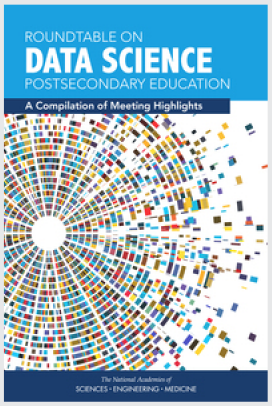




# Roundtable on Data Science Postsecondary Education: A Compilation of Meeting Highlights (2020)

## DETAILS

223 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-67770-7 | DOI 10.17226/25804

## CONTRIBUTORS

Linda Casola, Rapporteur; Board on Mathematical Sciences and Analytics; Committee on Applied and Theoretical Statistics; Computer Science and Telecommunications Board; Board on Science Education; Division on Engineering and Physical Sciences; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

## SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2020. *Roundtable on Data Science Postsecondary Education: A Compilation of Meeting Highlights*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25804.

# ROUNDTABLE ON DATA SCIENCE POSTSECONDARY EDUCATION

## A Compilation of Meeting Highlights

Linda Casola, Rapporteur

Board on Mathematical Sciences and Analytics

Committee on Applied and Theoretical Statistics

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

Board on Science Education

Division of Behavioral and Social Sciences and Education

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
*Washington, DC*
**www.nap.edu**

**THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001**

This activity was supported by Contract No. 5018 with the Gordon and Betty Moore Foundation, Contract No. HHSN263201200074I with the National Institutes of Health, the National Academy of Sciences W.K. Kellogg Foundation Fund, the Association for Computing Machinery, the American Statistical Association, and the Mathematical Association of America. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-67770-7
International Standard Book Number-10: 0-309-67770-X
Digital Object Identifier: https://doi.org/10.17226/25804

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.



Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2020. *Roundtable on Data Science Postsecondary Education: A Compilation of Meeting Highlights.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25804.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

## ROUNDTABLE ON DATA SCIENCE POSTSECONDARY EDUCATION

ERIC KOLACZYK, Boston University, *Co-Chair*
KATHLEEN R. McKEOWN, Columbia University, *Co-Chair*
JOHN M. ABOWD, U.S. Census Bureau
DEB AGARWAL, Lawrence Berkeley National Laboratory
RON BRACHMAN, Cornell University
JEFFREY BROCK, Yale University
BRIAN CAFFO, Johns Hopkins University (until December 2017)
ALOK CHOUDHARY, Northwestern University
RONALD COIFMAN, NAS,[1] Yale University (until December 2017)
MICHELLE DUNN, National Institutes of Health (until March 2017)
E. THOMAS EWING, Virginia Tech
EMILY FOX, University of Washington
JAMES FREW, University of California, Santa Barbara
CONSTANTINE GATSONIS, Brown University
JOHANNES GEHRKE, Cornell University (until December 2017)
LISE GETOOR, University of California, Santa Cruz
MARK L. GREEN, University of California, Los Angeles
ALFRED O. HERO III, University of Michigan
NICHOLAS J. HORTON, Amherst College
ERIC HORVITZ, NAE,[2] Microsoft Research
BILL HOWE, University of Washington
CHARLES ISBELL, Georgia Institute of Technology
MARK E. KRZYSKO, U.S. Department of Defense
DUNCAN TEMPLE LANG, University of California, Davis
RACHEL LEVY, Mathematical Association of America
BRANDEIS MARSHALL, Spelman College
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
ALEX PENTLAND, NAE, Massachusetts Institute of Technology (until December 2017)
CLAUDIA PERLICH, Dstillery
PATRICK PERRY, New York University
MEHRAN SAHAMI, Stanford University

[1] Member, National Academy of Sciences.
[2] Member, National Academy of Engineering.

VICTORIA STODDEN, University of Illinois, Urbana-Champaign
URI TREISMAN, University of Texas, Austin
MARK TYGERT, Facebook Artificial Intelligence Research
JEFFREY D. ULLMAN, NAE, Stanford University
JESSICA M. UTTS, University of California, Irvine
JANE YE, National Institutes of Health

**Staff**

TYLER KLOEFKORN, Program Officer, Board on Mathematical Sciences and Analytics
MICHELLE SCHWALBE, Director, Board on Mathematical Sciences and Analytics
BEN WENDER, Senior Program Officer, Board on Energy and Environmental Systems
SELAM ARAIA, Senior Program Assistant, Board on Mathematical Sciences and Analytics
LINDA CASOLA, Associate Program Officer, Board on Mathematical Sciences and Analytics
BETH DOLAN, Financial Manager (until May 2019)
CHRISTOPHER FU, Research Associate (until August 2019)
ADRIANNA HARGROVE, Financial Manager
RODNEY HOWARD, Administrative Assistant (until July 2018)
CATHERINE POLLACK, College Intern (until August 2018)

[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.
[3] Member, National Academy of Medicine.

## COMMITTEE ON APPLIED AND THEORETICAL STATISTICS

ALFRED O. HERO III, University of Michigan, *Chair*
ALICIA CARRIQUIRY, NAM,[1] Iowa State University
RONG CHEN, Rutgers University, State University of New Jersey
MICHAEL J. DANIELS, University of Florida
KATHERINE BENNETT ENSOR, Rice University
AMY H. HERRING, Duke University
TIM HESTERBERG, Google, Inc.
NICHOLAS J. HORTON, Amherst College
DAVID MADIGAN, Columbia University
XIAO-LI MENG, Harvard University
JOSÉ M.F. MOURA, NAE,[2] Carnegie Mellon University
RAQUEL PRADO, University of California, Santa Cruz
NANCY M. REID, NAS,[3] University of Toronto
CYNTHIA RUDIN, Duke University
AARTI SINGH, Carnegie Mellon University
ALYSON G. WILSON, North Carolina State University

### Staff

TYLER KLOEFKORN, Director
CARL-GUSTAV ANDERSON, Associate Program Officer
SELAM ARAIA, Senior Program Assistant
LINDA CASOLA, Associate Program Officer
ADRIANNA HARGROVE, Financial Manager

[1] Member, National Academy of Medicine.
[2] Member, National Academy of Engineering.
[3] Member, National Academy of Sciences.

## COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD



[1] Member, National Academy of Engineering.
[2] Member, National Academy of Sciences.

## BOARD ON SCIENCE EDUCATION



### Staff



[1] Member, National Academy of Sciences.

# Acknowledgment of Reviewers

This compilation was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each publication as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this document: Nicholas Horton, Amherst College, and Brian Kotz, Montgomery College. We also thank staff member Scott Weidman for reading and providing helpful comments on the manuscript.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the compilation nor did they see the final draft before its release. Neither did the members of the Roundtable on Data Science Postsecondary Education, nor the members of the Board on Mathematical Sciences and Analytics, Committee on Applied and Theoretical Statistics, Computer Science and Telecommunications Board, or Board on Science Education, whose role was limited to organizing and overseeing the roundtable. This compilation chronicles the presentations and discussions at roundtable meetings, and the statements and opinions contained herein are those of individual participants and are not necessarily endorsed by other participants or the National Academies. The review of this compilation

was overseen by Alicia Carriquiry, NAM,[1] Iowa State University. She was responsible for making certain that an independent examination of this compilation was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the rapporteur and the National Academies.

[1] Member, National Academy of Medicine.

# Contents

# 1

# Introduction

Established in December 2016, the National Academies of Sciences, Engineering, and Medicine's Roundtable on Data Science Postsecondary Education was charged with identifying the challenges of and highlighting best practices in postsecondary data science education. Convening quarterly for 3 years, representatives from academia, industry, and government gathered with other experts from across the nation to discuss various topics under this charge. Some stakeholders argue for data science to be described as a discipline, others as a domain, and still others as an umbrella. No matter the label, academia is in the midst of a transformation that will continue to have profound implications across society. In an effort to train postsecondary students effectively, institutions of higher education are (re)examining who is taught what and why, as well as how and by whom, and considering how to increase interactions with students' potential employers.

This introduction serves to orient readers to four central themes that emerged during the roundtable meetings: (1) foundations of data science; (2) data science across the postsecondary curriculum; (3) data science across society; and (4) ethics and data science. These themes are expanded in the chapters that follow, which contain detailed summaries of each roundtable meeting. These meeting summaries, as well as original meeting videos, are also available online.[1] These meeting recaps were prepared by the

[1] Watch meeting videos or download presentations at https://www.nationalacademies.org/our-work/roundtable-on-data-science-postsecondary-education, accessed February 13, 2020.

National Academies of Sciences, Engineering, and Medicine as informal records of issues that were discussed during meetings of the Roundtable on Data Science Postsecondary Education. All opinions presented are those of the individual participants and do not necessarily reflect the views of the National Academies or the roundtable sponsors.

## FOUNDATIONS OF DATA SCIENCE

Roundtable participants discussed the type(s) of training best suited for robust data science practice and considered what data science education could look like from a national perspective. No consensus opinion emerged as to how data science should be defined or what a data science degree should require. Nonetheless, roundtable participants agreed that this increasingly interdisciplinary field depends on foundational elements from many disciplines, including but not limited to statistics, computer science, engineering, and mathematics. Participants noted an abundance of foundational skills, techniques, and concepts—from one discipline or common to many—that are integral to proficiency in data science (see Chapter 2). As data science courses, programs, and degrees continue to evolve, consensus foundations may emerge but will depend on institutional contexts and opportunities.

## DATA SCIENCE ACROSS THE POSTSECONDARY CURRICULUM

Roundtable participants discussed the intersection of data science and domain sciences and contemplated how this interplay impacts the teaching of data science. Faculty are challenged to learn new skills, adapt methods, and find new (often multidisciplinary) approaches to teaching data science. At the same time, student demand for data skills continues to increase, and data, computation, and software tools are becoming pervasive. Several promising approaches to postsecondary data science education have emerged, such as integrating data science perspectives into existing data-intensive domain courses; creating new courses that integrate multiple perspectives, skills, and fields; and teaching collaboratively (see Chapter 3). These and other paths forward could be implemented successfully alongside the following strategies: training graduate teaching assistants across a range of skills; identifying and better supporting faculty who are willing to experiment with and assess new approaches; improving the understanding of disciplinary needs for data science; developing methods to introduce data science to students without quantitative training; integrating standard disciplinary data sets to support data science instruction; and lessening traditional seminar teaching and single-author monograph publishing. A shared goal of many data science

educators is the creation of investigators with deep expertise in either a domain or data science and with enough knowledge of the other to collaborate effectively (see Chapters 2 and 3).

### Ph.D. Programs

Ph.D. programs in data science are more nascent than undergraduate and master's programs in data science. Approaches to Ph.D. training in data science—such as a new entity created with existing faculty, an expansion of an existing entity, or an overlay model—often depend on where the Ph.D. program is housed within an academic institution and how it interacts with other departments. Each approach serves a unique purpose, and, collectively, these approaches are quickly creating options for advanced data science education. These approaches differ in their administrative mechanisms and application processes. For example, doctoral students could be directly admitted into a data science program or admitted into a home department; in some cases, admission into a data science program only happens after a student arrives on campus. Several programs compel students to complete all requirements in their home departments before completing additional requirements for data science. The ability to carry out a broader dissertation and research, often with interactions with multiple scientific domains, is advantageous in a data science Ph.D. program. Faculty flexibility and a willingness to work within the constraints of an institution are beneficial, especially given the challenges associated with starting new programs. In the future, it is likely that there will be some consolidation of the approaches taken, although it is unlikely (and undesirable) that one approach will fit all institutions. Evaluative measures could be useful to determine which programs flourished, how requirements varied across different programs, what types of dissertations were produced, and where graduates were hired (see Chapter 8).

### Two-Year Programs

Demand for employees with data science skills is expanding across industries, and some of today's data science jobs could be filled by individuals with 2-year (associate's) degrees. Many efforts are under way to better understand and align with the needs of employers, such as the development of data career pathways and expert worker profiles. Institutions are revising curricula accordingly to reflect the changing demands of the skilled workforce. Several 2-year colleges are developing courses, certificates, and associate's degrees in data science, data analytics, and data management. Incentivizing faculty training is key to the success of these

programs. Because funding and resource constraints make it difficult to implement new programs at 2-year colleges, there is value in establishing formal partnerships with nearby 4-year and master's-granting institutions. Two-year colleges are also considering the potential for transfer as they design their programs and are beginning to develop articulation agreements that could create smoother transitions for students in search of advanced training (see Chapter 12).

### Alternative Pathways

The past decade has seen substantial experimentation in how data science education is delivered both within and outside the classroom. Three alternative mechanisms within academia include certificate programs, practicums, and collaborative environments. These mechanisms challenge the traditional model, where practice—if incorporated at all into the curriculum—is more likely to be encountered through a class project or a capstone experience. Other unique educational opportunities include hackathons, boot camps, and activities in informal settings such as museums. Boot camps provide a way to fill the increasing gaps between degree-based programs in academia and on-the-job learning in industry, with focused, problem-based, team-oriented programs that build proficiency with the data science life cycle. Wider dissemination of successful efforts could be helpful for the efficient use of resources as well as to scale emerging best practices (see Chapter 5).

## DATA SCIENCE ACROSS SOCIETY

Roundtable participants further examined the development of data science expertise for the workplace. Effective teamwork and the ability to communicate clearly with diverse audiences are particularly important skills. High-quality, free, online training is readily available, and assessments have revealed that participants are incentivized by the opportunity to work with real data sets to solve real problems (see Chapter 4). Increased coordination between academia and industry (as well as between academia and government) could be key to the future of robust data science education and practice. Students and faculty could benefit from the opportunity to spend time in industry in the form of prolonged internships and postdoctoral assignments or with joint appointments, respectively. Academic institutions could stimulate successful partnerships by leveraging experiences from other disciplines; benchmarking and developing best practices; fostering continued interactions; providing firm financial support; offering resources and incentives to both

students and faculty; increasing diverse representation; developing synergistic relationships with neighboring institutions; building on credentials in well-established areas; creating legally binding master agreements; embracing open source, open data, and open science; and using cloud platforms (see Chapter 11).

## ETHICS AND DATA SCIENCE

Roundtable participants posed several questions about ethical and privacy issues related to data science education (see Chapter 6):

- What does it mean for an algorithm to be transparent, interpretable, or explainable?
- What rights should individuals have when they are the subjects of algorithms, and how do these rights interface with existing legislation?
- Who is responsible for the effects of how data are used?
- What information about fairness could data scientists supply that is suitable for a range of metrics for fairness? How does one close the feedback loop from metrics of fairness back to the design of algorithms?
- What rights should individuals have about keeping their data private?
- What are the sources of bias in algorithms and in data science more generally? Could they be eliminated or substantially lessened by explicit protocols and policies?
- How could those with technical knowledge most accurately and understandably present trade-offs? Would advance knowledge of trade-offs skew the results against privacy and fairness?
- What are the lessons learned from other disciplines?
- How and at what level could students be taught about ethics and privacy?

Rigorous approaches to these ethical questions are being implemented in research and in academic institutions through new courses on ethics in data science and through modules as part of other data science courses. In this time of innovation, making teaching materials widely and quickly available could help to expand ethical conversations in the classroom. The discussion of ethics and privacy in data science education could be broadened to include perspectives for social science, philosophy, industry, law, and policy as the research begins to delve deeper into issues of accountability, transparency, fairness, privacy, and bias (see Chapter 6).

### Reproducibility

Reproducibility in computationally enabled research has been an area of active discussion in the academic community for several decades; this discussion influences data science practice and teaching on both theoretical and practical levels. Computational tools (e.g., the Jupyter Notebook[2]) can help to address issues of reproducibility and transparency. Computational transparency permits not only the understanding of the reasoning behind scientific findings but also the comparison of results that may differ and yet claim to answer the same scientific question. These efforts inform modern practices in software development and coding for data science such as version control, the use of notebooks, and skills in specific languages (e.g., Python). The adaptation of techniques and tools from software engineering, database management, computing at scale, and statistical inference is essential to data science practice, but these techniques and tools do not guarantee the accuracy of the resulting scientific findings. Generally accepted standards for teaching computational transparency and reproducibility in data science could be useful, as could generally accepted standards for best practices in software engineering in data science applications (see Chapter 7).

### Social Good

Approaches to engaging students in meaningful projects with the potential for social impact are rapidly emerging and these efforts could help to attract and retain future data scientists. Questions remain about which types of institutions are able to provide these experiences; whether emerging programs are particularly resource-intensive, scalable, and conducive to academic or industry reward structures; and how to best prepare people with different levels of authority in academic and industrial settings to be able to raise and discuss ethical issues. The data science community more broadly could benefit from a process that builds trust between technologists and the social sector, increases attention to data collection and security, explains conclusions drawn from models, and plans for cases when harm is done to users (see Chapter 10).

### Diversity and Inclusion

Many of the same institutional barriers that have impeded equity and inclusion in STEM affect data science education—for example, the rigidity of the faculty reward system and implicit biases in faculty hiring and

[2] The website for the Jupyter Notebook is https://jupyter.org/, accessed February 13, 2020.

promotion and graduate school admission. The data science community has succeeded in raising awareness of the importance of inclusion, in part owing to a nationwide shortage of data scientists. Mentorship programs and cohort experiences have been particularly successful in recruiting and retaining underrepresented groups for data science education. Given that academic institutions are slow to change, especially with regard to rewarding faculty involvement in activities that do not result in peer-reviewed publication, partnership with industry could be a promising avenue to increase diverse participation in data science. Other potential paths toward success could include a more coordinated effort to involve teachers, counselors, and administrators in implementing change at the K-12 level; increased assessment and the sharing of best practices; and a stronger connection between inclusive academic programs and organizations working to increase inclusive participation in the field of data science (see Chapter 9).

# 2
# Meeting #1: The Foundations of Data Science from Statistics, Computer Science, Mathematics, and Engineering

The Roundtable on Data Science Postsecondary Education met on December 14, 2016, at the Keck Center of the National Academies of Sciences, Engineering, and Medicine in Washington, D.C. Stakeholders from data science training programs, funding agencies, professional societies, foundations, and industry came together to discuss data science education and practice, the needs of the community and employers, and ways to move forward. Roundtable members also examined foundations of data science from the fields of statistics, computer science, mathematics, and engineering and considered the needs of diverse data science communities. 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.

## FOUNDATIONS OF DATA SCIENCE

### Statistics

*Jessica Utts, University of California, Irvine*
*Nicholas Horton, Amherst College*

As a result of accelerating technological developments, larger bodies of available data, and increased interest in modeling and quantification, statistics is understood and taught quite differently today than it was

in the 1990s. According to the American Statistical Association (ASA), foundational data science should include the fields of database management, statistics, machine learning, and distributed parallel systems, and it should be introduced not only at the undergraduate level, but also at the K-12 levels. Statistics plays an important role in data science because it allows questions to be framed in a way that encourages better use of the data, inferences to aid in quantifying uncertainty, interventions to be identified by distinguishing between causation and correlation, methods to be used for prediction and estimation, and findings to be reproducible.

The cycle used to carry out statistical investigation includes the problem, the plan, the data, the analysis, and the conclusions (often abbreviated as the PPDAC cycle). The ASA notes that skills in computing, software, programming, data wrangling, algorithmic problem solving, and communication are needed to work with data and execute the PPDAC cycle, and thus should be part of the formal curriculum. With proper training, statisticians offer a valuable contribution to data science because they can understand context, account for variability, design and analyze data, understand inference, foster reproducibility, work in multidisciplinary teams, and make data-driven decisions.

### Computer Science

*Charles Isbell, Georgia Institute of Technology*

The three educational pillars of computing are as follows:

1. Basic foundations (e.g., understanding data through algorithms, machine learning, curation, visualization/modeling, and computational systems);
2. Advanced foundations (e.g., understanding large-scale data through high-performance computing and advanced machine learning); and
3. Practicums (e.g., applying knowledge to real-world problems through data engineering).

Models (containing data), languages, and machines are equally important, which reinforces the interdisciplinarity of data science. And because choices made while developing the algorithms may embed policy decisions or biases, ethics must also play a central role in any data science curriculum.

Bill Howe, University of Washington, noted that software engineering design is an important new component of computer science that should be tailored for data science education. Alok Choudhary, Northwestern University, raised the importance of applications and high-performance

computing for data science. John Abowd, U.S. Census Bureau, noted that disciplinary jargon is problematic; if computer scientists adopted more accessible language, their literature would be more easily understandable to a greater number of people. Victoria Stodden, University of Illinois, Urbana-Champaign, focused on the importance of developing standards and best practices for software, while Mark Tygert, Facebook Artificial Intelligence Research, wondered about the role of programming in future curricula.

### Engineering

*Alfred Hero, University of Michigan*

Engineers want to educate students to build reliable systems; however, the data-mining pipeline needs to be reimagined to make better decisions. Standards are an important part of this, including standards to deal with the growing number of citations to analysis software and the proliferation of software packages. Engineers view data science as a way to collect data (e.g., through sensing instruments and data repositories), to manage data (e.g., through resilient and protected databases), and to analyze data (e.g., with integrated computational algorithms). Data-enabled engineering, for example, is used in the materials genome, for precision medicine, and in cyber-physical networks. Data science is naturally multidisciplinary, and many disciplines rely on data science tools and principles that draw from mathematics (e.g., data as topological object); computer science (e.g., data as lists/graphs); statistics (e.g., data as random sample); informatics (e.g., data as interface); physics (e.g., data as natural phenomena); and engineering (e.g., data-to-decision).

The University of Michigan offers an undergraduate degree program in data science engineering, a graduate data science certificate program, an extracurricular data science student organization, and a weeklong summer camp for high school students. Because undergraduate students cannot be expected to become universal experts, it might make sense in the future to offer a B.A. or B.S. degree in data science with a concentration in a domain science.

### Mathematics

*Eric Kolaczyk, Boston University*
*Ronald Coifman (in absentia), Yale University*

Data science is typically divided into one of two categories: computational sciences (e.g., computer science, engineering, and statistics)

or domain sciences (e.g., genomics, neuroscience, text analysis). In both areas, there is a mathematical infrastructure: the computational sciences are supported by linear algebra, numerical analysis, and graph theory; the domain sciences are problem-specific, use physical and life sciences, and rely on physical models and mathematical analysis tools. Linear algebra, analysis, geometry, and optimization have always been essential tools used to model our world, and, with some adaptation, they will continue to be so. Mathematics can provide theoretical models, a conceptual framework, a language, and a related "calculus" for data science. A mathematical conceptualization of modern data science involves a blend of subfields in an integrative curriculum in which the varied mathematical tools are explained and jointly motivated. Moving forward, educators should consider how to evolve the mathematics curriculum to meet data science needs as well as how to better foster integrative teaching and learning.

### Open Discussion

Abowd opened the discussion by asking whether it is possible to develop a data science canon without having a mathematical model at the center. Kolaczyk posed a related question about the extent to which students need to understand mathematical structures relative to their tasks. Hero noted that current data science curricula are missing an analytical component; tools currently do not exist that are certified by the community as applicable to a variety of problems. Lou Gross, University of Tennessee, Knoxville, added that mathematics is a language of abstraction, and there is a key role for abstraction in data science. He continued that data science has the potential to create unity across disparate areas of mathematics. Tygert suggested that students would be better served if they were taught applied mathematics instead of traditional mathematics. Antonio Ortega, University of Southern California, highlighted the tension that exists in classrooms between mathematical concepts/methods and open-ended exploration. He wondered if it is possible to develop a more flexible educational model that allows more time for the latter. Patrick Perry, New York University, interjected that learning to use the tools and methods is essential to solve problems, but he agreed that there should be more room for experiential curricula. Gross noted that not all students follow similar career paths, so it is difficult to assess success in data science. Constantine Gatsonis, Brown University, mentioned the importance of extendable skills as the debate continues about whether data science is a discipline or a profession.

James Frew, University of California, Santa Barbara, noted the importance of distinguishing between repositories and resilient databases. Elaborating on this point, Hero explained that an increased exposure of

public data repositories emphasizes the need to develop standards, benchmarks, and principles for encoding databases to lessen misuse. David Rabinowitz asked whether there are tools that can serve unsophisticated users. Hero noted that the use of tools without a sufficient understanding of the data, underlying mechanisms, and limitations is risky. However, there is a need for a dashboard to navigate a suite of software tools so that sophisticated users can use tools with more authority. Steven Miller, IBM, talked about the difference between "human data scientists" and "machine data scientists" and suggested that the depth of computer science training required is less for human data scientists than for machine data scientists. Because of this distinction, he noted that applied data science programs have become more popular at undergraduate institutions across the country. Howe agreed that this is an important distinction, and he discussed the "transcriptable options" that are available at the University of Washington. For example, students can add a specialization in data science to their core major, which will appear on their transcript, thus making them more marketable when applying for jobs.

Gatsonis posed a question to the group: Do businesses prefer hiring one individual with all relevant skills or hiring a team of individuals, each with a unique skill? Mark Krzysko, U.S. Department of Defense, noted the difficulty of finding the "perfect" employee and emphasized the importance of a person's ability to communicate across disciplines and solve problems. Abowd suggested that the rules-driven approaches used by many large human resources organizations would benefit from incorporating particular data science tools into their hiring processes. Michelle Dunn, National Institutes of Health (NIH), noted that hiring is a concern across all government agencies, and there are currently teams in place developing better strategies for hiring data scientists.

Frew reminded participants to think about data science applications in a cross-disciplinary light. Isabel Cárdenas-Navia, Business-Higher Education Forum, asked participants to consider the importance of the liberal arts in the discussion of a data science curriculum (e.g., a liberal arts degree with a concentration in data science could prove valuable to hiring organizations), and Rebecca Nugent, Carnegie Mellon University, suggested that data science outreach efforts be directed toward humanities students. Hero mentioned that offering certificate programs tends to draw students from more diverse disciplines, but he also noted that student demand for data science courses is never an issue; what stifles enrollment is limited available faculty and course offerings. In support of additional cross-disciplinary efforts, Kolaczyk reiterated the importance of statistics students developing relationships with people in the disciplines with real problems that can be explored. Horton added that gender balance and diversity need to be considered when developing new curricula.

## NEEDS OF DATA SCIENCE COMMUNITIES

### Biomedical Research

*Michelle Dunn, National Institutes of Health*

As data science becomes crucial for biomedical research, five trends and related challenges have emerged in biomedical science:

1. Biomedical data science has been accepted as a field of study and departments have been created at institutions across the country, but there is a lack of clarity about its niche.
2. Biomedical data science training programs have been created with the help of Big Data to Knowledge (BD2K) funding, but a discussion about the core competencies of these programs is needed (e.g., almost all programs have courses in probability and statistics, while few have courses in reproducibility).
3. Data science has been deemed integral to biomedical research, so the next step is to identify and adopt best practices.
4. Demand for data science training among biomedical scientists continues to grow, and more massive open online courses (MOOCs) and short courses should be integrated into training programs.
5. Data science has increased visibility and impact at NIH—increased funding for data science exists, but continued leadership and integration is needed within NIH Institutes and Centers.

Lida Beninson, National Academies, noted that for those who are hired for R1 positions, the average age at which that first happens is 42. Because of this, it is crucial to ensure that training programs for the next generation of researchers include highly transferable skills. Dunn agreed that transferable skills are important, but she also hopes that those who want to stay in academia can do so and that some of NIH's initiatives will help lower the age of entry into academic careers. Jeffrey Ullman, Stanford University, asked whether it is feasible and desirable to align the curricula of bioinformatics and biostatistics in biomedical data science. Dunn responded that some alignment would be helpful, but this is also a matter of scale. She continued that programs should always have diverse offerings so that students can choose what will work best for them as individuals.

