experience summer immersion program called "Summer Bridge" that offers preparation for students who may not feel ready for "Data 8." There are also in-course adaptations available so that the course is accessible to all participants. In response to a question from Perry, Carson noted that different students struggle with different aspects of the course; for example, some students find coding to be difficult during the first few weeks of the class. Currently, students' receptivity to the course content is gauged ethnographically; however, Carson would like to see analytics used to measure student interest and success in the course in the future.

OPEN DISCUSSION

Considering Politics and Society

Stodden reiterated that the politics of a university are central to any discussion of course creation or modification. Institutional philosophies surrounding leadership and funding have the potential to make or break data science initiatives. Underwood agreed, adding that the hub and connector model at the University of California, Berkeley, provides an appealing gateway to increase the visibility of data science among humanities students. Stodden also expressed concern about a shortage of professors if the demand for data science courses continues to increase, but domain-based courses could alleviate this strain on faculty. Ullman and Isbell noted that it would be valuable to collect data on how different schools are handling various challenges. Isbell pointed out that a university's organizational structure adds another dimension to the decision-making process. For example, colleges within universities have their own standards and expectations. Thus, a "middle-out" approach may be more effective than a "bottom-up" approach in a state institution that has very different issues from a private institution. Deborah Nolan, University of California, Berkeley, highlighted the value of discussing the role of data science in community colleges, as they too will have unique political and organizational challenges.

Transforming Culture

Ullman highlighted the value of engaging professors in designing cross-disciplinary and experimental courses. At Stanford University, a small freshman seminar entitled "Big Data, Big Hype, Big Fallacies" was delivered in 2016, successfully linking computer science, humanities, and social science concepts. Stanford hopes to offer this as a regular course, open to all students, in future terms. Horton later added that any institutional plans to create cohorts of teaching faculty (with job security and professional development opportunities) need to be fast-tracked to address the challenges that data science curricula present. Carson noted the value of studying the many online data science courses that are already available before revising traditional undergraduate curricula. She also reiterated that a one-size-fits-all approach is impractical but emphasized that an open and collaborative culture can be grown on campuses. Kolaczyk noted the value of establishing local partnerships and encouraging face-to-face interactions when trying to build a culture of collaboration. Isbell cautioned, however, that it can take several years to change a campus culture.

Transitioning Platforms

Abowd discussed the enormous challenges that exist in conducting training for in-place workforces (especially government agencies) on in-place computing, data management, and software infrastructures. Anticipating which tools users will need in the workplace can also be difficult. Krzysko agreed that there are significant challenges in training and leading large, diverse workforces. Unlike the commercial world, government agencies face bureaucratic obstacles for deploying software. It could be beneficial for the education system to increase collaboration with both government agencies and policy makers to find more efficient ways to access new technologies. He also highlighted the misalignment between graduating students' creative aspirations about emerging data science opportunities and the realities of workforce capabilities: Universities instill a "what if" mantra in their job-seeking students, while parts of the current workforce respond with "you can't." Krzysko said that employers could work more closely with motivated and highly trained students to create pathways to middle management in the hopes of preserving their enthusiasm. Frew agreed that there needs to be a better relationship between universities and hiring bodies; at the University of California, Santa Barbara, the Master's program works closely with employers to understand what they view as shortcomings in new hires. Relying on a simple supply and demand philosophy, the university then uses this information to better train its students.

ABOUT THE ROUNDTABLE: The Roundtable on Data Science Post-Secondary Education is supported by the Gordon and Betty Moore Foundation, the National Institutes of Health Big Data to Knowledge, the National Academy of Sciences W. K. Kellogg Foundation Fund, the Association for Computing Machinery, and the American Statistical Association. Within the National Academies, this roundtable is organized by the Committee on Applied and Theoretical Statistics in conjunction with the Board on Mathematical Sciences and Their Applications (BMSA), the Computer Science and Telecommunications Board (CSTB), and the Board on Science Education (BOSE). Roundtable meetings will take place approximately four times per year. Please address any questions or comments to Ben Wender at **bwender@nas.edu**.

DISCLAIMER: This meeting recap was prepared by the National Academies of Sciences, Engineering, and Medicine as an informal record of issues that were discussed during the Roundtable on Data Science Post-Secondary Education at its second meeting on March 20, 2017. Any views expressed in this publication are those of the participants and do not necessarily reflect the views of the sponsors or the National Academies.

ROUNDTABLE MEMBERS PRESENT: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; John Abowd, U.S. Census Bureau; Ron Brachman, Cornell University; Alok Choudhary, Northwestern University; Emily Fox, University of Washington; James Frew, University of California, Santa Barbara; Nicholas Horton (via webcast), Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Chris Mentzel, Gordon and Betty Moore Foundation; Nina Mishra, Amazon; Deborah Nolan, University of California, Berkeley; Peter Norvig, Google; Antonio Ortega, University of Southern California; Patrick Perry, New York University; Victoria Stodden, University of Illinois, Urbana-Champaign; Mark Tygert, Facebook Artificial Intelligence Research; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

GUESTS PRESENT: Joshua Bloom, University of California, Berkeley; Cathryn Carson, University of California, Berkeley; Matthew Connelly (via teleconference), Columbia University; Kyle Stirling, Indiana University; and Ted Underwood, University of Illinois, Urbana-Champaign.

STAFF PRESENT: Linda Casola (via webcast), Research Associate, BMSA; Janel Dear, Senior Program Assistant, CSTB; Michelle Schwalbe, Board Director, BMSA; Ben Wender, Program Officer, BMSA.

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org

Copyright 2017 by the National Academy of Sciences. All rights reserved.