Cárdenas-Navia asked whether NIH targets any of its programs to undergraduates so that they get a sense of how data science is integral to the field and overcome "math phobia." Dunn noted that NIH has already spent approximately $1 million on K-12 initiatives and hopes to

fund programs at the undergraduate level as well. Gross noted that the attitude toward quantitative ideas has changed over the past 20 years and highlighted the importance of every member of a team having an understanding of quantitative ideas. Dunn added that although data science courses in biomedical programs provide the language to communicate with teammates, they do not provide the breadth for expertise. Nina Mishra, Amazon, offered the idea of a data science minor, and Dunn agreed that this possibility should be explored.

## Industry

*Nina Mishra, Amazon*

Mishra noted that students want to have solid foundations, to develop business acumen, to understand the nuances of data, and to be able to scrutinize experiments. She noted that data science has no clear definition, and she wondered whether the job category "data scientist" is one that will endure for decades. Ultimately, students are in need of a strong foundational understanding of probability, statistics, algorithms, linear algebra, and machine learning, and they need better critical scientific thinking and problem-solving skills to have long-term success in the workforce. Students need to learn how to frame a business problem before integrating their knowledge of data and algorithms, and they need to learn how to use data to make an argument. Students also need to understand bias in data, to question experimental results, and to know what tools do instead of just how to work them. Students would benefit from internships and mentorships in order to build better business acumen. Communities, on the other hand, want public data repositories and analytics, as well as ways to compare and rank data science programs.

Miller said that his preference would be for all new hires to be data literate. He highlighted the importance of individual institutions targeting different skills; it will not be useful to hiring organizations if all schools offer the exact same programs. Ortega agreed that the fundamentals still matter. He cautioned industry from continuing to send students the message that programming is the only important skill. Mishra noted that although programming is important to hiring groups at Amazon, many other skills and qualities are also valued. Ron Brachman, Cornell University, reiterated that data scientists are different from data engineers and that it is important to discuss varied career paths for students. Although everyone should be data literate, he does not see the value of having everyone enroll in data science programs. Stodden said that it might make sense to introduce the whole life cycle of data science in an introductory college course in order to draw greater appeal and understanding from students.

Howe cautioned against ranking data science programs; instead, he suggested that hiring organizations do research about candidates' institutional offerings prior to the interview to help determine the level of the candidate's preparedness. Gross asked whether "business acumen" differs from "data acumen." Krzysko said that "business acumen" extends beyond what is happening at universities because it relates to solving real problems. Perry explained that the survey of his colleagues' interests was similar to those of Mishra: domain experts teach data-intensive courses focused on problems, not methods. Christopher Malone, Winona State University, asked whether the agencies hire people with undergraduate degrees in data science. Krzysko said that acquisition capabilities developed in a graduate or doctoral program are often more desirable, but Abowd confirmed that agencies do hire people with undergraduate degrees.

### Government

*John Abowd, U.S. Census Bureau*

Abowd said that students need to develop four skills: designed data methodology, statistical/machine learning, hierarchical modeling, and curation and reproducibility. He noted that designed data are not the same as survey data and that although everything a statistical agency does should have a design, the data need not be from a survey. He also noted that inference is not just a prediction.

In the past, employee training at the U.S. Census Bureau, for example, involved a joint program in survey methodology, but now there is a need for data analysts to have expanded competencies. At the graduate and doctoral levels, there should be intense exposure to or an actual degree in a content area, such as economics or biostatistics, and every Ph.D. should have exposure to data science. The substantial increase in computing capacity required in government agencies can be difficult to manage. Data scientists can assist with both data management and infrastructure. Krzysko added that his group oversees a $1.7 trillion portfolio and, while infrastructure exists, questions remain about how to frame and guide those who need to deploy the infrastructure as well as how to look at the data and identify organizational/process applications. Krzysko reiterated that problem solving is the most important skill desired in employees.

### Open Discussion

Chris Mentzel, Gordon and Betty Moore Foundation, noted that the definition of data science, and whether or not it constitutes a discipline, still has not been formalized. He suggested keeping the definition flexible.

Rabinowitz noted that data science is a set of tools that will be universally applied; it does not need to be a separate discipline. Miller highlighted the challenge of building data "literacy" without defining specialties, and he also highlighted the importance of accreditation in any curricular discussions. Gatsonis suggested that the roundtable continue to discuss ways to teach data science both as a primary subject and as a concentration area.

In a discussion about the comparisons of operations research to data science, Mentzel noted that the pervasive application space for data science did not exist for operations research. Malone cautioned of the dangers in combining computer science and statistics and calling it data science. He also suggested that the roundtable pay particular attention to smaller colleges in its future discussions about data science programs, as well as to the expectations for graduates. Cárdenas-Navia reiterated the importance of attracting a diverse audience of students through careful course design and attentive advising.

# 3

# Meeting #2: Examining the Intersection of Domain Expertise and Data Science

The second Roundtable on Data Science Postsecondary Education met on March 20, 2017, at the Arnold and Mabel Beckman Center of the National Academies of Sciences, Engineering, and Medicine 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.

## EMERGING NEEDS AND OPPORTUNITIES 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 are 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 good questions and who answers these questions better and faster than her colleagues. This success, 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 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 catalogue 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 3.1) 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.


**FIGURE 3.1** Expecting postsecondary 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.

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 instead of attempting to train a person to develop both. 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, Columbia University*

Connelly explained that the social sciences are structured differently from the physical sciences, and this difference 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 co-authorship 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 subfield 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 whether 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 nonquantitative 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 reevaluating 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, because 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 that 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 3.2).

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 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.

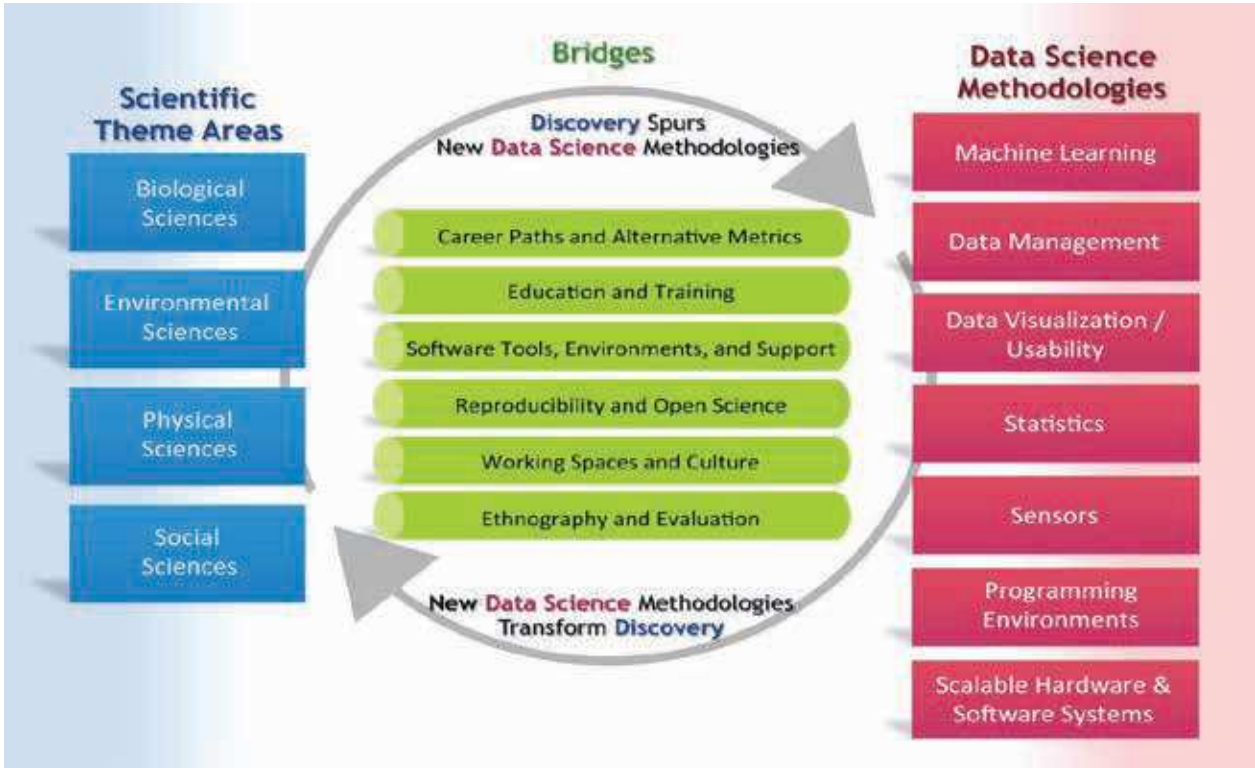


**FIGURE 3.2** The eScience Institute's data science cycle includes education and training as a means to connect domain science inquiries to methodological developments. SOURCE: Produced by Ed Lazowska, University of Washington, and Moore/Sloan Data Science Environments and presented by Bill Howe, University of Washington, to the roundtable.

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 four 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 define and are 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 nonscientific 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, such as Data: Past, Present, and Future, have been funded by the Collaboratory. This undergraduate course is taught by a historian and an applied mathematician, and it contains a core of knowledge that emphasizes 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 whether 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 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 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 want 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 3.3), 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 pedagogy 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

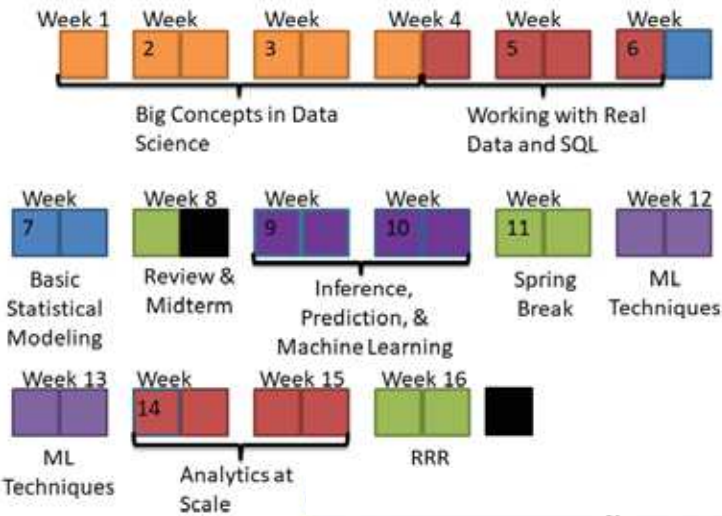


**FIGURE 3.3** The syllabus for a pilot Data Science 100 course, inspired by the data science life cycle. SOURCE: Produced by Professor Joseph Gonzalez, University of California, Berkeley, and presented by Cathryn Carson, University of California, Berkeley, to the roundtable.

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 whether 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, but 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 titled "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.

# 4

# Meeting #3: Data Science Education in the Workplace

The third Roundtable on Data Science Postsecondary Education met on May 1, 2017, at the Pew Research Center in Washington, D.C. Stakeholders from data science education programs, funding organizations, government agencies, professional societies, foundations, and industry convened to discuss data science training in the workplace. 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.

## DATA SCIENCE TRAINING IN THE WORKPLACE: GOVERNMENT

### Practicing Data Science in the Government

*Ron Prevost, U.S. Census Bureau*

Prevost explained that data produced by the U.S. Census Bureau are expected to be unbiased, statistically accurate, delivered quickly at low cost, useful to determine causality, reproducible, transparent, and protected. While striving to meet these expectations, statistical agencies confront many challenges, including greater than expected costs and lower than expected response rates for surveys, complex information requests, competition among data products and questions of product validity, new data sources and methodologies, and policy requirements.

The Census Bureau hopes to supplement survey data with data that have been repurposed from other sources. However, this data integration needs to be transparent and reliable, utilize quality measures, and ideally incorporate model-based estimation and data source acquisition and integration processes. In his view, the Census Bureau could advance this paradigm shift by taking several critical steps, including (1) consolidating business processes and systems and generalized solutions, (2) supplementing current business processes with new processes, (3) developing new products, (4) building new capabilities, and (5) optimizing current business processes. Prevost noted that institutional, budgetary, and security barriers can limit this type of large-scale transformation. For example, to address this transformation thoroughly, staff in information technology departments would have to learn new processes for managing, curating, and using data and metadata; to ensure that new software complies with government security protocols; and to organize, explore, and test real data in a collaborative environment often referred to as a "sandbox."

Many federal agencies are also exploring how increased opportunities for interdisciplinary teamwork and professional development could better equip employees for work that requires new computing techniques and new methodologies. The Census Bureau is evaluating program use cases to determine which skill sets will be needed by both current and future employees, as well as how projects will be funded. According to Prevost, current knowledge gaps include data science, business/data analytics, reproducibility, software design and engineering, data storage and retrieval models, and operations research. After extensive investigation, the Census Bureau created a catalogue of 80 programs (with a total of 600 courses) located in or near Washington, D.C., or available online, that offer degrees, certificates, or short courses in the needed content areas. Massive open online courses (MOOCs) may be a cost-effective alternative or complement to these more traditional training programs because they address specific agency needs quickly and flexibly.

Nicholas Horton, Amherst College, emphasized the need to train employees to understand the unique advantages and disadvantages of using different types of data in their work and encouraged the Census Bureau's emerging emphasis on fusion of found and designed survey data. He noted that issues of data ethics and cybersecurity are crucial areas for employee training. Jeffrey Ullman, Stanford University, suggested that there is a disconnect between training and real-world problems that could be eliminated with further development of core computer science skills. Prevost agreed that skills gaps exist but noted that the program use cases explored thus far were focused more on training for research analysts than for information technology specialists.

Patrick Perry, New York University, asked for clarification on what is driving the transition to model-based estimation and inference and whether new training is necessary to apply this type of methodology. Prevost responded that the Census Bureau sought new approaches to improve legacy products, given declining survey response rates and questions about bias in these products. He added that while the Census Bureau does provide training in big data and statistics to its employees, much of the current training is in project and budget management. Jordan Sellers, Howard University, suggested that the Census Bureau take the lead in establishing a professional development policy; however, Prevost noted that a formal, standardized training policy may not be effective because staff training evolves around mission-critical activities and rapidly changing technologies. Victoria Stodden, University of Illinois, Urbana-Champaign, asked how people can track the provenance of Census Bureau data sets. Prevost stated that all Census Bureau data products undergo numerous quality measurements related to collection methods, variance, and benchmarks. To learn more about Census Bureau data products, and how they compare to other data products, he recommended that researchers visit the Federal Statistical Research Data Centers[1] located throughout the United States.

### Training Government Employees in Data Science

*Drew Zachary, U.S. Department of Commerce*

Zachary noted that developing creative training initiatives is essential for federal agency managers who have limited funding or authority to offer education programs or hire new staff. When evaluating how to bring together the right set of data science skills, two models are useful for employees and managers to consider: (1) a "unicorn" model, in which one employee has all of the skills needed to complete a task; or (2) an "X-Men" model, in which people with diverse skills work together to complete a task.

The Commerce Data Academy[2] is an internal upskilling data science education initiative that relies on the Commerce Data Service, as well as extra-governmental instructors from organizations including General Assembly and Data Society, to train Department of Commerce colleagues in data science, data engineering, and web development skills. After training more than 1,500 Department of Commerce employees in 35 courses

[1] The website for the Federal Statistical Research Data Centers is https://www.census.gov/fsrdc, accessed February 13, 2020.

[2] The website for the Commerce Data Academy is https://www.commerce.gov/page/commerce-data-academy/, accessed February 13, 2020.

(both online and in-person) over the past 1.5 years, the Commerce Data Academy now invites employees from other federal agencies to enroll in its courses on an as-needed basis.

Initiated by the Office of Science and Technology Policy during the Obama Administration, Fellows in Innovation[3] reaches programs across government, representing 400 fellows and 30 agency divisions. Zachary noted that this program allows data professionals to apply their often underutilized technical skills to a policy problem, as well as to transfer these data science skills to their teammates. For example, a team used machine learning, digital mapping, and sentiment analysis to help understand neighborhood data and explore opportunities for economic development in high-poverty communities.

Supported by the General Services Administration, the Federal Data Cabinet creates a "community of practice" for data professionals in government to share best practices and success stories, as well as to discuss challenges faced throughout the data life cycle. One of the working groups within the Federal Data Cabinet, the Data Talent Working Group, plans to create a decision guide to help hiring managers and team leaders assemble teams and choose effective training models to best meet project needs.

Natassja Linzau, National Academies (formerly of the Department of Commerce), emphasized that all of the "teachers" in the Commerce Data Academy are sharing their time and expertise without additional compensation, and the "students" do not pay any fees to take their courses. Ullman wondered whether MOOCs could be used in the Commerce Data Academy in the future and whether there may be introductory courses that could be added to the list of offerings. Zachary and Linzau noted that many of their course materials and recordings are available on the Commerce Data Academy website so that anyone who is interested can use them to learn. They also plan to explore using MOOCs as a way to enhance future course offerings.

Louis Gross, University of Tennessee, Knoxville, wondered how a decision is made regarding whether to train employees or to hire consultants to solve particular problems. He suggested that agencies learn how to better use their talent pools, highlighting the Fellows in Innovation program as a good model, and Zachary noted that the Federal Data Cabinet could also serve as a repository for this information. Prevost added that mentorship programs could also be expanded to address this issue. In response to a question from Rebecca Nugent, Carnegie Mellon University, Zachary noted that sustained funding of the program is a concern, as is relating the benefits of the program to fellows' supervisors.

[3] The website for Fellows in Innovation is https://fellows-in-innovation.pif.gov/, accessed February 13, 2020.

## DATA SCIENCE TRAINING IN THE WORKPLACE: BUSINESS

### The Technology Sector

*Emily Plachy, IBM*

Plachy defined data scientists as "pioneers" who solve problems by relying on quantitative training, effective communication skills, business acumen, and various data and analytics tools and programming languages. She noted that data science continues to evolve in response to the era of cognitive computing. Data scientists now need skills in hybrid analytics, streaming data, artificial intelligence, application program interface–based analytical services, and cloud-based solutions. Data scientists often expect their employers to help them build upon their technical and business skills to keep pace with the evolving field, so Plachy suggested that it may be useful for employers to establish a certification roadmap. Offering workplace data science training not only improves employee performance but also may increase employee retention, according to Plachy. IBM created a Data Science Profession to encourage data scientists to continue to train and develop their skills; it uses "open badges" that contain metadata representing "skill tags" and accomplishments, both to signal and verify employees' skills and to improve social connections among colleagues.

Data science education opportunities for IBM employees include the following:

- *Data Science Bootcamp*—New data science employees can develop awareness of various data science concepts and form networks with other practitioners over 8 days.
- *Data Science Experience*—Scientists collaborate in sandboxes, using data analytics to solve problems.
- *Big Data and Analytics University*—Participants enroll in virtual data science courses at one of three expertise levels.
- *Analytics Product Course*—Short courses provide overviews of available IBM products.
- *Development Activities*—Employees select topics for monthly fundamentals courses.
- *Analytics Education Series*—Employees select from more than 30 1-hour videos of IBM expert lectures on topics such as natural language processing or spatiotemporal analytics.
- *Cognitive Academy*—Data scientists receive training in areas such as data visualization or machine learning.

- *Analytics Across the Enterprise: How IBM Realizes Business Value from Big Data and Analytics*—Textbook includes 32 case studies of problems solved using data analytics.

Plachy described the Chief Analytics Office, IBM's version of the "X-Men" model introduced by Zachary, in which 50-75 people with diverse technical skills form teams to solve business problems within IBM. She said that many recent hires at IBM have knowledge gaps in cognitive computing and the skills to better harness unstructured data to solve business problems; they could benefit from stronger quantitative foundations, better communication skills, and more curiosity and patience. Because data science will continue to evolve, it is unlikely that the conversation about knowledge gaps in data science education will ever end, and IBM may add apprenticeship programs in the future. Plachy suggested that it would be helpful if stakeholders created a public education system for data science where organizations could share ideas for workplace training.

Kristin Tolle, Microsoft, added that, in her organization, experimental design is a major knowledge gap among recent hires. Plachy agreed with Tolle about the importance of training in that area and noted that IBM hires experimental physicists to help colleagues with experimental design and also teaches design of experiments in a Six Sigma course. In response to a question from Ullman about gaps in current computer science degree programs, Plachy responded that she would like to see more preparation in artificial intelligence, deep learning, and natural language processing.

### The Consulting Perspective

*Ashley Lanier and Ashley Campana, Booz Allen Hamilton*

Lanier and Campana noted that the need to fill knowledge gaps in employee education is not a problem unique to the field of data science. At Booz Allen Hamilton, while employees without data science training need to learn how to use tools efficiently and to analyze and share data, employees with data science specialties need to learn "consulting skills" such as communicating, storytelling, working with clients, working in a team, understanding an audience, and choosing the right approaches.

Because there are unique infrastructure constraints in upskilling employees in consulting firms, Booz Allen Hamilton offers a variety of education programs to its employees, all of which include essential training in teamwork and presentation skills:

- *Data Science Bowl*—Approximately 2,000 teams from around the world participate in this 90-day online hackathon for social good.

- *Tech Tank*—Similar to a master's certificate program, with a math and a computer science track, 160 hours of training over 12 months are offered to employees with a scientific background and within 2 years of hire, upon nomination from a supervisor. In addition to technical training, participants receive training (based on personality test results) in communication skills and mentoring. Participants pitch to leaders acting as clients and work on a real problem during an apprenticeship.
- *Internship Program ("Summer Games")*—Approximately 300 undergraduate interns work on STEM-focused problems and pitch to Booz Allen Hamilton leadership over a 9-week session.
- *Data Science 5K Challenge*—Similar to Tech Tank, except that training is delivered by an external vendor instead of by Booz Allen Hamilton leadership. This allows more people to participate at the right level and helps the company to increase the total number of data scientists on staff.

Booz Allen Hamilton also offers a data science book club, Yammer groups, bi-monthly Hackathons, a distinguished speaker series, occasional boot camps, and a workshop series as additional, flexible ways for employees to become more engaged in data science. In response to a question from Stodden about additional data science problems that Booz Allen Hamilton interns and employees have helped clients to better understand or solve, Lanier and Campana highlighted the following projects: (1) applying analytics to cardiology to assess heart function, (2) using data analytics to increase adoption rates at animal shelters, (3) employing network analysis to better understand human trafficking in the United States, and (4) using data analytics and technology to help houses go off the electric grid. Gross asked whether the education programs at Booz Allen Hamilton have been formally assessed and whether those results have been published. Booz Allen Hamilton tracks billable hours, promotion, and retention of its Tech Tank participants to demonstrate the program's value, but that information is not shared externally. These assessments have also revealed that participants are more incentivized by the opportunity to make a difference solving real problems using real data sets than by the opportunity to earn social media "badges" or prize points for their work. Gross suggested that Booz Allen Hamilton publish future assessment results, as doing so could aid the larger data science community in its development of training.

Nugent encouraged increased collaboration between companies and universities, especially in terms of student skill assessments, so that companies are hiring the best-suited employees. In response to a question from Deborah Nolan, University of California, Berkeley, about the timeline for

skill cultivation, Lanier noted that new hires start developing communication, leadership, and presentation skills immediately. Doing so also helps determine with which projects new employees should be aligned. In response to a question from Ullman about gaps in current computer science degree programs, Lanier responded that she would like to see more preparation in machine learning and presentation skills. William Finzer, Concord Consortium, asked whether the emerging field of data science education research could address challenges in employee training. Lanier noted that Booz Allen Hamilton currently utilizes research collaboration sessions, rapid innovation workshops, and design thinking exercises to facilitate internal problem solving.

## DATA SCIENCE TRAINING IN THE WORKPLACE: EXECUTIVE EDUCATION

### Executive Education Online

*Brian Caffo, Johns Hopkins University*

Caffo described Johns Hopkins' Data Science Specialization,[4] delivered via Coursera, which includes the following courses: The Data Scientist's Toolbox, R Programming, Getting and Cleaning Data, Exploratory Data Analysis, Reproducible Research, Statistical Inference, Regression Models, Practical Machine Learning, Developing Data Products, and a Capstone Project done in collaboration with industry.

He explained that the program is unique in that it attempts to offer a complete data science curriculum through a large amount of bundled content; it provides all course notes on GitHub in R markdown and uses R almost exclusively; it utilizes Statistics with Interactive R Learning (Swirl[5]); it allows free course textbook downloads via Leanpub; and it offers a LinkedIn space for alumni to connect upon completion.

Because Caffo and his colleagues found that industry managers often have fewer technical skills than their junior-level employees, they realized the urgent need for a specific training program to equip executives with the right skills to manage their teams. Johns Hopkins adapted the Data Science Specialization to create the Executive Data Science Specialization, which provides an overview of data science management. The Executive Data Science Coursera curriculum includes four content courses designed to be completed in only 1 week each:

[4] The website for the Johns Hopkins Data Science Specialization is https://ep.jhu.edu/programs-and-courses/programs/data-science, accessed February 13, 2020.

[5] The website for Swirl is https://swirlstats.com/, accessed February 13, 2020.

1. *A Crash Course in Data Science*—High-level overview of statistics by example, machine learning, software engineering for data science, outputs of data science experiments, definitions of success, and the data science toolbox.
2. *Building a Data Science Team*—Overview of differences between types of data scientists and data engineers and how they can work together effectively.
3. *Managing Data Analysis*—Overview of types of questions asked by data scientists, qualities that make a sound question, exploratory analyses, inference, prediction, interpretation, modeling, and communication.
4. *Data Science in Real Life*—Ideal goals for data analysis, including clean data pulls, carefully designed experiments, and clear results, and strategies for when decisions are unclear or data products are ineffective.

Similar to the original program, the executive program emphasizes active learning and offers a Capstone Project (in partnership with Zillow and incorporating Swirl) upon completion of the coursework. Over the past year, 2,020 people completed the Capstone Project in the executive program, with 99 percent awarding it positive ratings.

Ullman asked whether the courses cover explainability of models, and Caffo responded that they circle around the topic of explainability by discussing knowledge creation, simple models, parsimony, and interpretability. In response to a question from Perry about the student demographics in the executive program, Caffo noted that the content is designed specifically with managers in mind; however, he cannot confirm whether managers are actually enrolling. Kathleen McKeown, Columbia University, inquired about the cost of the executive program, and Caffo noted that although the course videos and materials can be viewed for free, students have to pay to receive the certification upon completion. Mary Moynihan, Cape Cod Community College, mentioned that high-cost, for-credit online courses typically have only a 30 percent completion rate and wondered whether this is the best way to train people in data science. Because completion rate is not necessarily an accurate indicator of engagement and learning in free or low-cost online courses and MOOCs, Caffo suggested that these programs may need to be evaluated differently from high-cost online courses.

Horton suggested that professional development is needed for faculty who wish to deliver online course content effectively. Prevost added that there also needs to be an incentive for an employee to complete an executive course, whether it be a component of a performance review or a monetary award. Horton posed a related question: How do we encourage

people who do not have any incentive for further coursework? He noted that community colleges could play a role in training because of their low-cost, flexible offerings.

### Executive Education in Business Schools

*Claudia Perlich, Dstillery and New York University*

Perlich described a course she offers at New York University titled Data Mining for Business Intelligence that offers two tracks for master's in business administration (MBA) students: the technical and the managerial. The technical track is offered in collaboration with the Center for Data Science and the computer science department and is taught solely in Python, while the managerial track often enrolls students without any programming skills but who wish to learn how to manage data science. This course introduces data science (1) terminology; (2) methods (e.g., supervised and unsupervised learning, model evaluation, data processing); (3) applications (e.g., case studies, Weka[6]); and (4) management (e.g., deployment, hiring, interviewing, and proposal evaluation). This content is delivered via weekly lectures, guest speakers, homework assignments, a final exam, and a final team project. The course project requires students to identify a problem, find data, solve the problem, demonstrate business value, submit a written report, and present to the class—all of which are important data science skills (Figure 4.1).

In Perlich's view, students often have difficulty recognizing a predictive modeling problem, understanding the value of good baselines, translating a model into action, using precise language, and budgeting time for data preparation. However, upon completion of the managerial track, students are expected to be able to do the following:

- Approach business problems thoughtfully using data analytics to improve performance and know how to hire a data scientist;
- Understand that data preparation takes time but is necessary;
- Recognize that not all problems are data science problems;
- Think backward from a problem, not forward from the data;
- Know the basics of data mining processes, algorithms, and systems; and
- Have hands-on experience with mining data.

[6] The website for Weka is https://www.cs.waikato.ac.nz/ml/weka/, accessed February 13, 2020.

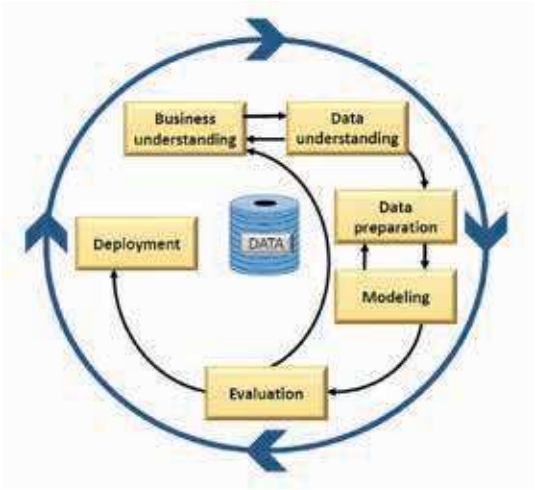


**FIGURE 4.1** The New York University course Data Mining for Business Intelligence follows an iterative data science process that emphasizes the formulation of a problem that can be addressed through data. SOURCE: Kenneth Jensen, 2012, "File:CRISP-DM Process Diagram.png," courtesy of CC-BY-SA-3.0.

In response to a question from Jessica Utts, University of California, Irvine, about helping students gain skills desired by employers, Perlich noted that it is incredibly difficult but necessary to teach communication skills and teamwork. John Abowd, U.S. Census Bureau, expressed concern about offering two separate tracks for the course, since the managerial-track students may not receive the same critical assessment experience as the technical-track students. Perlich responded that it would be difficult to cater to two audiences if faculty delivered this content via a single course, and she added that even if the tracks did not exist, students would likely self-select a course that best meets their knowledge and needs based on the syllabus content. While she sees value in offering a course without programming, she shares the concern about an overall decrease in technical content in data science curricula. In response to a question from Ullman about the course's attention to explainability of models, Perlich acknowledged that although both tracks discuss this topic, she is unconvinced that such discussions of transparency and explainability truly address issues of fairness and bias in data science.

## OPEN DISCUSSION

### Funding and Scaling Innovative Data Science Education in the U.S. Government

In response to a question from Alok Choudhary, Northwestern University, about data science education within the government, McKeown noted that training and retention are particularly important when hiring is constrained. Mark Krzysko, Department of Defense, said that government employees have unique challenges to becoming data literate, sharing data, and communicating across departments. He reiterated that the government has to work with what it has without overusing skilled employees. Krzysko noted that the government would benefit from an authoritative source of information on how data science can be used to help solve problems. Antonio Ortega, University of Southern California, highlighted USAFacts.org as an open-access repository that collects data from local, state, and federal government. He suggested that this be used as an entry point to address government data challenges. Horton suggested an independent statistical system that maintains high-quality data used to make better decisions in a nonpartisan way.

Ullman noted that cost should not be considered a barrier to education given the accessibility to MOOC video content; using such material, faculty can create short courses at low cost. Zachary responded that while data training can be free, it is still completely inaccessible to many, and continuing to offer training only to those with access broadens inequalities in our communities. Krzysko and Abowd observed that hiring opportunities in the public sector are more limited than in the private sector and commended Zachary's creative efforts in addressing training challenges for the Department of Commerce workforce. Prevost suggested that organizations ask themselves what will be needed to upskill core employees, as well as those around them, and develop a "product-training-process" cycle that can be implemented whenever a new problem related to staff training surfaces. Abowd remarked that, when it is possible to hire new employees with different skill sets, the government needs help creating job descriptions that attract appropriate candidates. McKeown encouraged organizations in the public sector to weigh the benefits and drawbacks of hiring new employees versus upskilling current employees and added that recruiting and retaining individuals in government jobs that pay less than industry jobs can be challenging. Abowd mentioned that interns could meet specific data science needs, albeit with short-term availability.

Stodden highlighted the potential role of OpenGov, which leads the government transparency movement, in solving data science problems or in helping to build pipelines for data scientists to do public service for

government agencies. She also suggested forming a group for data scientists, modeled after the Peace Corps or Teach for America, in which they can help communities or organizations solve large problems.

### Using Sandboxes Across Organizations to Better Facilitate Progress

David Levermore, University of Maryland, College Park, said that lack of access to data is problematic for faculty trying to create hands-on projects for students. He noted how helpful it would be if industry and government made their data available to universities for student coursework. Caffo explained that simply granting open access for faculty to use data in their classes is insufficient; faculty have to analyze and prepare the data first to align it with their curricular goals, which is a time-intensive process. Plachy remarked that some small IBM data sets are shared in a sandbox for public use, but Laura Haas, IBM, responded that data licensing can be challenging; most companies do not want to risk legal action for accidentally releasing copyrighted data, which creates a substantial barrier to sharing data with universities. She posited that government agencies may face similar obstacles to data sharing because some data assumed to be open could actually include copyrighted material. Abowd suggested that faculty check the Census Bureau data application program interface[7] for data they could freely use within the classroom.

Eric Kolaczyk, Boston University, explained that sandboxes spanning multiple organizations offer a holistic experience of working iteratively with experts in a team; the use of data repositories alone is inadequate. Creating successful sandbox experiences can be expensive, require energy and time, and depend on established relationships among stakeholders. Kolaczyk also wondered about the possibility of scaling sandbox experiences. Plachy referred to IBM's free beta version of the Data Science Experience because it provides teams a place to store data and collaborate. Kolaczyk suggested that this would be an even more useful platform if users had access to IBM data and IBM team members.

Abowd noted that the General Services Administration tried to execute sandboxes with GovCloud, but satisfying the agency-specific security requirements and completing the associated paperwork created implementation challenges. Kolaczyk reiterated that sandboxes are most useful when users have access to people, not just data. Abowd noted that government sandboxes will remain accessible to government employees for the time being, but he would like to see multi-organizational sandboxes offered in the future. Krzysko added that operational rules

[7] The website for the Census Bureau data application program interface is https://www.census.gov/data/developers/data-sets.html, accessed February 13, 2020.

and infrastructure do not yet exist in the government to support such an endeavor; however, a recent pilot program giving federally funded research and development centers access to data and a dissemination guide indicates good progress.

### Bridging Gaps in Knowledge and Perspectives Through Teamwork and Communication

Stodden wondered what makes communication skills for data scientists unique. Patrick Riley, Google, responded that while people working in traditional technical fields talk predominantly with other technical people, data scientists need to be able to explain difficult concepts to nontechnical audiences. Perlich agreed that students have to learn to frame problems clearly for nontechnical audiences. Riley suggested that students would benefit from practice exercises in which they have to present summaries of analyses to varied audiences. Levermore reiterated that the need for strong communication is not a new phenomenon; he suggested looking to the past, when computational sciences was a new field, and expanding those ideas to fit the even larger data science revolution. Gross said that methods for how to communicate scientific ideas to others could be integrated into any data science curriculum. He also suggested that educators focus on creating teams of varied backgrounds and perspectives, not just diverse knowledge levels. He pointed to educational approaches that can help reduce unconscious bias and teach others to speak effectively with one another, both of which are useful skills for teams composed of technical and nontechnical members. Andrew Zieffler, University of Minnesota, cautioned that definitions of "teamwork" and recommendations for team sizes vary in the literature across disciplines, institutions, and organizations and need to be researched carefully by faculty designing curricula. A university that teaches broad teamwork skills best prepares students for diverse work environments, and Zieffler explained that one way to do this is to give students problems that are impossible to solve individually. Choudhary added that it is important to involve students in experiential learning and to bring technical and nontechnical people together to define and refine problems. McKeown noted the value of exposing students to the unique vocabulary and approaches in varied disciplines so as to prepare them to work more cooperatively in interdisciplinary teams. David Culler, University of California, Berkeley, added that liberal arts skills (e.g., critical thinking, abstraction) aid in developing better data scientists.

Horton referenced a software engineering course at the University of California, Berkeley, as a model of teaching cross-disciplinary teamwork in which students used technology to solve important problems

for non-profit organizations. Culler believes that current students often want to be producers of knowledge instead of consumers of knowledge; they just need the right tools and experiences to make a difference. Perlich added that while there is no shortage of good will, there is a crucial lack of project management, especially in volunteer programs attracting data scientists. She thinks that a model to ensure that people with the right skill sets are brought together and that volunteers are doing work related to their areas of expertise is needed. Catherine Cramer, Hall of Science, discussed the early intervention program "Big Data for Little Kids," which works with young children from immigrant families to improve access to STEM education. Noting the value of community partnerships, Zachary added that it is challenging to translate technical capacity to a specific need. She noted that inequalities may continue to grow if data scientists do not engage with the community's problems.

# 5

# Meeting #4: Alternative Mechanisms for Data Science Education

The fourth Roundtable on Data Science Postsecondary Education met on October 20, 2017, at Northwestern University in Evanston, Illinois. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss alternative mechanisms in data science education. 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.

## STANFORD UNIVERSITY'S CERTIFICATE PROGRAMS

*Jeffrey Ullman, Stanford University*

Ullman shared the history of Stanford University's professional certificate programs. In the 1960s, the School of Engineering broadcast recorded lectures through the Stanford Instructional Television Network (SITN), and couriers delivered lecture notes to and collected homework assignments from local industry participants. Employers paid twice the tuition rate for their employees to complete a master's of science in engineering through the SITN. The SITN eventually became the Stanford Center for Professional Development (SCPD),[1] which now offers a vari-

[1] The website for the Stanford Center for Professional Development is http://scpd.stanford.edu/home, accessed February 13, 2020.

ety of courses and certificates worldwide via the internet. Participants do not have to apply to or enroll in the university to participate in SCPD programs.

He noted that although a graduate certificate is not equivalent to a diploma, it does hold more weight than a statement of completion because all certificate coursework is graded. The statistics department introduced the Data Mining and Applications certificate (three courses) in 2009, and the computer science department followed in 2010 with the Mining of Massive Data Sets certificate (four courses). Ullman emphasized the initial popularity of both programs but noted a decrease in enrollment since 2013, likely owing to the availability of more certificate programs in other disciplines of interest (e.g., artificial intelligence, cybersecurity). However, between 2009 and December 2017, there was a 50 percent increase in the total number of graduate certificates awarded across Stanford University's departments.

Ullman turned to a discussion of two of Stanford's approaches to data science. Although the computer science department does not offer a data science degree, students can complete a data science specialization at both the undergraduate and graduate levels. In conjunction with the Institute for Computational and Mathematical Engineering, the statistics department offers a master's of science in statistics: data science. Another difference, according to Ullman, is that computer scientists utilize algorithms to solve problems, while statisticians validate the soundness of solutions. He challenged Drew Conway's Data Science Venn Diagram (Conway, 2010), noting that it fails to acknowledge the value of computer science's understanding and implementation of algorithms, and he displayed his own version of the Venn diagram that removes mathematics and statistics from the core of data science.

Ron Brachman, Cornell Tech, asked whether matriculated Stanford graduate students are eligible to participate in certificate programs. Ullman noted that while it is possible, students are prohibited from cross-counting courses. Victoria Stodden, University of Illinois, Urbana-Champaign, asked Ullman about the role of university administration in sustaining the certificate model. He explained that individual faculty members propose content for certificate courses and emphasized that curriculum change occurs via bottom-up approaches. Challenging Ullman's version of the Venn diagram, Kathy McKeown, Columbia University, emphasized not only how much computer science and statistics overlap but also how important statistics and mathematics are to the study and practice of data science.

## BOSTON UNIVERSITY'S STATISTICS PRACTICUM

*Eric Kolaczyk, Boston University*

Kolaczyk described Boston University's commitment to developing students' data science skills, achieved through complementary top-down and bottom-up approaches to curricular innovation. In 2015, a master's of science in statistical practice (MSSP)[2] emerged, attracting a broad audience of quantitative students and producing holistically trained statisticians who have the foundational knowledge to work in an integrated data science environment. Participants enroll in eight courses and complete both a written portfolio and a two-semester statistics practicum.

He explained the primary motivations for developing the MSSP: (1) hiring organizations were increasingly demanding both degree completion and experience from their applicants; (2) employers wanted to hire people with both technical and communication skills; and (3) faculty were becoming dissatisfied with current course content. Thus, this revised statistics curriculum is practice-centric and requires the integration of diverse skills (Figure 5.1). Instead of adjusting existing infrastructure, MSSP faculty created a new organizational principle integrating practice and pedagogy and adopted a cohort-based system. The MSSP practicum's success is dependent upon a steady stream of real-world problems that are right-sized for student group work on various time scales, according to Kolaczyk.

The practicum is taught by a team of faculty members, fellows, and teaching assistants. Each class includes assigned readings, quizzes, discussion, and group work on topics such as data manipulation, visualization, modeling, and analysis; inquiry and interpretation; process management, workflow, and reproducibility; and communication. Statistical consulting is also available on walk-in, limited, and collaborative levels as part of the practicum. For their final projects, students collect data from various sources and must deliver a presentation, report, code, and data products—many of these projects focus on issues in the City of Boston (e.g., service quality, homelessness).

Kolaczyk described a number of challenges both in the curriculum and for the instructor, including balancing pedagogy and practice; emphasizing process; regulating project scope and timing; increasing students' independence; and elevating standards, goals, and accountability. He reiterated that the MSSP is a practice-centric, results-driven curriculum, and faculty must be prepared to modify course plans when necessary.

[2] The website for the master's of science in statistical practice is http://www.bu.edu/mssp/, accessed February 13, 2020.

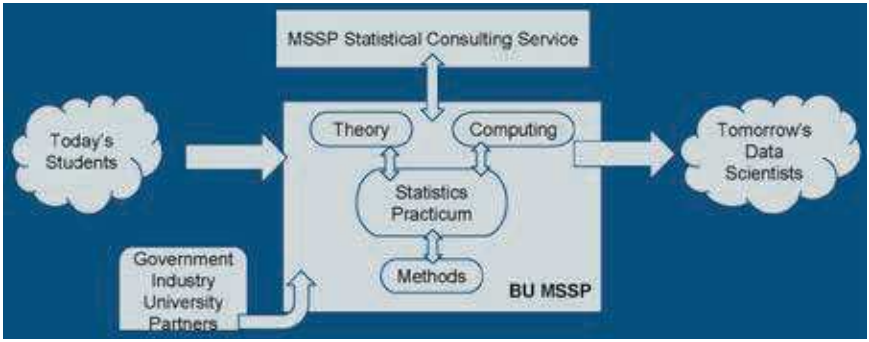


**FIGURE 5.1** Boston University's practice-centric approach to the master's of science in statistical practice. SOURCE: Eric Kolaczyk, Boston University, presentation to the roundtable.

In response to a question from a participant, Kolaczyk explained that students analyze scenarios as a way to discuss issues of ethics, fairness, and misuse. Katy Börner, Indiana University, suggested the use of an online learning system platform to deliver course content and reveal learning analytics. Kolaczyk noted that although it would be possible to move theory education to an online learning environment, this could create even more challenges in tracking students' results in the practicum. In response to a question from James Frew, University of California, Santa Barbara, Kolacyzk described that students work in small groups for nearly every component of the practicum, which provides good preparation for future workplace experiences. Kolaczyk relies on a strategy of "benevolent guidance" to arrange students in balanced groups. Another participant inquired about the program's success in workforce placement, and Kolaczyk noted that, anecdotally, students are being hired in a variety of data science positions and applying the skill sets developed in the program. In response to a suggestion from Karl Schmitt, Valparaiso University, Kolaczyk remarked that he hopes to disseminate educational materials from the program soon.

In a response to a question from Stodden about the role of university administration, Kolaczyk noted that university-wide changes can be challenging, especially for institutions that contain multiple schools. It is also important to recognize that infrastructure and hierarchies vary from campus to campus, which creates additional challenges when trying to scale programs. Alfred Hero, University of Michigan, suggested the creation of institutes or cross-school units to serve as brokers, identify commonalities, and resolve differences among departments. Schmitt commented that because data science is and can be done well at liberal arts colleges, which

are already set up to be integrative, many new technology hires come from institutions other than R1 research universities. He added that top-down approaches have been successful at Valparaiso, specifically, where co-teaching arrangements have been widely supported.

## CORNELL TECH AND THE JACOBS TECHNION-CORNELL INSTITUTE

*Ron Brachman, Cornell Tech*

Brachman highlighted New York City's 2010 call to universities to start or expand an applied science/engineering campus. Cornell University and the Technion–Israel Institute of Technology won the challenge, receiving $100 million in capital and a plot of city-owned land on which to build. Brachman emphasized that this new university, Cornell Tech, was developed from the ground up. Structured to offer practically oriented graduate degree programs, Cornell Tech opened in 2012 and, with an additional $130 million gift from the founder of Qualcomm, created the Jacobs Technion–Cornell Institute to focus specifically on application domain areas.

Brachman explained that Cornell Tech relies on a team-oriented "studio" environment that focuses on real problems to better prepare graduate students to enter (and lead) the digital technology economy. Interaction between academia and industry is key, and, because many graduates are hired locally, Cornell Tech has a direct impact on New York City's economy. He later added that Cornell Tech also employs a job placement staff, who contribute to the high placement rate of the graduates. A self-described "start-up company," Cornell Tech has graduated 324 students to date and currently has 250 master's students, 50 doctoral students, and 30 faculty on campus. According to Brachman, Cornell Tech plans to have 2,000 students and 200 faculty on campus by 2043.

Cornell Tech currently offers seven master's degree programs.[3] Brachman confirmed that although the coursework is structured differently, these degrees are equivalent to those conferred by Cornell University. All seven programs incorporate team-based, project-based learning, starting in the first semester—external companies with real problems ask the students to develop and manage products. This integrated studio education makes up approximately one-third of the total coursework for students enrolled in 1-year programs and includes alternative educational activities such as 24-hour project sprints, weekly critique sessions with

[3] For more information about these degree programs, see https://tech.cornell.edu/programs/masters-programs/, accessed February 13, 2020.

external practitioners, open studios, and opportunities to win monetary start-up awards. With such personalized attention and variety of rich experiences, he noted that this model could be challenging to scale.

Brachman commented that Cornell University does not yet have a specific degree or certification in data science. However, a cross-campus data science task force has emerged to evaluate current offerings and propose a new integrative structure for the future of "engaged data science." At Cornell Tech, specifically, faculty and administration are considering how the studio curriculum could integrate existing data science coursework into a degree or certificate program. In response to a question from a participant, Brachman remarked that, depending on the program, some incoming students come to Cornell Tech directly after receiving their undergraduate degrees, while others enroll after some amount of work experience. McKeown asked about the level of interest from external companies to engage more than once in Cornell Tech's studio projects and whether there are any intellectual property issues with the data they share. Brachman noted that companies continue to return, and the project list continues to grow. He added that companies are required to have a representative participate actively with the students, and they must agree that any work done by the students will become open source. Ullman raised a concern about students' open source data because venture capitalists who support their start-up companies may want to control that data. Brachman agreed that this issue warrants further discussion.

## AMERICAN STATISTICAL ASSOCIATION DATAFEST

*Andrew Bray, Reed College*

Bray described the American Statistical Association's DataFest[4] as a weekend-long competition held each spring on numerous campuses across the United States, Canada, and Germany. All participating host institutions must adhere to specified terms of use. During the competition, three to five undergraduate students—typically from the disciplines of computer science, statistics, engineering, business, social sciences, and natural sciences—work together to extract meaning from a complex data set (e.g., 10 million records from Expedia). The data set is not "revealed" until the first evening of the competition. The time that follows includes team time (all work must be done on site), support from on-site consultants, and optional workshops (similar to just-in-time teaching experiences). During the final evening of the competition, each team gives a

[4] The website for DataFest is https://ww2.amstat.org/education/datafest/, accessed February 13, 2020.

5- to 10-minute oral presentation of its findings to a panel of judges who are practicing data scientists from academia, industry, or the public sector. Two thousand students participated in 2017, and awards were given for best data visualization, best use of external data, and best insight.

Bray explained that DataFest gives students a sense of what it means to be statisticians as well as to be part of a community built around data. It is an opportunity for the students to practice the skills they have been taught in the classroom and to make connections with local data professionals. Students build technical and communication skills, learn to better generate and scope questions, and develop content for future job interviews. For faculty, DataFest can also be helpful in revealing some of the knowledge gaps that exist in the current academic curricula.

Bray acknowledged that there are a number of organizational challenges associated with DataFest. It can be difficult to find data that are of interest to the students, are not too specialized, are sharable, have multiple angles of inquiry, and are of appropriate size to be manipulated on a modern laptop. He cautioned that events such as DataFest could encourage irresponsible or ill-conceived analysis, so consultants interact with students throughout the event in an effort to combat such behavior. In the future, Bray hopes DataFest will coordinate a national competition, diversify data, and continue to increase student participation.

Brachman agreed that DataFest provides an excellent opportunity to expose students to the differences between responsible and irresponsible analysis, and Stodden suggested that examples from previous years' competitions be used as models, eliminating the need to embarrass any current participants who may be engaging in faulty analysis. Bray added that this topic could be integrated into a future DataFest workshop. McKeown asked Bray how DataFest prevents the exposure of private information. Bray commented that they utilize deanonymization and limit the number of covariates, as well as engage in lengthy discussion with participating companies' legal teams, but privacy continues to be challenging. He added that students will occasionally have to sign a nondisclosure agreement prior to participating in DataFest, or the company may lock up the data immediately after the competition concludes. David Ziganto, Metis, asked whether DataFest has considered using synthetic data instead to help avoid such privacy issues. Bray explained that the original intent was to give students as authentic an experience as possible with data as they exist in the wild, but he agreed with Ziganto that it is possible that even richer experiences could be had with synthetic data.

Hero asked what skills students need in order to participate in DataFest. Bray responded that students with some experience in a computational environment will be able to engage with their team and complete the challenge. He added that many students have worked in R, while

some have experience with Java, Python, MATLAB, or Stata, depending on their home disciplines. In response to a question from Ullman, Bray noted that industry representatives from companies that provided the data set do attend DataFest and occasionally will engage in follow-up work with a team that shared interesting findings. Börner remarked that a nonprofit organization without the resources to finance such a competition would benefit greatly from having students work on its data problems and offer solutions. Bray said that, historically, nonprofit organizations have not had the infrastructure to engage in DataFest, but he corroborated the value of having students do work with diverse data that could make a societal difference. Stodden asked about DataFest's level of integration with industry and wondered whether data science is being perceived as a scientific practice or as industry training. Bray explained that industry partners may support DataFest financially, and judges sometimes privilege a presentation with an actionable solution over one that is very scientific in nature.

## BOOT CAMPS

*David Ziganto, Metis*

Founded in 2013, Metis offered its first boot camp[5] in New York and now has locations in California, Illinois, and Washington. Ziganto explained that Metis's boot camp is the only one of its kind in the United States with endorsement from the Accrediting Council for Continuing Education and Training, though he hopes others will follow suit so as to improve the overall reputation of the boot camp model. In addition to a 12-week boot camp, Metis provides corporate training and online and evening courses in data science. He clarified that a boot camp is meant to bridge the gap between academia and industry and serve as a complement to other learning mechanisms. The boot camp model adjusts in real time to industry's demands for particular skills and technologies, while providing a fully immersive experience for participants.

Boot camp participants learn a combination of theoretical concepts and applications, including how to ask a solvable question, scope projects, collaborate and communicate with diverse groups, and use emerging tools and technologies. Ziganto added that Metis's boot camp allows students to work throughout the full data science pipeline on five projects (including posing the question and gathering the data), whereas in traditional academic settings, students typically enter the pipeline only when

[5] The website for Metis is https://www.thisismetis.com/data-science-bootcamps, accessed February 13, 2020.

it is time to explore and clean the data (Figure 5.2). In his view, the boot camp approach to the data science pipeline gives participants practice being data scientists before entering the workforce as entry-level data scientists.

Ziganto described boot camp participants in three ways: (1) fresh graduates without a portfolio; (2) career changers with a strong programming background, weaker math skills, and no portfolio; or (3) career changers with a strong analytical background, little programming experience, and no portfolio. Approximately 50 percent of participants have bachelor's degrees, while 49 percent have advanced degrees; 71 percent have industry experience, while 29 percent have experience in academia. He reiterated that a boot camp is meant to supplement hackathons, online courses, and advanced degrees, and he explained that successful boot camps have rigorous admission criteria, a rapidly evolving curriculum (partially influenced by employer feedback), instructors with industry experience, student-driven portfolio projects, and links to the data science community. For students enrolled in its boot camp, Metis provides career services to help with résumé development and networking. And for students who are not yet prepared to meet the admission criteria to enroll in a boot camp, Metis provides guidance for skill building.

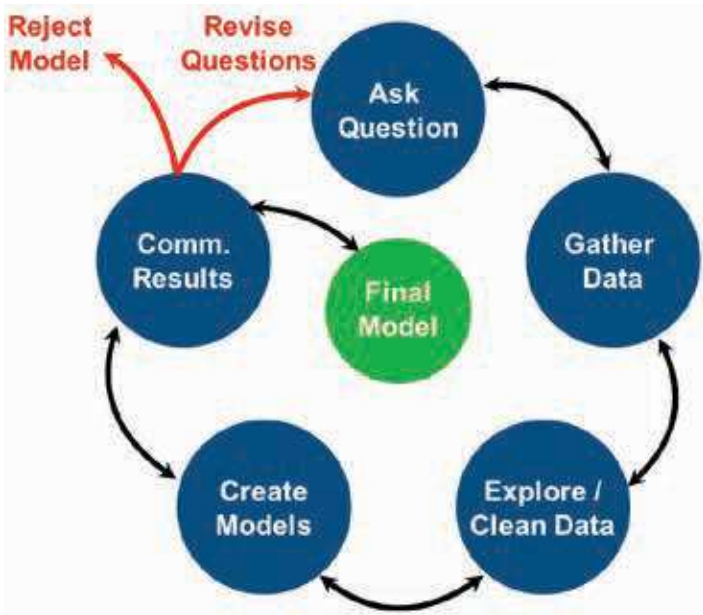


**FIGURE 5.2** Metis's comprehensive boot camp model of the data science pipeline. SOURCE: David Ziganto, Metis, presentation to the roundtable.

Stodden cautioned about the use of the term "boot camp" to describe such a program, concerned that the label can stifle curiosity or lead people to believe that data science is exclusive. She suggested the use of the phrase "quick start" or "jump start" instead. Another participant remarked that the term "boot camp" has an entirely different connotation—that of a remedial program intended to get participants up to speed on missing qualifications. Ziganto reiterated that Metis adopted the commonly used title because it symbolizes the fully immersive experience in which participants engage to refine their skills and build their portfolios. The participant added that both remedial and finishing programs serve equally important purposes, so it is important to clarify to participants what type of program is being offered when the term "boot camp" is used. Schmitt asked whether boot camps are targeted to particular sectors of industry, and Ziganto responded the boot camps focus on more sustainable data science fundamentals, while the sector-specific needs are addressed in Metis's corporate training programs. Ullman noted that algorithms, not models, solve data science problems, but Ziganto explained that the end goal in the data science process is to have an approximation, and a model is an approximation of reality.

## INFORMAL DATA SCIENCE EDUCATION

*Stephen Uzzo, New York Hall of Science*

Uzzo pointed out that while science practice has transformed, there has not been an equivalent revolution in science education. In his view, our ability to gather data has outstripped our ability to analyze it; new tools and techniques emerge rapidly; and data science pervades the science, technology, engineering, and mathematics learning ecosystem. Because data science problems are complex and interdisciplinary, data science has also transformed many other sectors of society. Yet, according to Uzzo, data science is generally not taught in any depth in the public school system, if at all, which ultimately threatens the pace of society's technological progress. This gap between data-driven science and technology practice and the understanding of science and big data for lifelong learning can be closed with big data literacy programs in informal educational settings, he explained. He noted that the abilities to adapt, innovate, collaborate, and analyze are essential in a data-driven society.

Uzzo explained that approximately 95 percent of learning happens outside a classroom, reinforcing the need for more informal science programming as well as for new technologies to access such educational opportunities (e.g., computational tools for visualization technologies).

The New York Hall of Science (NYSCI)[6] focuses on providing this needed data literacy to the public by offering knowledge when, how, and where the public can best engage with it. Science centers and museums exist, according to Uzzo, because most people learn better by doing, embodying abstract ideas, and engaging with phenomena (such as big data) through sight, touch, and creation. Core principles of museum experience and exhibit design include (1) placing people and play at the center, (2) envisioning visitors as creators, (3) introducing worthy problems with divergent solutions, and (4) issuing an open invitation to participate. NYSCI strives to create immersive experiences and share complex ideas to increase public interest and skills in science, which can be challenging for an audience of learners of various ages.

*Catherine Cramer, New York Hall of Science*

Cramer explained that NYSCI is situated in Corona, Queens, a community that is largely Spanish-speaking and includes 60,000 students—the largest school district in New York City. To support data literacy, the museum engages with local families, provides exhibits, offers public experiences, helps visitors understand new tools, organizes out-of-school programs, and hosts conferences. Cramer provided an overview of some of NYSCI's recent and upcoming activities:

- *Connections: The Nature of Networks*—Large floor exhibit, displayed 2004–2014.
- *Network Science for the Next Generation*—Three-year program pairing high school students from New York City and Boston with graduate students to create and present network science research projects.
- *Network Science in Education*—Hosts of annual international symposiums and teacher workshops, as well as authors of "Network Literacy: Essential Concepts and Core Ideas" (Cramer et al., 2015).
- *Big Data Fest*—2015 event in which 40 organizations provided data activities for the public.
- *Northeast Big Data Innovation Hub*—Effort to generate a collaborative inquiry process and a framework of principles for big data literacy.
- *Estuary Science Complexity*—Plans to develop a new science center that focuses on the data-dependent field of estuary science.

[6] The website for the New York Hall of Science is https://nysci.org/, accessed February 13, 2020.

- *Mobile City Science*—Program for students at New York's International High School who recently immigrated to the United States. Students used GoPro videos to map their community, identify problems, gather evidence, and propose solutions.
- *Big Data for Little Kids*—A current workshop designed to understand how 5- to 8-year-olds define, collect, represent, and interpret data, as well as how their caregivers engage with them in data inquiry activities such as variation, measurement error, data aggregation, interpretation, and prediction via a "make-your-own museum exhibit."
- *DataDive Exhibit*—Playful and personally meaningful experiences with data that help visitors understand patterns, algorithms, and machine learning processes.

*Katy Börner, Indiana University*

Börner shared her work in defining, measuring, and improving data visualization literacy—a combination of literacy, visual literacy, and data literacy that allows one to read, make, and explain data visualizations— which is critical for success in our data-intensive global society. In a study of 1,000 children and their caregivers who regularly visit a science museum, she found that most were unable to name, read, or interpret common data visualizations. She emphasized the need to bring more "macroscopes" to public spaces to help people make sense of large-scale data streams, identify patterns and outliers, and observe trends (Figure 5.3). She explained that macroscopes are not static instruments but rather continuously evolving bundles of software packages. She added that with numerous types of questions, varying experiences and knowledge of users, and different levels of abstractions, it can be challenging to create such toolkits.

One way to scale this education is through massive open online courses (MOOCs). Since 2012, students from 100 countries have participated with more than 350 faculty in Indiana University's Information Visualization MOOC.[7] Participants look at different workflows, run different types of analyses and visualizations, and learn to work collaboratively through algorithms to develop an actionable visualization. She also described a new project under way (joint among the National Science Foundation, Indiana University, the Science Museum of Minnesota, NYSCI, and the Center of Science and Industry in Columbus, Ohio) titled Data Visualization Literacy: Research and Tools That Advance

[7] The website for Indiana University's Information Visualization MOOC is https://ivmooc.cns.iu.edu/, accessed February 13, 2020.

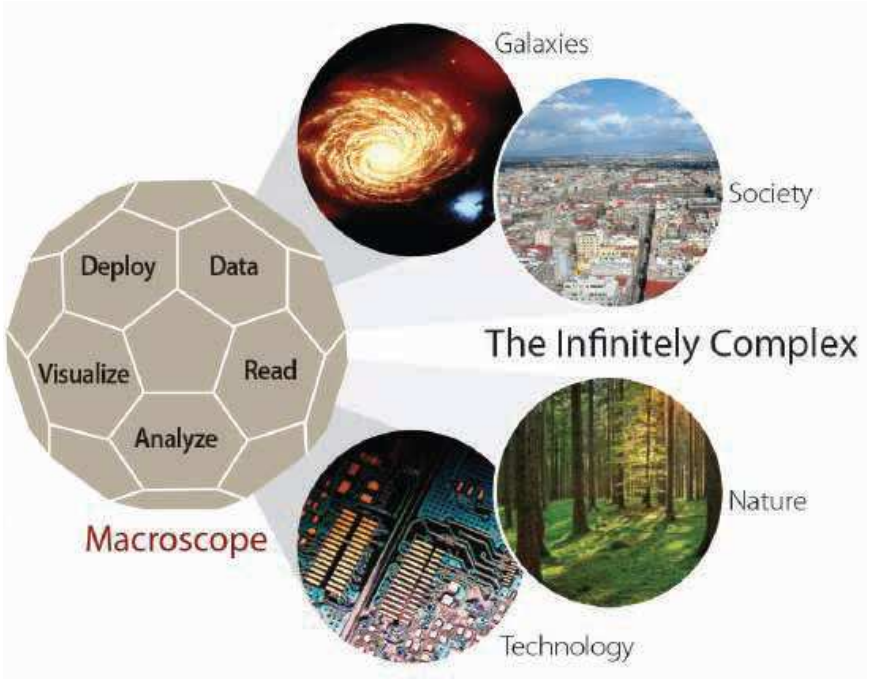


**FIGURE 5.3** While microscopes and telescopes only reveal the infinitely small and infinitely large, respectively, macroscopes allow one to study the infinitely complex. SOURCE: Katy Börner, Indiana University, presentation to the roundtable.

Public Understanding of Scientific Data. At the Science Museum of Minnesota, for example, a sports exhibit is available for children to explore and construct data visualizations after capturing their own race data and characteristics in a scatter plot.

Hero said that he has witnessed a decline in both data and visual literacy among high school students; he wondered how to reverse these trends and how to engage more students in science. Cramer noted that Network Science for the Next Generation students, for example, had little science training or interest in college prior to the program but became open-minded about their futures after the program. Even field trips to science museums can increase student interest in science, she added. Börner commented that many high schools are actively teaching visualization skills, and the global population of the Information Visualization MOOC has not demonstrated the decline in literacy skills that Hero described. She suggested that if U.S. schools continue to "teach to the test," data visualization questions can be added to those tests to increase data visualization literacy. Uzzo added that the Next Generation Science Standards

for K-12 students emphasize modeling (see NRC, 2011), suggesting that graphic literacy may be developed before high school begins. He also noted that network science is a field that appeals to students because of its focus on investigation; students can capture their interests from Harry Potter to human cells.

McKeown asked about strategies for diverse participation in informal settings, and Cramer noted that it took much work to engage her local community in NYSCI. Free museum entrance days and homework-help hours attract local families to the museum, which now has approximately 1,200 children visiting on a regular basis. Uzzo added that it is a challenge to appeal to and engage a wide age group in a single exhibit; however, many exhibits interest both adults and children when they simultaneously offer objects for children to manipulate and complex ideas for adults to ponder. He highlighted the importance of scaling the intellectual capacity of every space, especially because adults often accompany children to a museum. Börner supported intergenerational teaching and learning that happens outside a classroom setting, in which people of different ages and experiences share knowledge with each other.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Following a set of small-group discussions, Börner and McKeown shared considerations raised by their group for scaling data science programs. They noted the importance of trying to reach as many people as possible through varied methods of both formal and informal education. They suggested that libraries and museums serve as distribution systems for information that is not as readily accessible in rural areas as it is in urban environments. They cautioned about educational inequalities that exist owing to the economic circumstances of individuals or the resources of educational institutions. They lauded the value of experimentation and personalization in curriculum design. However, they noted that strategies that work in one setting may be difficult or inappropriate to scale in another. Last, they encouraged the development of top-down structures for program development.

Mark Krzysko, Department of Defense, asked the roundtable to consider carefully the definition of scale and the purpose of data science training. He encouraged start-up style thinking across campuses and emphasized that it is individuals who can bring cultural change to organizations and institutions. Kolaczyk agreed that cultural change is key, especially given that the term "data science" is so broad and relevant academic spaces are no longer so well defined. Bray noted the importance of protecting what academia has done well—teach durable skills that

outlast changing technologies. As the gap between theory and practice begins to close, however, and undergraduate programs and opportunities change drastically, he wondered whether master's programs will still be needed. Deborah Nolan, University of California, Berkeley, remarked that master's-level programs offer a deeper dive into the theory and methods previously learned and will adapt accordingly as the undergraduate programs change. She emphasized the need for undergraduate faculty to continue to focus on teaching fundamentals instead of emerging technologies. Börner suggested defining and surveying "timely knowledge" and "forever knowledge" for certain courses, as well as "theory" and "practice," to provide guidance for developing new curricula. Hero noted that while there are universities conducting these learning analytics, and balancing their use with student privacy considerations, others still rely on intuition and anecdotal evidence for course development.

On behalf of her discussion group, Nina Mishra, Amazon, discussed the values and challenges of project-oriented curricula. Her group emphasized the importance of understanding the purpose for incorporating student projects into a curriculum: Is the goal to prepare the students for industry jobs or to teach them how to use data to gain deep insight? She suggested that faculty avoid tailoring projects too closely to industry today, as most employers want to hire "big thinkers" who can solve tomorrow's problems. Students may be most successful if the project allows them to develop the analytic skills needed to work on future data projects. Mishra also explained that there is a spectrum of projects that serve different purposes, and those that require low student-to-faculty ratios may be difficult to scale. She also highlighted that it can be difficult to access company data for student projects, which can be a concern because students may not be as excited by the alternative option of working with public data. Mishra wondered whether collaborating more closely with industry or searching for new resources could alleviate this constraint. Last, Mishra noted the importance of carefully scoping the project problem with students so that there is a concrete question to be answered. This, in addition to active engagement from participating companies, can improve project outcomes. Hridesh Rajan, Iowa State University, suggested that because access to alumni networks and industrial partners (and thus projects) is limited on some campuses, it would be helpful if an open resource of projects were available to all institutions.

Kolaczyk noted that it is important to balance what industry wants and what students need. For example, when students persist about learning how to use a particular software package, it is the responsibility of the faculty to shift their mindsets by explaining that all of the technologies will change and that there are multiple languages in which to communicate data. According to Kolaczyk, some of this cultural change can happen

through facilitated group-based self-learning. Ziganto noted that there are only so many "big thinkers," and people who can make smaller changes are also essential—what is most valuable to employers is a student with the right fundamental knowledge to be able to learn quickly and adapt to new situations.

# 6

# Meeting #5: Integrating Ethical and Privacy Concerns into Data Science Education

The fifth Roundtable on Data Science Postsecondary Education was held on December 8, 2017, at the Keck Center of the National Academies of Sciences, Engineering, and Medicine in Washington, D.C. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss the integration of ethics and privacy concerns into data science education. 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.

Welcoming roundtable participants, co-chair Eric Kolaczyk, Boston University, noted that there are inherent ethical and privacy implications in the choices data scientists make while framing, obtaining, cleaning, manipulating, and interpreting data. He highlighted the value of integrating this context of data science practice into data science education, and he hoped that the conversations at this gathering of the roundtable would contribute to a more principled awareness of the ethics of data science.

## TEACHING ALGORITHMIC ACCOUNTABILITY IN DATA SCIENCE EDUCATION

*Cathy O'Neil, mathbabe.org*

O'Neil began her presentation to the roundtable by suggesting that data science ethics be reconceptualized as "algorithmic accountability."

She noted that although countless organizations use algorithms to score individuals (e.g., to estimate their propensity toward some desirable or undesirable behavior), their processes are not always scientific or ethical, and privacy and accountability may not be at the forefront of their concerns. What is most unfair, O'Neil described, is that recipients of such scores have no means to understand them, and there is often no mechanism in place to appeal decisions made as a result of these scoring systems. While these potentially destructive scoring algorithms rise to "secret laws," in O'Neil's point of view, she said that many companies have yet to find evidence that they are effective in reflecting the true likelihood of what they purport to score. An algorithm, according to O'Neil, makes predictions based on historical patterns. Although the definitions in an algorithm used to score individuals are crucial, these definitions are often determined secretly by those in power. Concerns also arise about the understanding of false positives and false negatives generated by the algorithm— balancing failures is just as important as having an accurate algorithm, O'Neil explained. She emphasized that it is already technically challenging to understand how and why various algorithms fail in different ways; it becomes even more difficult to hold algorithms accountable when they are optimized to a secret definition of success.

O'Neil provided three examples in which unaccountable, discriminatory algorithms are used in society:

1. Teacher assessment based on students' test scores. Such a scoring system relies on bad proxies (i.e., test scores), bad statistics (i.e., low correlations), and questionable practices.
2. Job application filters such as mental health assessments and gender. Such a scoring system is discriminatory, difficult to measure, and even more challenging to fix.
3. Police dispatch to neighborhoods with high arrest data or arrest of low-level criminals to prevent violent crime in the future. This unscientific system uses biased data and bad proxies (i.e., crime data are not the same as arrest data).

O'Neil commented that because lawyers and policy makers often do not have the appropriate levels of technical expertise, it is unreasonable to expect the legal system to keep pace with advances in data science. She encouraged academicians to address this issue of accountability in data science classrooms. She advocated for exposing future data scientists to these problems and teaching them to see themselves as accountable for ethically responsible products. She also suggested that, instead of only critiquing existing algorithms, data scientists who build algorithms

could help policy makers by producing white papers geared toward non-experts and could involve lawyers in the development of ethical guidelines for algorithms.

O'Neil hopes that university data science institutes will also play a larger role in the development of accountable algorithms. She noted the value of having a Hippocratic Oath for data science and encouraged data scientists to focus on their roles as translators of ethics instead of arbiters of truth. In response to a question from Solon Barocas, Cornell University, she suggested that data scientists reject jobs with organizations that do not build ethical (and legal) algorithms. For organizations utilizing algorithms for decision making, she suggested a scaffolding of monitors to ensure that algorithms are fair and legal and that data are clean. In response to a question from Patrick Perry, New York University, she elaborated that such monitors are valuable because they provide a continuous version of scientific algorithmic testing. She acknowledged that external data would be needed for validation throughout such testing.

Victoria Stodden, University of Illinois, Urbana-Champaign, asked O'Neil how she would teach these concepts at the graduate level. O'Neil responded that it is useful if every question to be addressed by an algorithm corresponds to a randomized experiment and if extreme mathematical cases are introduced. Aaron Roth, University of Pennsylvania, noted the bias that exists in data, even when humans make decisions, and wondered how machine learning is distinct from human decision making in terms of fairness. O'Neil highlighted the misconception that machine learning removes bias and encouraged humans to make their values explicit in the development of algorithms. Charles Isbell, Georgia Institute of Technology, asked how far the legal framework could be extended in algorithm development, and O'Neil responded that algorithms are already subject to the law; the questions that remain are whether these laws are enforced and when regulators will have the appropriate tools to measure legality. In response to a question from Perry about algorithmic definitions of success, O'Neil suggested having stakeholders complete an ethical matrix of their concerns about an algorithm. Such a matrix reveals that fairness is always a balancing act when trying to optimize with so many constraints. Perry countered that it seems implausible to determine the cost of making the wrong decision, but O'Neil reiterated that while considering the ethical implications is difficult, it is essential. Barocas identified this as another example in which challenges related to fairness still exist even when the data are reliable.

## UNCOVERING THE SUBSTANCE OF A DATA SCIENCE ETHICS EDUCATION

*Solon Barocas, Cornell University*

Barocas focused on the content of data science ethics education as opposed to the structure through which it is delivered (e.g., stand-alone courses versus integration throughout an entire course of study). He began by extending standard concepts of professional responsibility common to many fields—to do work that is valid, reliable, and transparent—to data science practice and education. Similarly, common professional virtues to strive to instill within future data scientists include skepticism about how models will perform, humility regarding the limits of the models that one develops, honesty to avoid misleading users, and vigilance to ensure that models work well after deployment. Standard ethical dilemmas can motivate students to question and develop their own moral agency and moral intuition. These generic approaches to professional ethics do provide value in the context of data science education, particularly in helping students to connect concepts of validity and reliability to questions of fairness and bias in algorithms with relative ease. However, Barocas commented that these approaches are not specific to data science and thus may be inadequate for data science ethics education.

Barocas remarked on the growing interest in the field of "data ethics," noting that it is unclear what this field entails. Standard approaches underscore privacy (i.e., adherence to the Fair Information Practice Principles [see FTC, 1998] and use of anonymization to safeguard personal information); however, clearly new ethical issues are arising in data science that fall outside of this narrow purview. The past few years have seen increased interest in adapting research ethics principles (i.e., autonomy, beneficence, and justice), which are historically designed to protect research participants, to the use of data analytic tools in companies. This is not a surprising approach, explained Barocas; however, research ethics still does not encompass the breadth and complexity of the ethical and normative questions that future data scientists will face.

Barocas described a new upper-level undergraduate elective at Cornell University—INFO 4270: Ethics and Policy in Data Science[1]—targeted toward aspiring data scientists from the disciplines of information

[1] The course website for INFO4270: Ethics and Policy in Data Science is https://docs.google.com/document/d/1GV97qqvjQNvyM2I01vuRaAwHe9pQAZ9pbP7KkKveg1o/edit, accessed February 13, 2020.

science, computer science, and quantitative social science. He mentioned that much of the syllabus grew from the annual Fairness, Accountability, and Transparency in Machine Learning Workshop,[2] which seeks to build a technical community interested in deeper normative questions in data science work. While interest in this conference and its subject has grown rapidly, Barocas worried that some researchers still mistakenly think that formalizing decision making through algorithms ensures fairness or prevents bias. In part, this is based on experience with human decision makers who do exhibit bias, which can be ameliorated through more formal decision processes (e.g., actuarial scoring tools). He emphasized that using machine learning does not ensure fairness and that misuse of data science can foster inequality in and prevent opportunity for segments of the population.

Ethics and Policy in Data Science challenges students to explore familiar technical problems—for example, detecting unobserved differences in model performance, coping with observed differences in model performance, and understanding the causes of differences in predicted outcomes—with greater ethical specificity. Focusing on an example of model validation, Barocas said that data scientists must make normative decisions during validation (e.g., to validate with respect to accuracy of predictions, with respect to differences in error rates, or with respect to differences in outcomes across subpopulations) and that data science ethics education can engage students in deliberation about the ethical implications associated with their modeling decisions. Regarding differences in outcomes, Barocas suggested that data scientists consider the historical events that shape algorithmic outputs about an individual (e.g., whether that person's family has a history of interaction with the criminal justice system) and to perhaps consider algorithmically aided decision making as a way to remedy past injustices.

Ethics and Policy in Data Science consists of 12 broad modules: cultivating a critical disposition in students toward data science and their own work; understanding bias in humans, algorithms, and data; case studies and opportunities in algorithmic auditing; formalizing fairness and trade-offs between different measures of fairness; individual agency and individualized assessment and the ethical dimensions of modeling individuals based on factors over which they have no control or based on their characteristics in reference to larger populations; moving from allocative to representational harms; transparency, interpretability, and explainability of algorithms and models from the perspective of policy makers or tool users; privacy protections and loss of privacy from precise, automated inference;

[2] The website for the Fairness, Accountability, and Transparency in Machine Learning Workshop is https://www.fatml.org/, accessed February 13, 2020.

price discrimination in marketing and insurance models; broader questions about algorithms in the public and their impact on democracy; and the ethics of autonomous experimentation by algorithms deployed in the real world. The final module in the course is about refusal and rejection, where data science students and practitioners explicitly choose not to pursue specific projects because they are ethically or practically objectionable. Barocas closed his presentation by appealing to senior data scientists to lead by example in refusing ethically questionable projects, which in turn will provide an example and protection for more junior researchers, practitioners, and students wishing to reject a project.

## RECOGNIZING AND ANALYZING FALSE CLAIMS FROM BIG DATA

*Jevin West, University of Washington*

West opened his presentation by noting that while many students excel in the execution of mechanics, they often lack the skills both to engage with ethical considerations for data analysis and to understand basic experimental design. In his classroom, West reveals to students, who may not appreciate the limits of technology, that machines make mistakes and harbor bias similar to humans. Instead of offering only a brief unit of study on ethics, he integrates these conversations throughout his curriculum. He encouraged faculty to adopt the humanities' approach to textual analysis, as future data scientists need to develop critical thinking skills to interrogate and interpret data.

West commented that society is drowning in false information, especially with the rise of charts and quantification in the news. In an effort to teach students to recognize and analyze false claims and to be able to communicate this information to broad audiences, West and his colleague Carl Bergstrom developed a course[3] at the University of Washington. The course includes topics in the following areas: false information and misrepresentation, causality, statistical traps and trickery, data visualization, big data manipulation, publication bias, predatory publishing and scientific misconduct, fake news and other shams, and refutation of falsehoods. Given that the course emphasizes data reasoning, West dedicates much instructional time to causation and refutation. Campuses across the country and abroad have adopted the course, and West and his colleagues also engage local middle and high school students in similar instructional sessions.

[3] The website for the course Calling Bullshit in the Age of Big Data is https://callingbullshit.org/, accessed February 13, 2020.

West next demonstrated a contrast between "old school bull"—empty phrases and circular reasoning are readily detected and disproved—and "new school bull"—scientific language and visualizations are presumed to be fact. While he acknowledged that the notion of the "black box" can be daunting to students, they can recognize misrepresentations by looking carefully at the data that are input into the algorithm as well as the output and the interpretation of an algorithm.

West suggested using real-world examples to create engaging classroom exercises that challenge students to identify instances in which an argument's methods or assumptions lead to absurd conclusions or causations. He shared a series of tips for spotting false claims: (1) Think about claims that seem too good to be true; (2) Beware of confirmation bias; (3) Recognize multiple working hypotheses; (4) Evaluate orders of magnitude; and (5) Be wary of unfair comparisons. West concluded by emphasizing the value of improving ethical data science education models at the secondary and postsecondary levels and engaging students and the broader public in data reasoning.

## OPEN DISCUSSION

### Bias and False Information

Jeffrey Ullman, Stanford University, described the "fake news" discussed in West's presentation as an intractable problem and asked for ideas to formally identify it. West admitted that, right now, it is impossible—it is important to arm machine learning consumers with the right skills and hope that artificial intelligence will catch up eventually. He explained that, unfortunately, for every algorithm created to identify fake news, there is another one designed to create fake news. Bill Howe, University of Washington, commended O'Neil's and Barocas's attention to validity but expressed concern that people may be under the false impression that simply building the perfect model solves all problems. He emphasized that the issues are far more complex. O'Neil said that the other, worse extreme is when people assume that nothing can be trusted and lose faith in technology entirely. She advocated emphasizing the science in data science by testing frameworks around algorithms so that they can be trusted. Barocas shared O'Neil's concerns but added that the foundation for seemingly objective work is actually subjective (i.e., nothing can be learned without some amount of bias). Howe also noted that data scientists have choices and power before training a model, and he emphasized the value of teaching students about these crucial data management steps. Barocas agreed and suggested starting courses with the question, "What is data?" Stodden observed that the topic of bias was central to all three presenters'

talks. Because bias is defined narrowly in entry-level statistics courses, and that definition may not translate well in larger discussions, she suggested that the data science community think about how to teach what bias is as well as how to think about data science more broadly.

### Preparation for Faculty and Students

Michael Fountane asked the presenters about senior-level faculty responsibilities in teaching future data practitioners. Barocas noted that, generally speaking, professors want to produce students who will do high-quality work. He added that competitive marketplaces should then reward those who become practitioners and avoid making statistical errors. O'Neil commented that, especially in financial trading, there is a strong incentive to be accurate so as to maximize profit but there are not nearly as many stakeholders as there are in data science spheres. The realm of data science is much more complicated because these many stakeholders have differing definitions of success, and their values have to be balanced against one another. Many people, she explained, either misunderstand this complexity or choose not to think about it. West reiterated that ethics instruction (i.e., a new way of thinking about and communicating the social elements of data) has to carry through all components of a data science education.

David Culler, University of California, Berkeley, wondered how to educate students to exercise good judgment. West noted that his course incorporates case studies and project-based work in which students are set up to fail; they quickly learn about the value of good judgment in such scenarios. O'Neil said that students can be taught to practice good judgment through exercises in which they work on one algorithm with multiple choices. Barocas discussed the importance of providing students with messy data so as to better prepare them for real-world experiences. Alfred Hero, University of Michigan, cautioned that although flagging false claims can energize students, it risks showing students that finger-pointing is always justified. Instead, Hero suggested teaching students to ask what evidence would be needed to make a true claim. He described this as a more constructive way to teach about the inadequacy of selected data and to increase appreciation for negative results, because this is how the scientific enterprise is motivated to continue its work. West noted that selection bias and reproducibility are topics of his course lectures, as are the civic and political implications, and he added that the field of data science could also learn from approaches used in applied psychology. Moses Namara, Clemson University, asked how to motivate people to scrutinize data, and West responded that students are both idealists and natural contrarians. He said that it is important for students to understand the

consequences of misusing data, but he cautioned against letting students believe that no truth exists anywhere. Nicholas Horton, Amherst College, commented that there is a clear need for a variety of approaches to and a spiraling curriculum for educating future data scientists. He emphasized addressing key concepts early and often in courses and encouraged the building of critical thinking skills at different levels. He urged faculty to identify learning outcomes related to data integration and data fusion and suggested enhanced faculty training. He described "data literacy for all" as a way for people to better understand the world around them without fear.

## MATHEMATICAL APPROACHES TO PRIVACY AND FAIRNESS

*Aaron Roth, University of Pennsylvania*

Roth presented two important social issues in technology: privacy and fairness. He emphasized the value of approaching these complicated issues from formal, mathematical perspectives. He noted that mathematical approaches to privacy already exist, but fairness is an emerging area of study with a recent explosion of research. A standard definition of and a quantitative approach to fairness would be useful, according to Roth, but both privacy and fairness require understanding trade-offs through formal reasoning.

Roth described privacy as the promise of freedom from harm. Privacy has been a public concern for decades; despite the use of de-identification techniques, people can still be connected to their data. He acknowledged that privacy is more complicated than hiding personally identifiable information or releasing only aggregate statistics. It is impossible for data analysts not to know anything more about a subject after analyzing that person's data when auxiliary information is present. Roth pointed out that if this instance is treated as a privacy violation, it becomes virtually impossible to do scientific research, because auxiliary information often reveals information data scientists want to learn. He alluded to an article by Dwork et al. (2006) that discussed the notion of differential privacy—a data set (in which each piece of data belongs to an individual) is input into a randomized algorithm, and even if the data are changed for one individual from the data set, the behavior of the algorithm should not change substantially.

Roth noted that many statistical problems can be solved privately with convex optimization, deep learning, spectral analysis, and synthetic data generation, for example. He emphasized that trade-offs will always exist (e.g., accuracy, sample sizes, and privacy level)—mathematical thinking simply allows one to better understand those trade-offs. He noted that although there is still much work to be done translating theory

into practice, organizations such as the U.S. Census Bureau already rely on differential privacy.

Roth next turned to a discussion of fairness, using a case study of COMPAS—the recidivism risk prediction software. Investigative journalists at ProPublica described the tool as unfair and biased against black people, owing to differences in false positive and false negative rates between black people and white people (see Angwin et al., 2016). COMPAS analysts responded that they used a different metric for fairness (see Dieterich et al., 2016). While Roth explained that both analyses offer reasonable definitions of fairness, no classification tool can simultaneously satisfy both conditions and equalize false negative rates if the base rates in the two populations differ. Equalizing false positive rates across subpopulations is only one measure of fairness, and it is unclear whether this is the appropriate metric. Roth concluded that the benefit of formalizing such fairness measures is that it allows better management of trade-offs, improved algorithmic design, and scientific progress toward more informed policy making.

## NAVIGATING HISTORY, PRIVILEGE, AND POWER IN INFORMATION AND DATA SCIENCE

*Anna Lauren Hoffmann, University of Washington*

Hoffmann encouraged the teaching of ethics in applied contexts—better decisions can be made about issues with moral impact if a combination of disciplinary, theoretical, and activist knowledge is considered. It is important for data science students to realize that different problem solving goals require unique considerations. She emphasized that historical and contextual information are essential in ethical decision making.

Hoffmann observed that data ethics is the intersection of moral, methodological, and practical concerns—data scientists need appropriate tools to balance these three areas. She emphasized the value of confronting these issues with disciplinary diversity, utilizing people with varied skill sets to solve complex problems. In her courses, Hoffmann approaches ethical considerations through a study of context, relevant history, key concepts, and the data life cycle. She suggested that ethical issues arise not only in algorithms and analysis, but also in data collection, and thus should be a part of the entire research life cycle. She noted the importance of teaching students to think about how platform design affects data as well as how to think critically and holistically about data and the problems that data can solve.

Hoffmann emphasized that, like any tool, data have affordances. Ultimately, data allow one to count, organize, and make decisions. She

emphasized that these processes are not wholly new—there is a canon of historical examples that expose the importance of research ethics. Discussion of Nazi experimentation and the Tuskegee studies, for example, help contemporary students understand ethical issues and determine how to apply these lessons to current case studies. She offered multiple examples, including the Henderson Roll—an illegal census in the 19th century of Native Americans—as evidence of historical precedent about the vulnerability of certain populations in the face of data-driven systems. She emphasized that such injustices could be perpetuated when people voluntarily provide records to the government (e.g., as discussions about Deferred Action for Childhood Arrivals continue in the United States) and reiterated the importance of thinking about uncovering issues in and using history to solve current problems in new ways.

## OPEN DISCUSSION

### Teaching Differential Privacy

John Abowd, U.S. Census Bureau, suggested that faculty focus on teaching the ratio of differentially private variance to the regular variance in their courses—the trade-off is clear and the privacy costs are revealed in that instance. Howe mentioned the tension between reductionism and interpretability. For example, he wondered how many people (especially lawyers responsible for decision making) can reliably understand and interpret differential privacy. Roth acknowledged that it will never be possible to write a mathematical constraint about privacy upon which everyone will agree, but formalization helps to reveal incompatible components. He noted that although one may not know a parameter for differential privacy, a quantitative discussion about privacy levels is possible and useful. Hoffmann added that reflective conversations about privacy and debates about trade-offs are more valuable than a focus on finding the "right answer." Hero noted that differential privacy and its measures place the analyst in the role of determining acceptable levels of privacy. But, he hypothesized, in the future, when individuals can select their own trade-offs, privacy may become a valuable, tradable asset. Roth clarified that the analyst does not set the privacy level and added that differential privacy is only a metric. He mentioned that there have not been many successful markets for private data in big data applications thus far because they are not very useful and are easily replaceable. He noted that people would need to alter the way they think about privacy before data markets would change.

### Balancing Trade-offs

Perry noted that privacy, fairness, and accuracy are all trade-offs that are at odds with one another. He wondered whether one should either place different weights on each factor and optimize for the objective function or explore the frontier first and then assign weights. He asked whether the latter approach is dangerous because it allows a decision to be made after seeing the trade-offs—in other words, sacrificing privacy and fairness for accuracy. Roth explained that it is the responsibility of the technologists to identify trade-offs and of the society to balance competing needs. It is only possible to make fully informed decisions after understanding the trade-offs, according to Roth. Hoffmann refuted this notion: by setting up values as trade-offs, one has already surrendered to certain inequalities. Roth responded by saying that while it is tempting to suggest that because the Constitution guarantees fairness it cannot be a trade-off—that is not the reality in which we live. Trade-offs have to be discussed in every case. Hoffmann acknowledged that discussions about trade-offs are useful unless a concession has already been made earlier in the process and a different set of trade-offs needs to be debated. Roth agreed that it is always reasonable for researchers to step back and evaluate what is most important.

In light of this discussion about fairness, Isbell commented that bias can be built into the data itself. Roth acknowledged that both data and algorithms can be problematic in terms of bias. He explained that problems with data are hard to measure, and, even if those problems were eliminated, fairness would still be an issue. He encouraged investments to be made in the study of both data and algorithms. To simplify the problem, he suggested first thinking about the data and algorithms in isolation. In response to a follow-up question from Isbell, Roth noted that it should be possible to formalize the problem in data collection and reiterated that fairness is only just beginning to be understood. Ullman asked about assumptions about right and wrong that faculty are making in their courses, as well as about how far rights to privacy extend, and wondered whether the conversation should focus only on the implementation of technologies. Roth reiterated that the technologist's role is to help discover and delineate trade-offs, not to make decisions about policy or morality. He added that it is possible to write definitions with parameters on which trade-offs will occur. Barocas added that the Fourth Amendment determines certain rights and noted that it is the responsibility of faculty to show students that long-standing issues are not new simply owing to the onset of data or technology. Hoffmann emphasized that when people are being harmed by data and software, science must progress to make changes.

### Navigating Social and Technical Concerns

Barocas asked Roth and Hoffmann to comment on one another's talks. Roth acknowledged that he has not yet overcome the obstacles of language differences across disciplines. He explained that even though he and his students are often operating with toy models for which the complicated problems of the world have been abstracted away, the complex problem of fairness persists. Roth continued that because fairness is complicated, it is critical to understand first how social and technical issues work in isolation and then how they work together. Hoffmann said that her work provides the broader social and political motivation for Roth's research. She reiterated that his and others' work has a larger framework; working ahistorically will only further fragment problems. Referencing Stodden's earlier observation about bias, Hoffmann recognized that communities contextualize bias differently— social theory and historical casework can orient people toward a positive vision about a socially acceptable definition. Mark Krzysko, Department of Defense, mentioned that his team regularly confronts many of the issues discussed in Hoffmann's presentation. He added that access and dissemination are also concerns for the Department of Defense and that it is important for future employees to understand and to engage in constructive dialogues about both data and institutional values.

### Educating Students

Kolaczyk asked how Hoffmann and Roth raise issues of ethics and fairness in the classroom. Hoffmann requires written assignments including memos, opinion pieces, blogs, and, during the data collection stage, reflective exercises. Roth responded that because he teaches mathematics to Ph.D. candidates, his focus is on equipping students with the skills to push research topics forward. He encourages students to look at popular media to explore real-world questions, but he does not teach interdisciplinary courses or content. Abowd asked how Roth would incorporate the notion of "privacy as a public good" into discussions about system design; Roth recognized the value in thinking about privacy in terms of economic quantities to be analyzed, and he supported further collaboration between scientists and economists.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Roundtable members and audience participants formed subgroups to discuss one or more of the following questions: (1) How could ethics

be integrated into the data science curriculum? (2) What mechanisms can educators use to help students navigate between the informal (e.g., fairness) and formal (e.g., algorithmic accuracy) terminology of data science? (3) How might educators teach students about the ethics of data science without radicalizing or paralyzing them with skepticism?

Isbell shared considerations raised by his group about integrating ethics into a data science curriculum. As most of the roundtable's discussion focused on real-world consequences of actions, he questioned whether "ethics" needed to be integrated at all. While it is possible to explore real-world consequences through the lens of ethics, there are other ways to do so, he continued. He noted that it is unrealistic to expect faculty with varied levels of expertise to integrate ethics units into their courses, and it would be equally unsuccessful to create an ethics course out of context. He instead suggested asking faculty to focus on problems assigned to students in each and every class meeting and then working toward a study of real-world consequences. This approach may be both easier to integrate across the curriculum and more interesting for students. An audience participant suggested developing a required ethics course, separate from core requirements and including guest lecturers from other departments, as a way to motivate students to think about the consequences of working with data. This participant also suggested adding a question about real-world consequences to every student project.

On behalf of his group, Fountane suggested incorporating ethics into curricula with the implementation of an orientation course on reasoning at a high meta-cognitive level (i.e., how to write/model ethical standards to peers). He described this as a practical way for students to engage in active reflection, which is a more socially valuable skill than calculation. Faculty may be more likely to buy in to this approach if ethics played a larger role in the professional data science discipline. For example, both the Association for Computing Machinery and the American Statistical Association already have guidelines for the inclusion of ethics in professional practice. An audience participant added that Bloom's Taxonomy (see Bloom, 1956) could be used to structure and integrate conversations on data ethics in either of the classroom models shared by Isbell or Fountane. She explained that Bloom's Taxonomy offers a way for undergraduates to develop evaluation and critical reasoning skills gradually through paced activities. Hoffmann shared her group's discussion, noting that educators do not yet have a good understanding of how much content from other courses might be useful in the development of ethics curricula. Referencing the National Academy of Engineering report *Infusing Ethics into the Development of Engineers* (NAE, 2016), she wondered which models already exist and how much data science educators should create

anew. She suggested that industry create incentives for academia to better understand and to better provide data science education. She also added that there are many structural barriers to developing new curricula in higher education—administrators and students alike often do not realize which skills industry values. An audience participant expressed concern about the lack of diversity in STEM Ph.D. programs and noted that it is crucial to discuss diversity in any conversation about ethics. She explained that students are, by nature, curious and can be motivated to study data science when courses are student- and project-centered. She emphasized that embedding ethical conversations and real-world problems into courses can also serve as a mechanism to improve diversity.

Summarizing his group's discussion, Perry noted that although many faculty may want to teach ethics, they may not know the best approach. For this reason, he cautioned against forcing faculty to teach ethics. He observed the need for faculty scripts for smoother incorporation of ethics without disrupting course material and for case studies relevant to material being taught. So that students do not become paralyzed with skepticism, it is important that these case studies show students possible solutions to problems. Because undergraduate students are often interested in exploring problems and nontechnical material, it may seem more feasible to incorporate ethics at this level than at the master's level, in which students are focused on building technical skills that can be applied in the workforce. Perhaps a way to introduce ethics at the master's level is to talk about the mathematics of differential privacy. Most importantly, Perry explained, it may be detrimental for students to believe that mathematics solves all problems. It is crucial that students are involved in hands-on exercises that show the consequences of fairness. For example, faculty could ask students to build a model to predict something that is relevant to them and randomly assign covariates. Such an exercise offers personal incentives (beyond the moral imperative) for students. Howe suggested replacing data sets in the classroom (e.g., use COMPAS instead of Iris) to start a conversation about fairness. He also wondered whether there is a way to obtain better curated data sets that could be crafted for teaching purposes, although he recognized that students are rarely excited about "fake" data. He referenced an effective and engaging assignment from danah boyd—the only correct way to complete it was to refuse to complete it on ethical grounds.

Kolaczyk, speaking on behalf of his group, suggested that the best practices used by the bioinformatics community for including ethics in the curriculum could be leveraged, if they are computationally motivated. He added that there are examples from the social sciences for integrating both the quantitative and the qualitative (e.g., survey and sampling taught together). He also noted that the integration of ethics depends

on the context—for example, Boston University's statistics practicum includes situational role-play, which may not be as effective in a theoretical course. Hero encouraged the inclusion of statistical consulting in engineering projects as a way for students to learn to interact with both people and data and become more aware of the consequences of their actions. Such an experience is personalized, offering a well-designed teachable moment. Constantine Gatsonis, Brown University, referenced a course he is designing: Case Studies in Health Data Science. The course invites speakers from Brown's School of Public Health to present real data sets and case studies. He will provide a template for privacy and ethical considerations around which speakers will organize their presentations. He hopes that this will be an effective means to generate useful discussion with the students. Stodden highlighted a forthcoming publication in *Communications of the ACM* that emerged from a working group of the Advisory Committee of the Computer and Information Science and Engineering directorate of the National Science Foundation. The document illustrates the life cycle of data science (i.e., acquire, clean, use, reuse, publish, preserve, and destroy), with ethical questions to be addressed from technical and mathematical perspectives at each stage, making it possible to start to frame concretely how ethics fits into data science.

# 7

# Meeting #6: Improving Reproducibility by Teaching Data Science as a Scientific Process

The sixth Roundtable on Data Science Postsecondary Education was held on March 23, 2018, at the Hotel Shattuck Plaza in Berkeley, California. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss how data science can be used to help understand and improve reproducibility of scientific research and to highlight several courses and training offerings in reproducible data science. 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.

Welcoming roundtable participants, co-chair Eric Kolaczyk, Boston University, noted that although replicability is a fundamental aspect of the scientific process, many have suggested that a "crisis in reproducibility"[1] currently exists. Recently published articles, such as "Why Most Published Research Findings Are False" (Ioannidis, 2005), have identified errors in research findings, and numerous workshops have been hosted on reproducibility. With data collection, management, analysis, and reasoning activities becoming pervasive throughout society, he said that the data science community is advocating that reproducibility be integrated

[1] Replicability "refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected," whereas reproducibility "refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" (NSF, 2015).

throughout the data science process. He suggested that academic institutions facilitate reproducibility as a mainstream practice.

## DATA SCIENCE AS A SCIENCE: METHODS AND TOOLS AT THE INTERSECTION OF DATA SCIENCE AND REPRODUCIBILITY

*Victoria Stodden, University of Illinois, Urbana-Champaign*

Stodden encouraged roundtable participants to frame data science as a science. She provided a brief historical overview of the tenets for scientific practice including (1) Robert Boyle's 17th century belief that any write-up of an experiment should be thorough enough for a reader to repeat the experiment; and (2) Robert Merton's 20th century emphasis on communalism, universalism, disinterestedness, and, most relevant to the current discussion of reproducibility, skepticism. However, she explained that scientific practice has changed significantly: high-dimensional data have become pervasive in society alongside improved methods and increased computational power. These advances have improved inference and simulation capabilities and present opportunities to ask new scientific questions.

Stodden noted that improved transparency in scientific computing will allow researchers to run more ambitious computational experiments at the same time that better infrastructure for computational experiments will allow researchers to be more transparent. She anticipates that new, efficient infrastructure in research environments, workflow systems, and dissemination platforms will enable both transparency and reproducibility. Even in a modern computational environment, Stodden explained, it is still possible to achieve Boyle's vision for transparent scientific practice. She suggested that contemporary researchers frame reproducibility in three ways—empirical, statistical, and computational.

Applying this expectation for practicing transparent science to the notion of teaching data science, Stodden commented that effective data science curricula would include training in computational methods and tools as well as in theory and computational techniques. She suggested thinking about both tool and curricula development in terms of the data life cycle (i.e., acquire, clean, use, reuse, publish, preserve, and destroy). Kolaczyk asked how faculty could modify their curricula based on the data life cycle. Stodden responded that using the data life cycle as a guide highlights where knowledge gaps exist and where new courses can be added in programs to address such gaps. Deb Agarwal, Lawrence Berkeley National Laboratory, elaborated that students should be trained to understand data as approximations of facts by considering how data sets are generated, examining uncertainty and underlying errors, and

evaluating how errors could affect algorithms. In response to a question from Timothy Gardner of Riffyn, Stodden said that the audience for her data science curriculum includes any student who wants to work in any aspect of the data life cycle—from departments of statistics, computer science, information, and library science, for example. She added that classes with the word "data" in the title are so popular that it would be useful to begin to refine the curricula appropriately for students who plan to enter industry or to continue in academia, respectively.

Jessica Utts, University of California, Irvine, inquired about the emerging practice of registering analysis plans with journals in advance of submission. Stodden replied that preregistration would not be needed if the right infrastructure for reproducibility were in place—for example, allowing any statistical tests performed during an experiment to be tracked—and she suggested the design of appropriate tools as an effective solution. Peter Norvig, Google, supported the notion of developing computing infrastructure to enable reproducible research and suggested disaggregating steps along the scientific life cycle. Stodden believes that such practices will be developed both for ethical reasons and out of necessity—it is difficult to train one person to be an expert in multiple areas of the data life cycle—and will lead to increased collaboration among researchers.

Mark Green, University of California, Los Angeles, asked how the framework Stodden described could be applied across domains. Stodden responded that the framework is narrowly defined to respond to the challenges that have emerged from the increase in computation-enabled research. Mechanisms for verification, validation, and uncertainty quantification will vary depending on the setting. Green asked how to conceptualize computational reproducibility given that many algorithms are randomized. Stodden replied that some randomizations can be deterministically repeated, but she is researching how uncertainty is influenced by the computational instrument itself. She explained that linking computation to scientific application is not a solved problem. Bill Howe, University of Washington, observed that the details of the computation or the exact code fail to capture the full nature of reproducibility. If the findings documented in a paper are so sensitive to even small changes in computing environments, they may not be generalizable to other contexts. Stodden agreed that generalizability is the end goal; she added that computational reproducibility is a subset of this issue, and transparency is a key part of the process.

## TEACHING REPRODUCIBLE DATA SCIENCE: LESSONS LEARNED FROM A COURSE AT BERKELEY

*Fernando Perez, University of California, Berkeley*

Perez opened his presentation with a description of reproducible research from Buckheit and Donoho (1995): "An article about computational science in a scientific publication is not the scholarship itself; it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures" (p. 5). Perez embodies this notion in his course Reproducible and Collaborative Data Science (STAT 159/259)[2] at the University of California, Berkeley. Cross-listed as an undergraduate and a graduate course in the Department of Statistics, participants are required to have completed courses on computing with data, probability, and statistics prior to enrolling in the course. Though many of the participants are majoring in statistics, the course attracts students from across the campus. The course most recently enrolled 50 undergraduate and 10 graduate students, who completed weekly readings, quizzes, homework, and three hands-on projects each, under the guidance of Perez and a graduate teaching assistant. The course focuses on data access, computation, statistical analysis, and publication as a way to underscore that reproducibility is an essential tenet of modern computational research. The course introduces the social and scientific implications of a lack of reproducibility, and students learn that reproducibility is an everyday practice that requires the development of skills and habits. Core skills include understanding version control, programming, process automation, data analysis, documentation, software testing, continuous integration, and the use of data repositories.

The course uses the Jupyter Notebook, which allows the combination of text, code, and mathematical language in a single document accessible via a web browser. The students' work environments include a personal installation on each of their devices, using Anaconda[3] for dependency management, and an installation hosted by the Department of Statistics; this mimics real-world settings in which data science practitioners may have to use remote servers or the cloud. The course and its materials can be accessed via GitHub,[4] which provides a natural workflow for content management. Working with this software allows students to develop habits for good "computational hygiene," according to Perez. Students

[2] The course website is https://berkeley-stat159-f17.github.io/stat159-f17/, accessed February 13, 2020.

[3] The website for Anaconda is https://www.anaconda.com/, accessed February 13, 2020.

[4] The website for GitHub is https://github.com/, accessed February 13, 2020.

learn how to automate tasks with the Make tool, using tutorials developed by Software Carpentry,[5] as well as how to do continuous integration with validation using Travis.[6] Students attempt to replicate real-world research in their first hands-on projects; then, they develop a practical "playbook" for reproducible research and use Binder[7] to share a live, executable version of their completed work. For their final course project, Perez explained, students find their own data and conduct and document their analyses using the playbook they created earlier in the course.

Before concluding his presentation, Perez noted that the University of California, Berkeley, has other data science courses targeted toward first-year (Data 8[8]) and upper-division (Data 100[9]) students, which rely on interactive Jupyter Notebooks and are some of the fastest growing courses in the university's history. In summary, Perez described the tenets of a successful data science course: an actionable template for reproducible research, adequate coverage of tools and skills, a heterogeneous group of students (in terms of computational background), applicability of skills to multiple disciplines, and experience with real-world problems and data.

Brandeis Marshall, Spelman College, asked how such a course could be adapted on a smaller campus without similar staffing capacity. Perez noted that discussions are under way with the National Science Foundation's big data regional innovation hubs to address this issue. Stodden noticed that many of the tools Perez uses in his course come from outside the academic community and have been repurposed for scientific work. She wondered whether this trend should continue or whether the academic community should shift research and funding priorities to develop its own tools. Perez responded that while it makes sense for the academic community to develop its own tools in the case of specific research questions, much is gained from establishing industry partnerships and integrating industry-developed tools. He noted that it is important for students to be comfortable with a variety of tools, not just those found in academia, because many students seek jobs in industry after graduation. Kathleen McKeown, Columbia University, asked whether computer science and statistics should be taught separately or in blended data science courses. Deborah Nolan, University of California, Berkeley, replied that students benefit more when courses are co-taught and the content is

[5] The website for Software Carpentry is https://software-carpentry.org/, accessed February 13, 2020.

[6] The website for Travis is https://travis-ci.org/, accessed February 13, 2020.

[7] The website for Binder is https://mybinder.org/, accessed February 13, 2020.

[8] The course website for Data 8 is http://data8.org/, accessed February 13, 2020.

[9] The course website for Data 100 is https://data.berkeley.edu/education/courses/data-100, accessed February 13, 2020.

integrated because they can learn more about how to use computational skills in the context of data analysis.

## REPRODUCIBLE MACHINE LEARNING—THE TEAM DATA SCIENCE PROCESS

*Buck Woody, Microsoft Research and AI*

According to Woody's survey of practicing data scientists, teamwork among individuals with varied expertise is becoming essential in the workplace to better solve problems. Survey participants also observed that while practicing data scientists have established processes for data mining—for example, based on the CRISP-DM framework (see Chapman et al., 2000)—recent graduates entering the workforce are not familiar with such processes, in part because many undergraduate projects use only clean data. Furthermore, many organizations also utilize project plans to complete and monitor their business processes, and they expect data science projects to align with corporate platforms and practices.

Woody emphasized the need for a formal process in data science in which each participant considers all other project life cycle steps, including the needs of the end user. Implementing a standard process eliminates problems, motivates repetition, fosters communication, encourages collaboration, enhances security, and allows encapsulation of experiments. Woody described Microsoft's Team Data Science Process methodology that aims to improve team collaboration and learning:

- During the first phase of this process, *business understanding*, the team defines objectives and identifies data sources. Woody explained that defining a problem is one of the most difficult aspects of data science practice, and he added that many problems are not best solved with machine learning.
- During the second phase, *data acquisition and understanding*, scientists ingest, explore, and update the data.
- The third phase, *modeling*, encompasses feature selection as well as the creation and training of a model.
- The fourth phase is *deployment*.
- The final phase focuses on *customer acceptance*, which includes testing and validation, hand-off, retraining, and rescoring (Microsoft Azure, 2017).

A comment from Gardner highlighted the undercurrent of need-driven development in the Team Data Science Process and emphasized that product failure drives the desire for reproducibility. He described

this business motivation as very different from that in academic research. Woody agreed that a distinction exists between scientific reproducibility and industry reproducibility, because the latter is focused on finding a solution to a problem rather than repeating an experiment. Woody suggested that students be exposed to industry reproducibility so as to better prepare them for future workplace opportunities. Howe suggested an additional life cycle, specifically for a research question—the aspect with which students and scientists most often struggle. Woody noted that the Team Data Science Process includes a subprocess for defining the problem, a step in which domain expertise is crucial. Green advocated for new training opportunities that include industrial internships for students. Such experiences allow students to understand problem solving both in terms of a customer's needs and a business's objectives. Woody described such an internship program at the University of Washington that paired students with data scientists at Boeing. He also described an effective partnership in which the University of Washington paired students with nongovernmental organizations to work on specific problems. Kolaczyk highlighted similar alternative learning mechanisms at Boston University and Cornell Tech. Stodden asked whether the Team Data Science Process helps to increase efficiency, especially in instances of employee turnover. Woody noted that these issues are monitored and addressed within the development and operations framework of the process.

## OPEN DISCUSSION

### Incentive and Reward Structures

Nicholas Horton, Amherst College, wondered how incentive structures in academia could be modified to encourage faculty to teach data science courses and to develop data science tools. Gardner described the fundamental difference between incentive structures in academia (e.g., publishing results and earning grants) and in industry (e.g., creating products that work for a customer). The incentive structure in industry better drives collaboration and innovation, Gardner explained. Agarwal said that the reward system in academia has not yet emphasized team-based investigation over individually driven investigation. She suggested working to prioritize team-based investigation in the culture and in the practice of science by giving appropriate credit to everyone who participates in any part of the research and analysis process. She also noted that the people involved in reproducible scientific research are just as important as the mechanisms of reproducibility because myriad decisions get made over the course of an analysis. She encouraged recognizing and rewarding people for the contributions they make, specifically in

the middle of their careers. Green and McKeown added that curriculum development is not incentivized or rewarded as much as it could be at many public research universities. Green suggested encouraging faculty to develop contracts with their deans that formalize reward structures for course development as well as educating students early about the importance of team-oriented and goal-oriented approaches so as to begin to change the culture.

Perez noted that software artifacts do have intellectual value and thus deserve to be recognized accordingly. He encouraged developing a more relevant definition of intellectual value that also emphasizes teamwork. Tracy Teal, Data Carpentry, added her support for revised incentive structures. She objected to the current framework of "service" that exists around software development in academia—software is indeed a "research" product. If the incentive structure does not change, Teal cautioned, those individuals who develop software in academia may seek new employment in industry, where they will receive the recognition they deserve. Duncan Temple Lang, University of California, Davis, noted that software development that allows experimentation and brings in new ideas deserves to be rewarded but that not all software development fits in this category. He advocated for educating faculty on different types of software and redefining incentive structures. Mark Tygert, Facebook Artificial Intelligence Research, encouraged academic institutions to promote individuals with "nonstandard" résumés. Nolan suggested that faculty consult their institutions' academic personnel manuals: the language is often broad enough to encompass creative development of products and educational materials, and so faculty can be more proactive in making a case for promotion.

### Reproducible Research

Marshall commented that different audiences (i.e., undergraduates, graduate students, professionals) have varied needs and will benefit from diverse approaches to data science education, which will continue to evolve alongside emerging tools and software. Alfred Hero III, University of Michigan, encouraged the roundtable to think about the relationship between teaching students best practices for reproducibility and teaching students about ethical behaviors. Perez added that his students learned much about this relationship from discussing real-world cases with massive social impacts. Kolaczyk said that perfect reproducibility is difficult and occasionally impossible, so discussions of limitations may be necessary.

Antonio Ortega, University of Southern California, noted that the nature of software is changing. He suggested that deep learning systems

be treated as experiments so as to better capture the process of arriving at a result, thus enhancing reproducibility. Green commented that conversations about reproducibility should also include discussions of Bayesian techniques. Tom Treynor, Treynor Consulting, explained that it is more exciting to use science to predict the future than to retrospectively evaluate whether a finding is reproducible. He wondered why one would focus on preregistering an analysis instead of demonstrating, for example, the reproducibility of the result. He added that most trained scientists using data of all sizes could not provide a good definition of reproducibility (e.g., getting the same result in a predicted window), and he noted the importance of educating students about confidence intervals instead of p-values.

## TRAINING AS A PATHWAY TO IMPROVE REPRODUCIBILITY

*Tracy Teal, The Carpentries*

Teal described an increasing awareness around the need for reproducibility in research as well as a new appreciation for working reproducibly. She noted that working reproducibly requires additional computational and data science skills and novel ways of working, which can be a difficult shift for people to make. To be successful, researchers would need to connect the theory of reproducibility with practical skills and application. In other words, reproducible research emerges from the combination of a motivated researcher and relevant training. According to a survey of NSF principal investigators in biology, the majority of them are eager to learn new data analytics skills (Barone et al., 2017).

Because data are pervasive, it can be difficult to scale training alongside data production and to reach all audiences. For those already in the workplace, graduate students, or active researchers, Teal suggested (1) training "in the gaps," (2) developing collaborative and open educational resources, and (3) building communities of practice. She described successful training as

- Accessible for all learners, in all locations, for a reasonable duration;
- Approachable no matter the knowledge level, by creating an empowering, respectful, and motivating learning environment with faculty who understand educational pedagogy;
- Aligned with domain interests and current needs; and
- Applicable to people's current job tasks.

These four goals can be achieved, according to Teal, by revising existing courses, hosting short courses and workshops, developing massive open online courses, or offering just-in-time training. Teal suggested that

educational resources be built collaboratively, reused, and continually updated. Based on her experience, these materials would be most useful if made discoverable and open, and they are most effective when aligned with the needs and goals of the individual learners.

Teal explained that the Carpentries[10] is a "non-profit organization that develops curriculum, trains instructors, and teaches workshops on the skills and perspectives to work effectively and reproducibly with software and data." The Carpentries offers 2-day active learning workshops led by trained instructors. In these workshops, students receive formative feedback, have opportunities to collaborate with one another and with instructors, and develop skills applicable to data workflow and software development best practices. Teal recounted that the Carpentries hopes that students recognize the possibilities for data-driven discovery, develop confidence in using computational and data science skills, and will continue learning upon completion of a workshop. The Carpentries has hosted more than 1,300 workshops on seven continents with 1,300 volunteer instructors for 35,000 learners. Teal noted that the Carpentries conducts both short- and long-term pre- and post-workshop surveys to gauge participant interest and success. Responses to these surveys indicate that, overall, students have improved their attitudes toward reproducible research and use the skills they have acquired on a regular basis.

In response to a question from Woody, Teal said that the Carpentries recently created a new data curriculum to meet the needs of more entry-level learners. In response to a question from McKeown about the Carpentries' cost model, Teal noted that the nonprofit organization previously had a grant from the Gordon and Betty Moore Foundation and currently has a grant from the Alfred P. Sloan Foundation. They also support operations through a Member Organization and workshop fee model. Organizations can become Member Organizations at the Gold, Silver, or Bronze level for instructor training and workshops to build local capacity for training. Individual sites can request a workshop for a $2,500 workshop coordination fee, and fee waivers can be available.

## PERSPECTIVES ON ENHANCING RIGOR AND REPRODUCIBILITY IN BIOMEDICAL RESEARCH THROUGH TRAINING

*Alison Gammie, National Institute of General Medical Sciences*

Gammie explained that because issues of scientific rigor and transparency (especially in the field of biomedical research) are being discussed

[10] The website for the Carpentries is https://carpentries.org/, accessed February 13, 2020.

more frequently in the popular press, representatives of Congress are now paying more attention to the notion of reproducibility. Surveys conducted by *Nature* revealed a number of causes that contribute to irreproducible results, the top three of which are selective reporting, pressure to publish, and low statistical power or poor analysis (Baker, 2016). The biomedical research incentive structure, in particular, represents an underlying systemic factor that can affect reproducibility. Academic researchers are under constant pressure to secure funding, innovate, publish, and gain tenure. These issues are complicated by the fact that only 10 percent of National Institute of Health (NIH)-funded principal investigators receive greater than 40 percent of NIH funding, according to Gammie. Through its program called Maximizing Investigators' Research Awards,[11] the National Institute of General Medical Sciences (NIGMS) works to better distribute these funds among researchers and enhance scientific discovery.

Gammie described a case study in cell culture—highlighting issues of cell line contamination and misidentification, genomic instability, infections in stocks, and variability of growth conditions—to demonstrate the challenges of reproducibility in biomedical research. NIH is starting small business initiatives to develop inexpensive tools that can help authenticate biological materials and thus encourage more rigorous work. Another initiative involves drafting new grant guidelines, which focus on enhancing rigor and transparency by emphasizing premise, design, variables, and authentication in the review criteria.

Gammie explained that increased training is one pathway to enhance reproducibility. NIGMS developed a clearinghouse[12] for new training resources that contribute to rigor and transparency, as well as multiple funding announcements to develop training modules in enhanced reproducibility or local courses in experimental design and analysis. NIH also offers a predoctoral training grant program[13] to ensure that rigor and transparency are threaded throughout the graduate curriculum and reinforced in the laboratory. The principal investigator and program faculty on these grants are required to have a record of doing rigorous and transparent science and to submit a specific plan for how the instruction will enhance reproducibility. Such programs will help trainees develop the technical, operational, and professional skills needed to enter the

[11] The website for the Maximizing Investigators' Research Awards program is https://www.nigms.nih.gov/Research/mechanisms/MIRA/Pages/default.aspx, accessed February 13, 2020.

[12] The website for the NIGMS clearinghouse is https://www.nigms.nih.gov/training/pages/clearinghouse-for-training-modules-to-enhance-data-reproducibility.aspx, accessed February 13, 2020.

[13] The website for the NIH predoctoral training program is https://grants.nih.gov/grants/guide/pa-files/PAR-17-096.html, accessed February 13, 2020.

biomedical research workforce. Gammie emphasized the need for academic institutions to recognize training and mentoring activities in tenure and promotion packages and to decrease the pressures on principal investigators that negatively impact the research culture. Gammie concluded by reiterating that rigor and transparency, responsible and safe conduct of research, and diversity and inclusion are integral to excellence in training.

In response to a question from Stodden about the lack of reference to software in the description of the training grant programs, Gammie noted that software could be covered in areas of data analysis and interpretation, but institutions should provide input to funding agencies on what skills are needed in data science training and write them into their specific aims. The funding agencies will then support training in those areas and hold the institutions to the standards they set for themselves. McKeown mentioned that while training grants are pervasive in biomedical research, few equivalent opportunities exist in other domain areas. Gammie replied that NIH training programs are becoming more interdisciplinary as the scientific culture changes and as graduate studies become increasingly interdisciplinary. Hero said that NIH's role in funding data science training and research is uncertain given that its Big Data to Knowledge initiative has ended. Gammie encouraged data scientists who can demonstrate a robust training program that meets the basic science mission of NIGMS to continue to apply for training grants, as many fundamental skills cross disciplines. Hero suggested that it would be useful if predoctoral data science training programs had funding for and openness toward application areas. Kolaczyk noted that it remains to be seen where computational infrastructures fit in the broader scientific view of reproducibility as well as in the larger ecosystem of training grants.

## BURIED IN DATA, STARVING FOR INFORMATION: HOW MEASUREMENT NOISE IS BLOCKING SCIENTIFIC PROGRESS

*Timothy Gardner, Riffyn*

Gardner commented that it is important to bridge the gap between industry and academia. Riffyn's mission is to help scientists deliver reusable data and trustworthy results. He emphasized the value of focusing on the fundamental causes of irreproducibility rather than the symptoms, and he explained that researchers are failing to harness reproducibility lessons learned more than 50 years ago and apply them to scientific research. More than $420 billion is spent on research and development globally each year, and, if even 25 percent of the results are irreproducible, $105 billion will be lost each year. He continued that researchers hope to achieve a world

of science in which published results can be built upon, but this goal has not yet been realized, primarily because researchers spend 80 percent of their time cleaning and organizing data instead of learning from them. He categorized data-related challenges in research and development in terms of data quality, access, integration, interpretation, and system flexibility. Gardner agreed with Agarwal that data are only approximations, not facts. Clean data begins with quality experiments, and it is important to teach principles, develop tools, and build a culture of quality in research and development throughout foundational undergraduate curricula.

He presented multiple examples of data evaluation and quality assurance efforts that lead to improved reproducibility and productivity in biotechnology processes, although the problems and principles are generalizable. Gardner worked with a company identifying new cell lines for further development, but the high level of noise and variability in assay results, even when looking at only a single cell line, prevented any significant conclusions about the relative performance of different lines. In another case, he described how better tracking and control of variables, including factors such as temperature and the choice of growth medium, explained why so few candidate strains had been proven to be more effective than the control. Gardner found that scientists must control and qualify their assays before applying them. In another example, he described a company's attempt to massively scale up a fermentation process using an engineered yeast but was stuck in part because of high levels of noise and variance in assays. Reducing the error in measurements allowed the company to identify the critical parameters that had to be maintained and ultimately enabled it to scale up manufacturing while maintaining performance. His final example of how data quality assurance and control can drive process improvement featured a company that reduced the relative error of its assays sixfold, which allowed it to reproducibly identify and build upon small incremental improvements that were otherwise lost in the noise. This doubled the rate of strain improvement, and Gardner described this as a paradigm for reproducible science—if each individual can make an incremental improvement, society can make scientific discoveries much faster.

He reiterated the value of learning from history. For example, the automobile industry recognized that reduced decision-making error through improved data quality assurance accelerates manufacturing and improves results. Valuable best practices of manufacturing quality can be transferred to scientific research and development, including designing experiments, measuring, analyzing and improving the experimental process, sharing, and iterating.

Howe questioned the analogy of scientific research to the automobile manufacturing process—it is difficult to transfer lessons about

reproducibility because the two contexts are so different. He also explained that he would rather have access to noisy, unstructured data, which prompt further innovation, than rely on "complete, accurate, and permanent data." Gardner responded that the examples he shared depended on determining the reliability of the assays. While important steps such as these can add time, he asked, "Would you rather have a result that you can't trust or take an extra week to qualify an assay?" Gardner added that he does not advocate that data be withheld from analyses but rather that all data used are appropriately qualified and linked to the various experimental parameters across the chain. Treynor explained that signal-to-noise ratio in many industry experiments is on the same order as the accuracy of the measurement systems, further motivating the adoption of the automobile industry's best practices. He added that he prefers structured data no matter how good or bad they are, but fundamental principles of data management and organization are not currently taught in enough depth to accommodate this preference.

Hero emphasized the importance of teaching data science students to consider the data collection process and the potential value of metadata. Gardner noted that "metadata" is a misleading term—metadata are of utmost importance and should be structured so that statistical learning, machine learning, and regression analyses can be applied to better understand their relationship with the primary data. Teal commended Riffyn for its work to improve data quality and observed that its incentive structure helps achieve that goal. She described a specific challenge in the genomics arena: because the data users are not data producers, they cannot easily impact data quality. Gardner said that that problem is universal: if no consumer exists to determine when a product is inadequate—and many academic products do not have direct consumers—no pressure exists to improve it. Green noted that although reproducibility of experiments and reproducibility of data analyses may have different challenges, they do overlap in the role of domain knowledge.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Roundtable participants divided into two groups to discuss key questions that emerged earlier in the day. On behalf of his group, Green summarized discussions in response to the following questions: *How could reproducibility be taught within a particular course or program? What are the implications of resource limitations, class size, teaching structure, and other incentives? Should reproducibility be taught on its own or integrated into other topics?* Green described his group's discussion of how to balance programming with statistics education in data science courses. In

reflecting on Perez's presentation, Green posited that perhaps only one-fifth of the curriculum would focus on programming, while the remainder would focus on issues such as testing and validation. He added, however, that such a curricular decision would vary by audience and that several approaches such as the following exist:

- Create a prerequisite course sequence with programming and software engineering before data science;
- Require a data literacy course (e.g., Data 8 at the University of California, Berkeley) as a prerequisite to a data science course;
- Eliminate introductory computer science courses and replace them with data literacy courses; and
- Develop a course that enables data literacy at the level of dialogue as opposed to a course that attempts to teach mastery.

The group also discussed the potential for institutions with large, established programs to provide packages to help institutions with limited staffing to implement such courses and make data science more widely available. Green emphasized that even with such tools and resources, faculty members need a certain level of training and knowledge, and graduate student instructors play a crucial role. He suggested that national funding could support programs for graduate students to assist undergraduate students at other institutions remotely. Another suggestion included developing a GitHub for teaching materials. The group considered whether chemistry, biology, and economics departments should each have their own data science courses. Green noted that one option could be to have a required core course that includes foundational knowledge in statistics and computer science. This model could unfold as a foundations course with additional sessions that teach data science tailored to particular domains (similar to the connector courses at the University of California, Berkeley). He continued that online courses could serve as bridges for people in other disciplines and for students enrolled in smaller colleges and that different classes can be combined to satisfy prerequisites. Green noted that the group discussed the need for reproducibility of analysis to be taught in an integrated fashion, although he added that reproducibility of data is somewhat domain-dependent and may need to be taught independently. The group's last topic of discussion considered how much preparation time is needed to become a well-trained data scientist. Green commented that the time would be substantial as well as dependent upon the needed technical depth and the rapidly evolving world of data science.

On behalf of her group, Agarwal summarized discussions in response to the following question: *Key factors (such as software system development*

*and statistical uncertainty estimates) may contribute to reproducibility challenges. In which ways can data science education be modified to make the most impact?* She noted that her group chose to discuss this question from the perspective of the entire data life cycle because reproducibility is truly a life cycle problem. She referred to the notion highlighted in Woody's presentation about understanding and considering issues that surround an analysis or another single component of the data life cycle. Agarwal also noted that her group was inspired by Gardner's reflections on the evolutionary aspect of reproducibility—students have to be taught to understand that achieving reproducibility is not a one-step process; rather, it is gradual evolution. She highlighted academic programs that incorporate consulting as a way for students to begin to recognize the value of these processes. Agarwal's group noted that although conversations about reproducibility and the data life cycle often focus on the data producers and the first users of data and analyses, the decision maker is also a critical part of the process. Stodden shared her approach to teaching students about reproducibility: Students first work in pairs to try to reproduce results from literature. Later in the semester, students will try to reproduce the results of their partners' outputs in the class and write a memo about this experience. This adds an instructional component on the process of peer review and the value of professional communication about research. Agarwal reiterated that such personal experiences are often effective for students to learn and become more conscientious about the challenges of reproducibility.

# 8

# Meeting #7: Programs and Approaches for Data Science Education at the Ph.D. Level

The seventh Roundtable on Data Science Postsecondary Education was held on June 13, 2018, at the National Academy of Sciences Building in Washington, D.C. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to explore the content and organization of new and emerging data science Ph.D. programs and to discuss alternatives for structuring Ph.D. programs, including stand-alone degrees, domain-based concentrations, and activities begun under the National Science Foundation's (NSF's) former Integrative Graduate Education and Research Traineeship program. 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.

Welcoming roundtable participants, co-chair Kathy McKeown, Columbia University, noted that while many universities have focused on the development of undergraduate- and master's-level data science education, fewer Ph.D. programs in data science have been established. She emphasized the value of discussing curriculum requirements, levels of interdisciplinarity, departmental designations, institutional barriers, degree types, and research opportunities when evaluating or developing Ph.D. programs.

## THE PH.D. PROGRAM IN DATA SCIENCE AT NEW YORK UNIVERSITY

*Vasant Dhar, New York University*

Dhar explained that New York University's (NYU's) Center for Data Science (CDS)[1] was created in 2012 with support from representatives across campus. By creating a separate unit, NYU demonstrated its commitment to data science as a distinct area of study that integrates many disciplines. Although NYU ultimately plans to create full professorships in data science, current faculty appointments are joint between data science and another department.

NYU's Ph.D. program in data science admitted its first cohort—4 students—in 2017. From a well-qualified applicant pool of 400, the 2018 cohort includes 15 students who are diverse in geographic region, gender, and academic discipline. While all applicants had uniformly high quantitative Graduate Record Examination (GRE) scores, admitted applicants had higher verbal GRE scores. He emphasized the added value of strong written and verbal communication skills as well as the ability to conduct scientific inquiry as preparation for data science study. The Ph.D. curriculum is structured in a way that blends engineering and social science and gives students flexibility and time to develop a thesis topic and find an appropriate advisor. The curriculum requires five core CDS courses—Introduction to Data Science, Probability and Statistics for Data Science, Machine Learning, Big Data, and Inference and Representation—and a multitude of electives from across the university. Over the course of the program, students participate in formal research rotations with faculty, take a qualifying exam and a comprehensive exam, and complete a dissertation. Dhar expects that the curriculum will continue to evolve in the future, in part driven by new faculty developing courses in their areas of expertise.

Daniel Spielman, Yale University, asked how NYU determines whether students need certain courses. Dhar explained that students can take placement exams, but he would prefer to see those decisions made by faculty on a case-by-case basis. In response to a question from Nicholas Horton, Amherst College, Dhar said that the Ph.D. program's five core courses are also offered at the master's level. James Frew, University of California, Santa Barbara, asked about the workplace experience of NYU's Ph.D. students, and Dhar estimated that at least half enter the Ph.D. program directly after completing a bachelor's or master's program. He

[1] The website for the Center for Data Science is https://cds.nyu.edu/, accessed February 13, 2020.

noted, in response to a follow-up question from an audience participant, that students in NYU's data science Ph.D. program are funded by a combination of university and external fellowships.

Jeffrey Ullman, Stanford University, asked how a Ph.D. in data science compares to a Ph.D. in computer science for a student seeking employment in artificial intelligence. Dhar responded that if such a student is sufficiently motivated, the student could attain the equivalent training with the Ph.D. in computer science as well; however, the interdisciplinary nature of NYU's data science program gives students a broad exposure across methods and domains and leads to research questions they might not ask in a typical computer science department. Jeffrey Brock, Brown University, posited that the differentiating factor between the Ph.D. programs in data science and computer science could be mathematical foundations. Dhar commented that while some differences exist in the types of mathematical foundations in each program, more substantial differences can be found in the overall breadth of problem types that one encounters in a data science program, which can lead to methodological innovation. In response to a question from an audience participant, Dhar said that the Ph.D. programs in computer science and data science at NYU require the same total number of credits.

Devavrat Shah, Massachusetts Institute of Technology, wondered how faculty members balance their time developing courses for data science and teaching in their home departments. Dhar noted that currently those types of decisions are negotiated by the provost and the dean, although such processes will likely become formalized in the future. Despite the burden placed on faculty to contribute in both areas, Dhar reiterated the value of collaborating across disciplines and the excitement of working in an emerging field. Charles Isbell, Georgia Institute of Technology, asked how NYU manages culture clashes commonly found in interdisciplinary programs. Dhar replied that CDS has a positive outlook and has thus far avoided such clashes; participants acknowledge the value of interdisciplinarity and appreciate what they can learn from one another. In response to a question from Abani Patra, University at Buffalo, Dhar said that faculty with joint appointments will be reviewed and evaluated for tenure by both the home department and CDS.

## YALE'S PH.D. PROGRAM IN STATISTICS AND DATA SCIENCE

*Daniel Spielman, Yale University*

Spielman explained that Yale's Department of Statistics became the Department of Statistics and Data Science in 2017 and hosts both an undergraduate major and a Ph.D. program. The Ph.D. program is

structured in a way that reflects this evolutionary approach. To foster interdisciplinarity, some new faculty hires at Yale are being offered "half slots" in the Department of Statistics and Data Science; although resources and responsibilities come from both the Department of Statistics and Data Science and the faculty member's home department, the faculty member completes the tenure process only in the home department. The Department of Statistics and Data Science also offers secondary faculty appointments, which provide opportunities for collaboration on student data projects without teaching obligations from the department.

Yale's Ph.D. in statistics and data science requires 12 courses, which help to define what it means to be a data scientist and to create a common culture among students practicing data science. Spielman noted that the Ph.D. program should take students approximately 5 years to complete, 2 of which will be dedicated to coursework. Requirements include a course and a qualifying exam in probability; a course and a qualifying exam in statistics; coursework in computation; studies in practical data analysis;[2] and a research oral exam. In response to a question from Ullman about whether requiring a qualifying exam in statistics but not computation emphasized data analysis over problem solving, Spielman explained that the coursework requires successful problem solving, as does the practical data exam. Although Ph.D. students can choose advisors from other departments, the thesis is supervised at least in part by a member of the Department of Statistics and Data Science.[3]

Yale plans to increase the size of the incoming class of Ph.D. students from four to six and to revise the grant structure for students. Spielman commented that once a truly coherent culture is developed in the Ph.D. program, the Department of Statistics and Data Science might scale back course requirements as well as consider an alternative name that would better embrace the broad spectrum of data science. Brock highlighted the important roles that administrators and funding agencies play in making these programs successful. A Ph.D. program in statistics and data science may motivate faculty to collaborate beyond their disciplinary silos, which is crucial for the future of science. Further, NSF is creating conduits for graduate students to work in a domain area and data science, as well as promoting discussions across university boundaries. He added that establishing industry–university partnerships is essential as the data science landscape continues to evolve.

[2] The studies in practical data analysis include a case studies course, a practical exam with a data problem that must be solved within 1 week, and practical work through a semester-long project with a faculty member in another department.

[3] The website for the Department of Statistics and Data Science is https://statistics.yale.edu, accessed February 13, 2020.

Alfred Hero III, University of Michigan, asked about industry's perspective of a Ph.D. in statistics and data science. Spielman said that industry has high demand for students with undergraduate degrees in statistics, computer science, and applied mathematics, so he expects the same to be true for Ph.D.'s in statistics and data science because they further develop these skill sets. Alok Choudhary, Northwestern University, asked whether the Ph.D. program teaches students how to build scalable software, and Spielman explained that individual graduates will emerge with varied skills and strengths. This will best prepare them to be productive members of data science teams in the workplace, he continued. Philip Bourne, University of Virginia, emphasized the importance of breaking down traditional disciplinary silos and transferring best practices across departments and institutions, both in the United States and abroad. Spielman agreed that it is important to engage faculty from other departments and universities to create intellectual diversity and introduce new methods.

## INTRODUCTION TO STATISTICS AND DATA SCIENCE AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

*Devavrat Shah, Massachusetts Institute of Technology*

Shah described the Statistics and Data Science Center (SDSC),[4] which is part of the Massachusetts Institute of Technology's (MIT's) Institute for Data, Systems, and Society,[5] as an interdisciplinary academic center with the mission to advance statistics and data science programs and research activities across campus. The SDSC encourages connections with the social sciences, life sciences, and computational sciences.

The SDSC began offering an undergraduate minor in statistics and data science in 2016, professional education in data science in 2016, and an interdisciplinary Ph.D. in statistics[6] in 2018, and will launch an online micro-master in statistics and data science for professionals in fall 2018. Shah said that MIT hosts the interdisciplinary Ph.D. through its five schools—Architecture and Planning; Engineering; Humanities, Arts, and Social Sciences; Management; and Science—because students need to be trained in statistics, computation, and data science in order to be successful, and no single unit at MIT could achieve this. The Ph.D. program

[4] The website for the Statistics and Data Science Center is https://stat.mit.edu/, accessed February 13, 2020.

[5] The website for the Institute for Data, Systems, and Society is https://idss.mit.edu/, accessed February 13, 2020.

[6] The website for the interdisciplinary Ph.D. in statistics is https://stat.mit.edu/academics/idps/, accessed February 13, 2020.

is managed by an institute-wide standing committee, with representatives from and within each academic unit. Shah emphasized community-building as an essential part of the program, with weekly activities and annual events (e.g., SDSCon[7]) sponsored by the SDSC as well as a required semester-long advanced research seminar.

Shah explained that students must be admitted to a home unit before becoming eligible to apply to the interdisciplinary statistics Ph.D. program in a subsequent semester; that admission decision will be made by the home unit first and then by the institute-wide standing committee. In addition to course requirements from the students' home units, Shah continued, courses across four foundational areas (i.e., probability, statistics, computation and statistics, and data analysis) are required. While the probability and statistics courses share a common curriculum, the computation and statistics and the data analysis courses may vary across domains. Shah added that a student's thesis must be relevant to both statistics and data science in order to earn the interdisciplinary Ph.D.

In response to a concern from Mark Tygert, Facebook Artificial Intelligence Research, Shah said that prospective students are aware that acceptance into the interdisciplinary Ph.D. program is not guaranteed. Replying to questions from Isbell and Spielman, Shah noted that a graduate of this program would receive a degree that reads, "Ph.D. in 'X' and 'statistics and data science.' " Because of the community that is developed and the work that is required to complete the program, this degree signifies more than a "badge." Frew wondered about the administrative management of such a program, and Shah explained that the interdisciplinary program is relatively straightforward to manage because all units are provided a clear set of checkpoint guidelines. In response to a question from Choudhary, Shah commented that students can take courses in the interdisciplinary program without obtaining the interdisciplinary Ph.D. and added that qualification is determined by the individual units, not a centrally administered exam. Dhar asked what volume of students is expected for the program, and Shah replied that because the burden on students is substantial with six additional courses, only one or two students at a time are expected to apply to the interdisciplinary Ph.D. from each participating unit.

## OPEN DISCUSSION

### Diversity and Interdisciplinarity

An online participant asked how the diversity of students' backgrounds impacts the curricula of graduate programs. Dhar noted that

[7] The website for SDSCon is https://stat.mit.edu/calendar/sdscon-statistics-data-science-center-conference/, accessed February 13, 2020.

although NYU's Ph.D. applicants come from many disciplines, no formal pathways have been created. This decision will be evaluated as the program evolves. Spielman noted that the statistics Ph.D. at Yale historically accepted and trained students with diverse academic backgrounds, and he is hopeful that the same can be done for participants in the Ph.D. program in statistics and data science. Shah explained that because MIT's program is interdisciplinary by design, students are expected to be heterogeneous. He believes this level of diverse experience attracts students to the program and ensures the best contributions from each. Choudhary asked whether the thesis in each of these Ph.D. programs is driven by domain data. Spielman replied that acceptable theses come in many forms: some are driven by data, some develop methods, and others prove a theorem without data. Dhar and Shah emphasized that a Ph.D. in data science allows for broad inquiry. Horton suggested that roundtable members read the National Academies' report *Graduate STEM Education for the 21st Century* (NASEM, 2018a) and underscored the importance of implementing evaluation and assessment, fostering a community, offering faculty development, and promoting diversity and inclusion in emerging data science Ph.D. programs.

### Ethics and Curriculum Development

Lise Getoor, University of California, Santa Cruz, asked how these Ph.D. programs integrate responsible data science practice and data science ethics. Dhar said that CDS is collaborating with AI Now[8] on this issue; although courses that discuss such topics are already available, it would be beneficial to create a formal course requirement in ethical data science for the Ph.D. program. Spielman noted that Yale has begun a search to hire faculty with the expertise to integrate ethics into the program, but the university does not yet offer a formal course beyond what is covered in a graduate case studies class. Shah asserted the need to involve social scientists in this conversation. He added that although such topics have been introduced in some courses, there is no single course in the ethics of data science at MIT. Mark Krzysko, Department of Defense, emphasized that a framework is needed around social and domain norms for data science practice. Tygert encouraged people to engage Facebook in these conversations about ethics, as the company continues to explore similar questions. Brock described a master's program at Brown University that includes a course on data and society, whose students noted that they do not believe privacy is important. Because such students do not experience data and the world in the same way as faculty, faculty have

[8] The website for AI Now is https://ainowinstitute.org/, accessed February 13, 2020.

an added challenge in understanding how this dichotomy of viewpoints should affect course content and delivery.

## DATA SCIENCE AT THE UNIVERSITY OF CALIFORNIA, DAVIS

*Duncan Temple Lang, University of California, Davis*

Temple Lang shared that the University of California (UC), Davis, perceives data science as a distinct academic discipline that focuses on the process of data-enabled research; explores the breadth of the entire data pipeline; and integrates mathematics, computer science, and statistics. The data science curriculum focuses on applying data science across the domains as well as solving problems within the domains. UC Davis's goal is to engage and impact all disciplines from engineering to religious studies.

The Data Science Initiative[9] began at UC Davis in 2014, when the provost provided funding to explore the best structure for data science education, including collaborative research, community building across disciplines, training and consulting opportunities, and dedicated space in the campus library to connect people across diverse areas. Temple Lang explained that a new academic unit for data science, led by a multidisciplinary coalition of faculty, will be in place in the 2018-2019 academic year. This unit will provide an opportunity for a new perspective and culture in research and education and will serve as a complement to the mathematics, statistics, and computer science departments.

UC Davis will ultimately offer three types of doctoral study in data science: a Ph.D. in data science; a Ph.D. in computer science, mathematics, or statistics; and a Ph.D. in a domain discipline. The latter two are of greatest focus for UC Davis currently, Temple Lang noted, because they attract the largest number of students. To provide educational opportunities for these currently enrolled Ph.D. students, UC Davis plans to offer two types of add-on data science credentials: a "designated emphasis" and a "graduate academic certificate." Both credentials require an additional four courses: Survey of Statistical Machine Learning; Data Technologies and Computational Reasoning; an elective; and a capstone project. The designated emphasis also requires a data science-related thesis and qualifying exams. Both credential programs are especially attractive to students in computer science, statistics, mathematics, and domain sciences, Temple Lang continued, because they give students practice with real data science problems. Both programs prepare graduates who seek employment outside of academia as well as graduates who may want to teach data science in a discipline.

[9] The website for the Data Science Initiative (now known as DataLab: Data Science and Informatics) is https://datalab.ucdavis.edu, accessed February 13, 2020.

Temple Lang is often asked, "Why offer a data science Ph.D. if all Ph.D.'s use data to do science?" He reiterated that data science has a unique culture and concept; therefore, an academic home that emphasizes the data science process, the entire data science pipeline, and multidisciplinarity is essential. Such a home encourages students to engage in systematic research in workflows, data science problem-framing, computational environments for data analysis, data visualization, data sources and fusion, reproducibility, and ethics. He added that UC Davis is committed to its acceptance of interdisciplinarity—for example, faculty can advise students in many different Ph.D. programs beyond those in their home departments. In addition to developing the new academic unit in data science, the Ph.D. in data science, and the add-on Ph.D. data science credentials, UC Davis also plans to continue its complementary data science initiative as well as develop a data science undergraduate major, data science minors with varied foci, and a data science master's degree.

Ullman expressed his skepticism of data science as a unique intellectual domain. Temple Lang responded that the core of data science is the composition of the process: framing data science problems, enabling qualitatively different research in existing fields, and communicating about data. Although there is overlap in the content of data science and other disciplines, he continued, data science has a unique focus. In response to a question from Hero about the role of information scientists in building the Ph.D. program in data science at UC Davis, Temple Lang noted that although UC Davis does not have a school of information, faculty with such expertise could find a home in the new academic unit for data science. Brock commented that Temple Lang's systematic research topics frame data as primary and domains as essential; these are the types of topics with which a data science Ph.D. student would engage. Magdalena Balazinska, University of Washington, mentioned that as data science departments emerge, computer science and statistics departments are evolving—data science departments often play an important role in uniting all of these efforts.

## SOCIAL SCIENTIFIC DATA SCIENCE: BUILDING THE PENN STATE PH.D. IN SOCIAL DATA ANALYTICS

*Burt Monroe, Pennsylvania State University*

Monroe discussed Social Data Analytics (SoDA)[10] at Penn State, which aims to integrate social science and data science approaches to better understand human interactions. SoDA resulted from an NSF-funded

[10] The website for Social Data Analytics is https://bdss.la.psu.edu/soda/graduate-program-in-social-data-analytics-soda, accessed February 13, 2020.

Big Data Social Science Integrative Graduate Education and Research Traineeship Program (BDSS-IGERT),[11] and it hosts a dual-title Ph.D. (in cooperation with six departments), a doctoral minor, and a bachelor's of science degree. According to Monroe, with its focus on socially relevant problems, SoDA excels in attracting and recruiting women and underrepresented groups. Monroe provided a historical overview of the data science education-related efforts at Penn State that led to the development of SoDA, starting with the Social Science Statistics Partnership (SSSP) in 2004. This initiative began in an effort to raise the level of methodology within the political science and sociology departments. With funding from the College of Liberal Arts, SSSP expanded into the campus-wide Quantitative Social Science Initiative in 2006. The BDSS-IGERT grant of $3 million in 2012 allowed for 2-year academic research rotations in interdisciplinary projects and summer externships for students, initial plans to create the SoDA curriculum, and community building through the establishment of the "Databasement"—a central campus location where SoDA students meet.

Monroe explained that the dual-title Ph.D. in SoDA is structured to offer interdisciplinary programs without creating new departments, similar to the model used at MIT. Penn State's program differs from MIT's, however, in that it is possible for a student to be accepted simultaneously into the home department and the SoDA program. Students complete a series of requirements in their home disciplines and an additional four courses to satisfy SoDA requirements (e.g., two data approaches and issues seminars and two courses from approved options in analytical, social, quantitative, and computational sciences). Monroe discussed some program design challenges, including agreeing upon the number and type of requirements for the Ph.D. program, navigating boundaries between social science and non-social science disciplines that think about data in different ways, achieving true interdisciplinarity, and balancing the levels of faculty ownership for the program.

Horton wondered whether this model is similar to a data science + X degree program. Monroe responded that it is different because it extends beyond substantive engagement with domain theories to exploration of methodological approaches unique to social science. He noted that no one model of data science education will meet everyone's needs. Benjamin Ryan, Gallup, Inc., asked Monroe about his ideal relationship with industry, and Monroe said that SoDA has an industry advisory board and often invites industry speakers to campus. He emphasized that not all Ph.D.

[11] The website for the Big Data Social Science Integrative Graduate Education and Research Traineeship Program is https://www.nsf.gov/awardsearch/showAward?AWD_ID=1144860, accessed February 13, 2020.

graduates should become university professors; if a graduate secures employment in any position that requires Ph.D. training, Monroe considers that a success.

## DATA SCIENCE SPECIALIZATIONS IN PH.D. PROGRAMS AT THE UNIVERSITY OF WASHINGTON

*Magdalena Balazinska, University of Washington*

Balazinska explained that the University of Washington's (UW's) eScience Institute[12] was founded in 2008 and has become a permanent fixture on campus, owing to funding from UW and the Washington state legislature. As UW's neutral hub of data science activity, the eScience Institute strives to empower students and faculty to accelerate discovery and leverage data, no matter how complex. It aims to build community, further research, and educate. The eScience Institute includes more than 100 affiliated faculty from across the university as well as a number of postdoctoral and Ph.D. students from a 2013 NSF IGERT award, and it extends open office hours to anyone on campus with a data problem.

The motto of the eScience Institute is "data science for all," Balazinska continued. The eScience Education Working Group makes data science education available to any interested student through formal programs, short courses, domain-themed hack weeks, workshops, and seminars and encourages an interdisciplinary community for students. Because the students generally fall into two categories—those who want to use data science tools and those who want to build data science tools—varied educational approaches are needed.

Balazinska commented that UW's formal data science education programs include a Ph.D. in a discipline with either an "advanced data science option"[13] or a "data science option"[14]; an undergraduate degree with a data science option; a professional data science master's degree; and a variety of professional certificates. To enroll in either of the Ph.D. options, students are first admitted to their participating home departments[15]

[12] The website for the eScience Institute is https://escience.washington.edu/, accessed February 13, 2020.

[13] The website for the advanced data science option is https://escience.washington.edu/education/Ph.D./advanced-Ph.D.-data-science-option/, accessed February 13, 2020.

[14] The website for the data science option is https://escience.washington.edu/education/Ph.D./data-science-graduate-option/, accessed February 13, 2020.

[15] Departments of astronomy, chemical engineering, genome sciences, and psychology currently offer the data science option; departments of applied mathematics, astronomy, biology, chemical engineering, computer science and engineering, genome sciences, mathematics, oceanography, psychology, and statistics currently offer the advanced data science option.

and then can simply elect an option. The options are managed by the individual departments, although a single framework under the eScience umbrella is used and a central steering committee oversees the process. The advanced data science option, which is intended for data science tool developers, requires students to complete three of four courses in basic statistics, machine learning, data management, and data visualization, in addition to any home department requirements. Students also take four quarters of an eScience seminar, Balazinska said. In the data science option, which is designed for data science tool users, students and departments have a bit more flexibility in the course requirements. More than 60 students currently participate in the Ph.D. options. These Ph.D. options evolved out of an NSF-IGERT program that had additional requirements: IGERT students are co-advised by faculty in data science methods and in domain sciences, encouraged to participate in internships, and regularly attend seminars. Several networking activities are also available for students interested in data science, Balazinska explained, such as an annual retreat, student-led seminars, lunches, summits, program evaluation, and a career fair, all facilitated by the eScience Institute infrastructure and resources as the IGERT grant draws to a close.

In response to a question from McKeown, Balazinska confirmed that UW would like to expand its data science options in the humanities and social sciences. Replying to Atma Sahu, Coppin State University, Balazinska said that the core domain framework for both options was initially developed by the eScience Education Working Group and continues to evolve. Balazinska added that departments play a central role in developing and maintaining the options, with special consideration for issues of accreditation. An audience participant asked about prerequisites for the data science options, and Balazinska reiterated that the two levels of data science options target different audiences, depending on their needs and interests (i.e., some courses in the data science option have minimal or no prerequisites). Hero inquired about the interdepartmental partnerships that are required to develop successful Ph.D. programs in data science. Balazinska explained that UW tries to increase its capacity within individual departments by hiring additional faculty. She also said that UW has various departments teaching data science courses; as a result, departments are starting to specialize in certain areas and offer more courses.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Roundtable participants divided into two groups to discuss key questions that emerged earlier in the day. On behalf of his group, Bourne

summarized conversations surrounding the following questions: *From an employer's point of view, what are the anticipated advantages of a Ph.D. in data science in contrast to a Ph.D. in a domain? More broadly, as asked by Temple Lang, "Why [offer] a data science Ph.D. if all Ph.D.'s use data to do science?"* Bourne noted that because data science skills are in such high demand in industry, graduates of either type of program are likely to gain employment. A number of factors are important: if employers are seeking knowledge of the complete data life cycle (which he defined as acquisition, engineering, analytics, visualization, dissemination, ethics), a Ph.D. in data science would be more useful than a Ph.D. in a domain. Bourne observed that the unique cultures of different fields also play a role in educational preparation and hiring decisions—industry focuses on a product, academia focuses on knowledge creation, and government focuses on policy. The scope, scale, and topic of a particular project would also influence the type of knowledge and training best suited for success.

On behalf of his group, Frew summarized discussions in response to the following question: *Data science education at the Ph.D. level is multifaceted, and institutions are coming up with many different approaches. Is it possible to identify emerging best practices to common process challenges?* Frew identified three models for emerging Ph.D. programs: (1) a start-up (i.e., a new entity created with existing faculty); (2) an expansion of an existing entity; or (3) an overlay (i.e., data science superimposed on top of preexisting departments). Best practices may vary by model. No matter which model is chosen, Frew continued, it is vital to recognize that contributing disciplines have diverse perspectives and to react to those appropriately. Institutions themselves also have varying levels of ease in piloting new programs. Frew added that all three models would benefit from the inclusion of a physical space, independent from any specific department, which allows for cross-disciplinary interactions and collaborations at an appropriate level. When implementing new models, Frew explained, it is important for institutions to incentivize cross-disciplinary collaboration. For example, Stanford University allows faculty members to serve as advisors of record for Ph.D. students in any department. Frew emphasized that cross-disciplinarity should not be seen as a barrier to tenure.

# 9

# Meeting #8: Challenges and Opportunities to Better Engage Women and Minorities in Data Science Education

The eighth Roundtable on Data Science Postsecondary Education was held on September 17, 2018, at the Georgia Institute of Technology in Atlanta, Georgia. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss existing efforts in computing, statistics, and mathematics societies to improve core fields' engagement with underrepresented populations and to learn about several new programs focused on broadening participation in data science. 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.

Welcoming roundtable participants, co-chair Eric Kolaczyk, Boston University, described a profound lack of representation of women and minorities in science, technology, engineering, and mathematics (STEM) fields. He noted that tremendous challenges and opportunities exist to improve equity and diversity in STEM education programs and workplaces. He suggested that the emergence of data science, with its focus on new paradigms, has the potential to create a watershed moment to better engage women and minorities in STEM fields and beyond. The presentations and discussions that followed detailed best practices and possible strategies for creating opportunities in STEM for underrepresented populations.

## THE CONSTELLATIONS CENTER FOR EQUITY IN COMPUTING

*Kamau Bobb, Georgia Institute of Technology*

Bobb, along with Charles Isbell, Georgia Institute of Technology (Georgia Tech), developed the Constellations Center for Equity in Computing[1] in an attempt to address some of the structural challenges that students, particularly students of color, experience both in the city of Atlanta and throughout the nation. Despite the fact that computer science skills are central to decision making in a modern digital economy, Bobb noted a dearth of computer science educators in both the K-12 and postsecondary spheres—for example, Georgia has more than 528,000 students enrolled in public high schools but only 93 teachers certified to teach computer science. With low teacher pay, limited professional development opportunities, and industry pull for recent college graduates, this educator shortage will likely continue even as student interest in computer science education increases, he explained. In response, the Constellations Center built a structural tool to deliver computer science content through a hybrid infrastructure: skills are delivered online, and classroom teachers facilitate learning. This model has the potential to increase equitable access to computer science skills for minority and low-income students.

Bobb described inequities in access to computing education and their impacts on undergraduate enrollments of underrepresented minorities. While Atlanta's population is greater than 50 percent African American, only three African American students are enrolled in Advanced Placement computer science courses in local public high schools, and this population is similarly underrepresented in Georgia Tech's College of Computing, according to Bobb. This year, three fellows from the Constellations Center are going into six public high schools in Atlanta to teach Advanced Placement Computer Science Principles to the students, while the classroom teacher observes. In the future, a virtual course will take the place of the fellow, and the classroom teacher will facilitate. Scale is the most challenging aspect of this model because it is impossible to deploy fellows to all schools; however, Bobb noted that the Constellations Center's work continues to receive support from the National Science Foundation (NSF) and various independent organizations.

Victoria Stodden, University of Illinois, Urbana-Champaign, asked about next steps for research and resources as well as how this problem of access relates to data science specifically. Bobb responded that the

[1] The website for the Constellations Center for Equity in Computing is http://www.constellations.gatech.edu/, accessed February 13, 2020.

dominant problem for students of color is access to higher education in general; by prioritizing access to computational skills in particular, students will be exposed to data science and able to pursue any number of computational-type fields in college and beyond. Jeffrey Ullman, Stanford University, said that the number of students who took the 2018 Advanced Placement Computer Science Exam had increased substantially, including those in rural areas. He wondered whether the problem of access is being resolved across the country. Bobb replied that while increases in the numbers and types of students taking the exam are important achievements, there is still much progress to be made in terms of the numbers and types of students passing the exam. He explained that the subject matter is still not being deployed at even a minimal level in many parts of the United States. Renata Rawlings-Goss, South Big Data Hub, asked how teachers are selected for participation in the hybrid program. Bobb said that his team currently asks local principals to suggest teachers with the interest and the aptitude. Another avenue involves identifying teachers who lead courses in Georgia's Career, Technical, and Agricultural Education infrastructure's computer science and information technology pathway.[2] During a later discussion, Uri Treisman, University of Texas, Austin, posed Bobb's hybrid approach as a public policy question: Is it a public good, and, if so, who should pay for it?

## PANEL PRESENTATIONS ON EXISTING PROFESSIONAL SOCIETY EFFORTS TO INCREASE DIVERSITY

### Student-Centered Interventions to Retain Women, Underrepresented Minorities, and Persons with Disabilities in Computing

*Ayanna Howard, Georgia Institute of Technology and Computing Research Association*

Before beginning her presentation, Howard mentioned that the Computing Research Association—Women (CRA–W)[3] will soon change its name and mission statement to include all underrepresented populations, including persons with disabilities. She showed a brief video of CRA's 2018 graduate cohort for underrepresented minorities and persons with

[2] For more information about Georgia's Career, Technical, and Agricultural Education, see https://www.gadoe.org/Curriculum-Instruction-and-Assessment/CTAE/Pages/default.aspx, accessed February 13, 2020.

[3] The website for the Computing Research Association—Women (which has since been changed to the Computing Research Association—Widening Participation) is https://cra.org/cra-wp/, accessed February 13, 2020.

disabilities (URMD).[4] The 2018 URMD cohort enrolled 90 people from 60 institutions, all of whom were sponsored. Howard noted that CRA will host another URMD cohort in March 2019 and a graduate cohort specifically for women in April 2019. All of CRA's programs rely on the cohort model, which incorporate opportunities for participants to learn both from one another and from senior-level mentors. CRA offers two undergraduate programs—the Distributed Research Experience[5] and the Collaborative Research Experience[6]—both of which include student research, stipends, and mentorship. CRA also hosts Discipline-Specific Workshops,[7] Distinguished Lecture Series,[8] and Virtual Undergraduate Town Hall[9] events.

### Celebrating Women in Statistics and Data Science: Goals, Creation, Implementation, and Outcomes

*Dalene Stangl, Carnegie Mellon University and American Statistical Association Committee on Women in Statistics*

Motivated by the words of Susan Ambrose and Barbara Lazarus at Carnegie Mellon in 1992—that traditional pedagogical approaches emphasizing male patterns of behavior have restricted teaching and learning for women—Stangl and a team of women in STEM at Duke University committed to "disrupting the hierarchy." In particular, Stangl's participation in the Grace Hopper Conference on Women and Computing,[10] which today attracts more than 20,000 female participants annually, illuminated the different educational and professional experiences of men and women. With the help of a $10,000 grant from the American Statistical Association (ASA), Stangl initiated Celebrating Women in Statistics and Data Science, which gives women a "place to learn, understand, and voice what they value whether it agrees with or goes against a mainstream

[4] The website for the graduate cohort for underrepresented minorities and persons with disabilities is https://cra.org/cra-wp/grad-cohort-for-urmd/, accessed February 13, 2020.

[5] The website for the Distributed Research Experience is https://cra.org/cra-wp/dreu/, accessed February 13, 2020.

[6] The website for the Collaborative Research Experience is https://cra.org/cra-wp/creu/, accessed February 13, 2020.

[7] The website for the Discipline-Specific Workshops is https://cra.org/cra-wp/discipline-specific-mentoring-workshops-dsw/, accessed February 13, 2020.

[8] The website for the Distinguished Lecture Series is https://cra.org/cra-wp/distinguished-lecture-series-dls/, accessed February 13, 2020.

[9] The website for the Virtual Undergraduate Town Hall is https://cra.org/cra-wp/undergrad-town-hall-series/, accessed February 13, 2020.

[10] The website for the Grace Hopper Conference on Women and Computing is https://ghc.anitab.org/, accessed February 13, 2020.

work culture." The group hosted its first Women in Statistics and Data Science conference[11] in 2014; ASA has now taken over hosting this annual conference, offering technical presentations, professional development, and networking opportunities for those new to the field and those with more experience.

### Collaboration, Cohorts, and Comfort Zones: The Three Cs of Community

*Ami Radunskaya, Pomona College and Association for Women in Mathematics*

Radunskaya said that although women have made progress in terms of representation in mathematics, more work is needed. The Association for Women in Mathematics (AWM)[12] supports women and girls all along the pipeline through enrichment programs and with the assistance of 200 volunteers. AWM's programs for middle and high school girls include essay contests, mathematics days, and mentorship, and more than 200 AWM student chapters are located on college campuses across the country, she explained. For women who are more advanced in their careers, AWM offers travel grants, semiannual conferences, workshops, prizes, and distinguished lectureships. AWM also partners with NSF's ADVANCE program[13] on career advancement for women through research-focused networks. AWM's goal, according to Radunskaya, is to increase recognition of women at all levels with tiered mentoring and supportive collaboration—20 collaboration networks have already been established. Radunskaya has also been involved for 20 years with the Enhancing Diversity in Graduate Education (EDGE) program,[14] a comprehensive mentoring program to encourage women to stay in graduate mathematics programs. EDGE offers a summer immersion program, online mentoring, "difficult dialogues" sessions, support for research and travel, summer symposia, and regional clusters, and participants have become leaders in their fields across the country. Radunskaya reiterated the value of a cohort program such as EDGE in forming large networks of women.

[11] The website for the Women in Statistics and Data Science conference is https://ww2.amstat.org/meetings/wsds/2018/, accessed February 13, 2020.

[12] The website for the Association for Women in Mathematics is https://awm-math.org/, accessed February 13, 2020.

[13] The website for the ADVANCE program is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5383, accessed February 13, 2020.

[14] The website for the Enhancing Diversity in Graduate Education program is https://www.edgeforwomen.org/about-edge/, accessed February 13, 2020.

## PANEL DISCUSSION

### Initiatives for Data Science

In response to a question about whether the initiatives presented by the panelists could be duplicated for data science, Radunskaya remarked that they should be replicated for data science because data science is computing, modeling, and solving problems. She added that early collaborations with industry would be particularly useful in data science, along with mentorship opportunities. She also mentioned that the NSF Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science (INCLUDES) initiative[15] is designed to enhance U.S. leadership in STEM discoveries and innovations by focusing on broadening participation in these fields at scale. Across these various programs, Radunskaya noted that mentoring is repeatedly described as essential. Howard said that many of the existing initiatives could be duplicated in data science programs.

### Investment and Research Strategies

Panelists were asked what initiatives they would like to implement if resources were unlimited. Stangl said that the social stratification problems in elementary and high schools should be addressed first. Howard emphasized the need for time resources, in addition to financial resources, especially at the K-12 levels and for students at under-resourced postsecondary institutions. Radunskaya noted the value of dedicating time and financial resources to middle school programs, camps, 1-day events, and other partnerships to motivate children to study STEM, and she emphasized the importance of respecting the people involved in organizing such programs. She also suggested that funding be allocated to research experiences for undergraduates that intentionally engage underrepresented minorities.

Stodden wondered whether a more established research agenda would help prioritize issues of diversity and accessibility. Treisman noted that high-quality research already exists (see, e.g., Meyer et al. [2015] on the underrepresentation of women in STEM fields) and that the next step is for such research to inform classroom practice. He emphasized the need for a systemic, institution-wide approach to issues of inequity and injustice instead of simply having a few people offer useful programming.

[15] The website for the NSF Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science initiative is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505289, accessed February 13, 2020.

Radunskaya agreed that excellent research is available but that many practitioners often do not understand the research, how to implement it, or how to engage faculty in relevant professional development experiences. Brandeis Marshall, Spelman College, agreed with Treisman about the need to overcome the isolation of programs and singular points of advocacy. She explained that initiatives should be discussed with all relevant audiences (e.g., parents of elementary and high school students should be invited into conversations about the value of computational thinking for their children). She also suggested an investment beyond the academic institutional environment—for example, if the media shows women of color and people with disabilities working in computational fields, participation may increase in those areas.

### Underrepresented Populations in STEM

In response to a question about issues facing women in STEM today, Stangl stated that many of the problems that existed 25 years ago remain widespread. She said that structural changes (e.g., more flexible teaching) would better accommodate the different ways in which individuals learn, thus broadening participation in STEM communities. Treisman said that, in the past 20 years, some fields have experienced a dramatic increase in the numbers of women earning Ph.D.s (e.g., molecular biology), while others remain lacking (e.g., physics). He wondered whether cultural features of disciplines have generated the dramatic shifts and whether lessons learned could be leveraged for other fields. Stangl noted that the field of life sciences seems to have more growth for women than non-life sciences, and she added that the higher the percentage of women in departments, the lower the pay. Radunskaya observed that the breakdown of mathematics publications suggests that women are more interested in areas of study that allow for collaboration. This bodes well for gender diversity in data science, she continued, because data science is an inherently collaborative field. She also emphasized the need to abandon the myth that a field such as mathematics requires innate ability, as that is another deterrent to broad participation. Emily Fox, University of Washington, added that stratification exists even within fields (e.g., the number of women studying computational neuroscience is abysmal, while in neuroscience female representation is strong). She speculated that women often enter emerging fields later than men, which may contribute to their initial underrepresentation.

Referring to Stangl's comments about structural issues in the education system, Fox asked the panelists what universities could be doing on a regular basis to address these inequities. Howard said that a university's response should depend upon its demographic of interest. She

encouraged academic institutions to offer family leave and to include bias training in faculty hiring initiatives. She also emphasized the need to engage students at the undergraduate level in computing by offering various flavors of introductory computer science courses. Radunskaya noted that the EDGE program model could be used in both undergraduate and graduate programs, and she championed the "Uri Treisman model" for cohort building at the undergraduate level. Stangl said that flipped classrooms are effective for undergraduates, though such an innovation may be difficult at large public universities with fewer resources per student. Ron Brachman, Cornell Tech, wondered whether inclusive strategies for persons with disabilities differ from those for members of other underrepresented populations. Howard said that being an underrepresented person is a shared quality. In terms of best educational practices, she emphasized accessibility at the postsecondary level, and she encouraged participants to sign the "Computer Science for All" Accessibility Pledge[16] to make computer science materials more accessible.

## OPEN DISCUSSION

### Diversity in the Professoriate

Isbell described Diversifying Future Leadership in the Professoriate (FLIP),[17] a consortium working to change the process for graduate school admittance into computing programs and to improve representation of minorities in the professoriate. Kathleen McKeown, Columbia University, said that representation in the professoriate influences institutional changes and mentorship opportunities, and she supported Isbell's commitment to fixing problems of underrepresentation in the professoriate. Over the past 5 years, Columbia increased the number of women in leadership positions, specifically in the School of Engineering, which led to changes in hiring processes. In response to a question from Brachman about the proportion of underrepresented graduate students who become faculty, Isbell noted that biases drive decisions about whether students are motivated to pursue faculty positions. Isbell explained that the way in which people are pushed down the pipeline is flawed, and risk-averse faculty hiring practices place underrepresented minorities at a disadvantage.

[16] The website for the "Computer Science for All" Accessibility Pledge is https://www.csforall.org/projects_and_programs/accessibility-pledge/, accessed February 13, 2020.

[17] The website for Diversifying Future Leadership in the Professoriate is http://www.cmd-it.org/programs/current/flip-alliance/, accessed February 13, 2020.

### Data Science for K-12 and Postsecondary Students

Nicholas Horton, Amherst College, wondered about the extent to which data science initiatives could be implemented at the K-12 levels in order to improve participation at the postsecondary level. Rawlings-Goss said that there is abundant opportunity to be involved in the K-12 space. She said that this conversation should involve data scientists who can help create a people-driven solution informed by data. McKeown suggested asking capstone students to work on some of the data-rich city problems that Bobb discussed, and Rawlings-Goss proposed including high school students in these capstone experiences as a way to introduce them to data science and as an opportunity for mentorship. Rachel Levy, Mathematical Association of America, noted that the newness of data science in K-12 could allow teachers to reimagine themselves as mathematics doers and statistical thinkers and to help them empower their students to develop computational thinking skills. She added that engaging with students who have different kinds of learning abilities could reveal new, more accessible strategies for teaching all students more effectively. Marshall said that understanding how to get students involved in computational thinking is an ongoing conversation throughout the community. She emphasized the value of considering all individuals, not just underrepresented individuals, in order to take socially responsible actions in education. Treisman noted that high school students comprise 25 percent of the student body at 40 percent of the community colleges in Texas. No standard offering exists for such dual enrollment in mathematics, he continued, and that creates a space for the introduction of data science.

## KEEPING DATA SCIENCE BROAD

*Renata Rawlings-Goss, South Big Data Hub*

Rawlings-Goss explained that the four big data hubs are part of an NSF initiative to bring together academic, industry, and government researchers and practitioners in the data science space for the benefit of U.S. economic and social well-being. She noted that 563 data science programs exist at the undergraduate and graduate levels in academic institutions across the United States. The South Big Data Hub's "Keeping Data Science Broad" project encompassed three webinars and a workshop with 60 participants. Workshop participants included faculty from historically minority-serving institutions, community colleges, and 4-year liberal arts schools interested in creating data science programs, as well as representatives from government and industry. The project's first webinar featured speakers from campuses that already have data science programs and

focused on structural topics such as whether to require prerequisites, tips to build successful programs, and strategies to train different types of data scientists. The second webinar focused on alternative avenues to data science education such as industry programs, museum experiences, and academic programs outside of traditional STEM disciplines.

A written consensus report including challenge topics, visions for the future, top asks, and next steps for introducing data science emerged from the project's activities. The report was released in January 2018, followed by the project's final webinar. Challenges and visions discussed in the report were categorized in terms of (1) access to data; (2) assessment and evaluation; (3) curriculum; (4) data literacy; (5) diversity, inclusion, and equity; (6) ethics; (7) faculty, staffing, and collaboration; and (8) the pipeline to higher education (Rawlings-Goss et al., 2018). Rawlings-Goss commented that discussions on diversity, inclusion, and equity, in particular, revealed that a one-size-fits-all solution does not apply in all academic communities. The report revealed that implicit bias training for faculty, staff, and institutions; culturally relevant, high-quality curricula; and respect for the role that 2-year institutions, minority-serving institutions, and K-12 schools play in program development are essential.

Rawlings-Goss also described the DataUp program (hosted by the South Big Data Hub in partnership with the Carpentries),[18] which offers introductory "train the trainers" workshops. In these workshops, participants develop skills that will be useful for training their academic colleagues. The South Big Data Hub hosts the Data Science for Social Good program as well, in which graduate students work with undergraduates on local problems. Rawlings-Goss hopes that these programs will increase inclusivity and diversity in the field of data science.

Kolaczyk asked about follow-up and quality control measures for the DataUp program, and Rawlings-Goss explained that because the program is only in its first year, evaluation is still evolving. She said that upfront training reduces quality control issues and added that assessment will be conducted to understand how this program ultimately affects the trainers' institutions. Treisman turned the focus of the conversation to how to scale such efforts, and he emphasized the need to think beyond simply spreading programs. Instead, a scaling framework used in mathematics that could be helpful for data science programs includes four dimensions: spread, depth, and ownership of the program, as well as normative changes in policy and practice. He emphasized the criticality of a shift in ownership. He also highlighted the notion that "diversity," "inclusion," and "broadening participation" mean different things to different people,

[18] The website for DataUp is https://southbigdatahub.org/programs/dataup/, accessed February 13, 2020.

which can make achieving the desired social justice implications of educational programming difficult. Accordingly, Rawlings-Goss wondered about whether the DataUp program should focus on spreading to more institutions or increasing the depth at institutions currently involved.

## THE DSX PROJECT: A FIRST LOOK AT DATA SCIENCE EDUCATION ON SPELMAN AND MOREHOUSE CAMPUSES

*Brandeis Marshall, Spelman College*

Marshall described the Data Science Extension (DSX) Project, which is funded by NSF, as a 3-year targeted infusion project between Spelman and Morehouse Colleges—both of which are private, minority-serving, baccalaureate, liberal arts institutions that have mathematics and computer science departments but do not have statistics departments. DSX focuses on faculty and their impacts on students through curriculum, with the objectives of (1) sharing the power of data in context and (2) increasing access to and participation in data science practices for Spelman and Morehouse students. Marshall explained that DSX embeds one or two data science concepts into the existing curriculum. Faculty meet for 2 weeks in the summer and monthly throughout the year, giving them the time and space to consider the connection between their disciplines and data science. Faculty training revolves around interdisciplinarity, competency building, and knowledge transfer to students.

Marshall explained that the project is challenging because it requires in-house faculty development in data science (which is especially difficult at small institutions where faculty are already overburdened), technological and computing infrastructure, availability of relevant course offerings for students, and sustainability planning of courses. She described the project's benefits for students as exposure to data science in the core, cognate, and elective courses during sophomore, junior, and senior years; applicability to a variety of disciplines; and incentives to examine career opportunities in data science.

Ullman asked whether data processing could be infused into any of the courses, and Marshall replied that it depends on the course. Some courses might integrate units on data ethics or data storytelling, while others are more hands-on (e.g., an environmental science course integrated a unit on data processing). She noted that faculty do not make any assumptions about their students' previous experiences with coding or computing, and she added that many tools exist to help with those aspects of data science. In response to a question from Kolaczyk, Marshall explained that this model is still a pilot, so it will continue to be evaluated and the measures of success will remain varied (e.g., whether the

faculty training is effective, how many students are impacted, whether the infused content is valuable in the course). In response to a question from Jessica Utts, University of California, Irvine, Marshall said that the materials from the project's faculty retreats will be posted publicly in mid-2019.

McKeown asked which fields Spelman and Morehouse students pursue after completing their undergraduate degrees. Marshall said that approximately 50 percent attend graduate school. In response to a question from Kolaczyk about the management style of the project, Marshall explained that faculty participants use the Piazza platform during the summer retreats and rely on email, meetings, seminars, and small-group conversations during the academic term. Rawlings-Goss asked about differences in the infusion process at Spelman and Morehouse, and Marshall said that she has not yet observed any differences but will continue to process the data. Treisman asked whether Marshall has surveyed the faculty and students on their levels of comfort with data science tools such as Python, and Marshall noted that the vast majority of incoming students at Spelman arrive without any computational knowledge.

## HISPANICS AND NATIVE AMERICANS IN COMPUTER SCIENCE: PATTERNS, PRESSURES, AND PROGRAMS

*Lydia Tapia, University of New Mexico*

Tapia showed a series of graphs from CRA's Taulbee Survey to demonstrate that Hispanic and Native American students continue to be significantly underrepresented in computer science bachelor's, master's, and doctoral programs (Zweben and Bizot, 2018). In addition, the rate of change for the degree production is not matching the rate of change for the population of the Hispanic community. She noted that 15 years ago, fewer and fewer Hispanic students were in each stage of the computer science pipeline (i.e., bachelor's degree through full professorship). Ten years ago there were only small gains, and, 5 years ago, both small increases and decreases were evident at various stages of the pipeline. Overall, she stated that this demonstrates that not enough progress has been made for underrepresented populations in computer science.

Tapia provided an overview of her educational path to becoming a faculty member at the University of New Mexico, which began with a supercomputing challenge in high school, included an internship at Sandia National Laboratories with continued mentorship, and concluded with a doctoral degree in computer science from Texas A&M University. She noted that members of underrepresented groups often lack technology resources, endure pressures to stay close to home after high school and to contribute to the family (sometimes financially), lack an understanding of

graduate school, and experience difficulties with travel. With all of these challenges, Tapia continued, intervention programs throughout the pipeline, starting as early as kindergarten, are crucial. Successful programs at the K-12 levels include the New Mexico Supercomputing Challenge,[19] which is an expo of student computing projects with mentorship opportunities; the Tapia Lab[20] Demos; and the New Mexico CS4All program,[21] which trains students in dual enrollment classes and trains teachers who will be working with those students. At the undergraduate level, the NASA Swarmathon[22] and the Robot Guru both engage underrepresented students in computer science. The CRA–W URMD grad cohort, discussed by Howard, is beneficial for graduate students, Tapia added. At the faculty level, career mentoring workshops and proposal writing workshops are two methods to improve retention in the field. Moving forward, it is important to consider additional ways to increase participation at all stages of the pipeline.

Marshall asked Tapia why a drain exists at every level of the pipeline for a variety of demographics. Tapia replied that while encouragement to pursue a bachelor's degree may be strong in Hispanic communities, for example, motivation to attend graduate school is lower because such a degree is not required to secure employment. This explains the first significant drop in the numbers on the pipeline, although the larger drop occurs at the Ph.D. level. Tapia believes that mentoring is the best way to overcome that gap. Kolaczyk asked whether mentoring could be expanded to better address local cultural considerations, and Tapia responded that although families may be reluctant to listen to advice from a stranger, it could be valuable for students' mentors to interact with their families. Treisman later cautioned against stereotyping students and viewing them through a deficit-focused lens, as many students came from families who support their educational aspirations.

Radunskaya noted that minorities often carry a larger faculty service burden, and she wondered how successful Tapia has been at convincing her colleagues to support her in these endeavors. While she has supportive colleagues, Tapia acknowledged that this is a challenge and that sometimes service is downplayed relative to research in tenure reviews. McKeown commended Tapia for highlighting this pervasive struggle in

[19] The website for the New Mexico Supercomputing Challenge is https://supercomputingchallenge.org/18-19/, accessed February 13, 2020.

[20] The website for the Tapia Lab is https://www.cs.unm.edu/tapialab/, accessed February 13, 2020.

[21] The website for the New Mexico CS4All program is https://cs4all.cs.unm.edu/, accessed February 13, 2020.

[22] The website for the NASA Swarmathon is http://nasaswarmathon.com/, accessed February 13, 2020.

the tenure process, and she wondered whether it is time to rethink the notion of "credit" in academia. Levy agreed that work that is not "publishable" often does not receive the credit it is due and suggested that the roundtable continue a conversation about new approaches to recognizing important work. Treisman offered his support for such a conversation.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Participants divided into two groups to discuss issues of diversity and access in greater detail. The first group focused its discussion on leveraging programs that have been successful for underrepresented populations in other fields. On behalf of his group, Kolaczyk reported that because "data science" is an amorphous term, it can prove challenging to recruit undergraduates into the field. Thus, he continued, it is imperative that students know what data science opportunities exist and which skill sets are needed for particular career paths. He added that it could be possible to leverage the efforts of a field such as neuroscience, which has been successful in recruiting and retaining women, by explaining that data science overlaps with that particular field. He also described the University of California, Berkeley, introductory data science course, Data 8, which requires no prerequisites, enrolls approximately 2,000 students each semester, and is complemented by "connector courses" in various disciplines. Data 8 participants are exposed to valuable technical skills even if they choose not to follow a technical career path. Regarding professional organizations, Kolaczyk said that they could promote data science activities of interest to students (e.g., ASA's DataFest) and could fund regional events during which students could learn more about various data science careers and the training they require. He suggested that data science organizations collaborate with organizations that are actively engaging women and minorities in other fields. Kolaczyk's group also discussed the importance of scaling up mentorship opportunities. As an example, ASA works directly with high school counselors to encourage early participation in data science activities. He added that the student chapters of ASA or AWM could host networking events or organize panels of faculty, undergraduates, and graduate students to provide information to high school students.

The second group focused its discussion on promising opportunities for investment if resources were unlimited. On behalf of his group, Brachman said that it is critical to consider the current research and to evaluate the largest potential marginal payoff before making funding decisions. Resources could be dedicated to making improvements along the pipeline by first engaging the media to better portray diverse

individuals in technical fields. A partnership with an organization such as the Geena Davis Institute on Gender in Media may present interesting opportunities, he continued. Once people are attracted to the field of data science, Brachman explained, they need to be given technical resources and creative opportunities to learn before they enter high school. He added that students have the potential to become most engaged in data science if they are presented with choices for both school and extracurricular programming. He noted that PK-12 teachers could be given stipends, perhaps from industry, to participate in data science initiatives and to develop innovative educational programs. At the postsecondary level, scholarships would assist students who do not have the means to complete their degrees, although complications may arise if these funds come from companies that have expectations for the students after graduation. Brachman's group also discussed the need to balance mentorship and sponsorship as well as the importance of resource allocation toward sponsorship and collaboration opportunities. He added that collaboration should be built into curricula to prepare students for the work that awaits them in the field of data science. Overall, he continued, it is important to think about ways to revise curriculum and pedagogy to encourage cross-disciplinary work and incorporate studies of data ethics. This includes increased attention to faculty training in new tools and resources, perhaps during summer institutes. Industry could also be involved in departmental reviews so that faculty can improve the way they train students for future industry jobs.

# 10

# Meeting #9: Motivating Data Science Education Through Social Good

The ninth Roundtable on Data Science Postsecondary Education was held on December 10, 2018, at the Keck Center of the National Academies of Sciences, Engineering, and Medicine in Washington, D.C. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to learn about academic, government, nonprofit, and private-sector projects promoting data science for socially desirable outcomes and their intersection with education and hiring, and to explore how socially motivated projects and topics can engage and excite students. 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.

Welcoming roundtable participants, Kathleen McKeown, Columbia University, commented that many students in data science, computer science, and statistics courses are eager to "give back" to their communities through the practice of data science for social good. She highlighted ethical concerns raised during previous meetings of the roundtable, such as potential bias in machine learning and fair artificial intelligence (AI), which are important to revisit in discussions of how data could be used for social impact.

## AN INFORMAL DISCUSSION ABOUT DATA SCIENCE FOR SOCIAL GOOD

*D.J. Patil, Devoted Health and Former Chief Data Scientist, White House Office of Science and Technology Policy*

Patil explained that the Chief Data Scientist for the United States, a role that is currently vacant, works to ensure that data are used responsibly to benefit all people uniformly instead of to divide or oppress individuals and communities. As data officer positions have been created within the federal government, at state and city levels across the country, and throughout the world, Patil is hopeful that the current administration will find a way to leverage this role.

While many researchers are focused on AI and algorithmic bias, Patil noted that data collection, use, safety, and security, as well as appropriate policy making around data, are basic concepts worthy of increased attention. *Ethics and Data Science* (Patil et al., 2018) identifies ethical constructs lacking in organizations, such as a dissent channel, a checklist for product launches, and standard principles for ethical data use. He suggested that ethics and security be integrated throughout data science curricula and that future data scientists receive increased liberal arts training. He championed the role of 2-year institutions in offering introductions to data science for social good, and he advocated for Congress to support free education from 2-year institutions for all Americans.

Patil identified ways in which data science could be used to benefit society. For example, various technologies could have been used during Hurricane Katrina to predict how many people would evacuate and from which areas in March 2019, to detect where bridges were washed away, to locate people sheltering on rooftops, and to direct boats engaged in search and rescue missions. Data science could also be used to help police departments compare data across state lines, because no infrastructure currently exists to do so. However, Patil emphasized that transparency remains an issue, especially for applications in the criminal justice system. The mental health space is already benefiting from data science applications with the development of a crisis text-line to help meet the demand of mental health emergencies. Patil added that data science and AI could have a substantial impact on basic logistics and transportation problems. He emphasized that one does not need access to a large data set to impact society and suggested contacting local food banks or shelters to find out whether their challenges could be addressed with data science.

Uri Treisman, University of Texas, Austin, observed that when government agencies fail to manage crises, citizens often organize responses. However, getting data quickly and optimizing resources remains a

challenge; local volunteers need to be trained to use data science in emergency situations. Patil agreed that these "digital humanitarians" need guidance on how best to create infrastructure and coordinate so as to be most effective. Jessica Utts, University of California, Irvine, asked Patil for advice on structuring a data ethics course. Patil encouraged faculty to integrate ethics throughout the curriculum—referencing Professor Ed Felten's, Princeton University, case study approach as a model—instead of offering only one course on ethics and security. He directed participants to view and contribute to a collection of curricula[1] from faculty across the country. Mehran Sahami, Stanford University, asked Patil to talk more about the importance of liberal arts education and the most useful tools for data science. Patil described liberal arts' emphasis on formalism, creativity, and framework development as invaluable in preparing to solve industry and societal problems.

## FROM CLASSROOM TO CLINIC: DATA SCIENCE FOR SOCIAL GOOD FELLOWSHIPS AND THE LESSONS DATA SCIENCE EDUCATORS CAN LEARN FROM THE MEDICAL PROFESSION

*Matt Gee, University of Chicago and BrightHive*

Gee described the University of Chicago's Data Science for Social Good (DSSG) program[2] as an immersive fellowship in which aspiring data scientists learn how to map data methods and tools to social problems in partnership with a government agency or nonprofit organization. Gee said that DSSG builds a community of open, ethical, collaborative data science practice through research and development, lectures, workshops, and events. In its first year, DSSG received more than 600 applications but chose only 36 fellows to participate in the program. DSSG looked for partner organizations with important problems, leadership buy-in, access to data, staff capacity to work with data, and a commitment to implementing solutions. After defining goals, determining what actions would be taken, identifying what data were available internally and what data would be needed, deciding what analysis needed to be done and how it would be validated, 14 projects emerged and the first cohort of fellows arrived in May 2013. In working with both their partner

[1] The website for this collection of curricula is https://docs.google.com/spreadsheets/d/1jWIrA8jHz5fYAW4h9CkUD8gKS5V98PDJDymRf8d9vKI/edit#gid=0, accessed February 13, 2020.

[2] The website for Data Science for Social Good is http://www.dssgfellowship.org/, accessed February 13, 2020.

organizations and their DSSG mentors, fellows learned to consider the social and ethical implications of data used in decision making as well as strategies to communicate with diverse audiences. Project outcomes have included solutions to predict heart attacks, to anticipate school dropout rates and improve graduation rates, to help state governments save money on energy bills, and to help aid organizations respond to crises faster. Partner organizations emerged with an understanding of how to view data as an asset, Gee said, while fellows learned that data science tools, when used responsibly, may amplify one's ability to do good. During its 6 years, DSSG has engaged more than 224 fellows from all over the world in 70 projects.

Although the program has made great progress, Gee explained that the title "Data Science for Social Good" implies a moral superiority for data science that helps nonprofits and government agencies, and reduces data science for social good to something done in one's spare time. He emphasized that all data science should be grounded in a sense of the good; instead of doing data science for good, professionals should continually do good data science. As educators consider the future of data science training, Gee suggested turning to long-established professions, such as medicine, and learning from their experiences. He referenced Paul Starr's *The Social Transformation of American Medicine* in his rationale for new data science pedagogy. First, he explained that because data science has gained popularity, economic power, and cultural cachet quickly, data scientists are often unaware of the potential consequences of their work. Data science education is currently failing in that it is taught at a distance, with clean data sets separated from social context. Instead, Gee continued, students need to be taught about developing personal accountability and avoiding algorithmic tyranny, in which algorithms lead rather than inform decision making. For example, DSSG fellows spend the first 2 weeks of the program working without data, talking with project partners, and gathering context. Second, he explained that data science would benefit from the development of professional norms—for instance, choosing service over profit when the two conflict, so that consumers know that their best interests are considered when working with their data. Gee referenced The Global Data Ethics Project as an example of the profession's attempt to adopt ethical principles. Third, he commented that it is important for the data science profession to attract and retain the best and brightest minds.

Moving forward, postsecondary educators could add clinical practice requirements to data science programs. Although this could be both complicated and expensive, Gee commented that this would allow students to explore the social context of where data are generated and will be used, developing the analogue to medicine's "bedside manner" for

data science. Educators could add written and verbal discussions of the social and ethical implications of data sets and models into problem sets in data science coursework instead of relegating ethical conversations to a single course. Educators could also provide guidance to employers for incorporating ethics case studies into hiring, apprenticeship, and mentorship opportunities. Taking these steps to improve data science training, Gee said, could render data science as more of a "healing profession with deep purpose and moral authority." Michael Pearson, Mathematical Association of America, asked Gee whether DSSG includes discussions of how data science will inform policy or hold policy makers accountable for data misuse. Gee acknowledged that the program would benefit from more discussions about "data misuse" as well as "data missed use."

## TEACHING DATA THAT MATTERS

*Rahul Bhargava, Massachusetts Institute of Technology Media Lab*

Bhargava began with a moment of silence to acknowledge that many people still face discrimination working in the space of data science and to honor the history of Title IX, which has provided instruments to help address this problem. In discussing the concept of data storytelling, Bhargava noted that how information is presented to an audience impacts how it will be understood—viewers are often distanced from the lived reality of data. He described the separation that exists between the desire to do something valuable with data and the respect for the experience of the person represented by the data. This notion of respect is accompanied by a question of responsibility: Is an algorithm designer responsible for what happens to an algorithm user?

Bhargava explained that powerful people have used data to subjugate those without power throughout history. For instance, Egyptian leaders created a census to catalogue laborers for the construction of pyramids. This history has to be acknowledged and challenged by those who wish to use data for good, he continued. Both historic and contemporary counter efforts exist: Predictive models were developed in the 17th century to prevent the Bubonic Plague, and W.E.B. DuBois used infographics to catalogue and share the life experiences of former slaves. Currently, the Data for Black Lives organization[3] works to eliminate the presence of bias in data. In all of these cases, data were used to tell alternative stories about matters of social importance. Teaching "data that matters" presents an opportunity for students to better use and understand real data, to ask hard questions and take risks, and to balance learning objectives with

[3] The website for Data for Black Lives is http://d4bl.org/, accessed February 13, 2020.

personal interests. Bhargava teaches a cross-disciplinary course, hosted by the MIT Humanities department, called Data Storytelling Studio,[4] in which students "consider the emotional, aesthetic, and practical effects of different [data] presentation methods." This course is offered via MIT's Open Courseware.[5]

He described three student projects from this course in which data sets were put into context to inform actions: (1) a board game, comprised of refugee data, that people "play" at a fundraiser to better understand the refugee experience and hopefully donate to the cause; (2) an inverted map of real stop-and-frisk data, accompanied by a satirical data journalism story; and (3) a data-driven game, based on Food and Drug Administration data, to teach children about the roles that bees play in the environment. Bhargava said that his classroom is a "playground" where students "flex their data muscles" in a safe learning space. So that other educators can access hands-on data-storytelling activities, this open source content is available through the Data Culture Project.[6] Alfred Hero, University of Michigan, wondered how Bhargava achieves a convergence between his course and more traditional data science methodology courses because many students enrolled in the latter may not enroll in the former. Bhargava said that he recruits students for his 30-person course; those students then advertise the course in their departments.

## OPEN DISCUSSION

### Program Development

Nicholas Horton, Amherst College, noted that DSSG serves as a model of integrated co-curricular experiences, but he wondered about the barriers to rolling out similar programs at less-well-resourced institutions. Gee said that while challenges vary by institution, few institutions offer a clear home for such a program or the faculty and budget lines to support it. He emphasized the value of creating a dedicated co-curricular space. Bhargava noted that MIT's Media Lab, known for its anti-disciplinarity, has positioned itself at the intersection of numerous fields, is well supported, and attracts great students and faculty. He noted that no single recipe for success exists for all institutions. Bill Howe, University of Washington, wondered whether emphasizing the liberal arts and injecting

[4] The website for the Data Storytelling Studio is http://datastudio2018.datatherapy.org/, accessed February 13, 2020.

[5] The website for Open Courseware is https://ocw.mit.edu/index.htm, accessed February 13, 2020.

[6] The website for the Data Culture Project is https://databasic.io/en/culture/, accessed February 13, 2020.

more social context into data science programs could cause some students to lose interest in the courses, either because they differ from what the students imagined or because they require messy project work. Bhargava said that truths about fields are always evolving; faculty should help students reset their assumptions and build a new knowledge base. He addresses similar student concerns through team design, pairing students with different perspectives, learning goals, and work habits. Gee said that some fellows consider leaving the program each year because they dislike the amount of time spent talking with project partners or navigating team politics; however, most ultimately realize that this "messiness" is the benefit of doing clinical practice.

### Community Partnerships

Deb Agarwal, Lawrence Berkeley National Laboratory, commented that although short-term problem-solving engagements have some value, she asked whether DSSG fellows have the opportunity to study and learn from the efforts of previous cohorts. Gee responded that the fellows discuss why projects were not chosen, which helps them understand "messy" issues they may face; however, he noted that he might implement Agarwal's idea with a future cohort. Rachel Levy, Mathematical Association of America, wondered whether DSSG project partners have the capacity to test and use the solution provided by the fellows and have the independence to modify it. She emphasized the value of thinking about tools as opportunities not only for the fellows but also for the project partners. Gee described three possible considerations to build better capacity within the partner organizations: (1) right-size the project to the course and the timeline; (2) provide cross-semester or cross-year continuity for a project; and (3) ensure that training for the partner is built into the curriculum. Bhargava mentioned that he no longer develops community partnerships in his course because one semester is insufficient to cultivate such relationships.

### Ethical Considerations

Sahami said that many computer science faculty are uncomfortable teaching ethics because they lack the relevant training. While a philosophy department could offer a multidisciplinary course, he wondered what other strategies could be used to teach ethics in a meaningful, balanced way. Gee and Bhargava suggested that simply exposing students to the appropriate set of questions, without necessarily providing the foundational text, helps prepare them to continue to learn on their own. Treisman said that because of the power and potential of data science, a

rich liberal arts background should be embedded in data science education. Students have to understand how to enter into the social worlds in which they will use data if the objective is to empower people, he continued. Treisman emphasized that all academic departments, not just philosophy, have an obligation to attend to the social, ethical, and moral development of students.

Jeffrey Ullman, Stanford University, questioned whether educators and researchers have the right slant on the matter of data consent, as the notion of data privacy is a modern construct. He said that because Google and Facebook are free platforms, consent is a difficult concept; if the companies were to charge users to opt out of data collection, lower-income users would lose access to privacy protection. He emphasized the need to think carefully about allowing data consent, pointing to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) as an example of how a codification of privacy rights can have unintended consequences (e.g., in HIPAA's case, complicating patients' ability to communicate with medical professionals). An audience participant noted that many more stakeholders exist in data science than medicine, and negative consequences are possible for many stakeholders.

## DATA, DESIGN, AND ENGAGEMENT: LESSONS FROM 30+ DATA SCIENCE FOR SOCIAL GOOD PROJECTS

*Peter Bull, DrivenData*

DrivenData[7] has worked on more than 50 projects with nonprofits, social enterprises, and corporate social responsibility groups, Bull explained, and it tries to figure out how to solve organizations' problems with machine learning or data science tools, using the data assets that they already have. An organization's problem is posted online, and a community of data scientists proposes algorithms to solve it. DrivenData selects the best-performing algorithm and assists the organization with implementation. DrivenData has run more than 30 competitions during the past 5 years, with participation from a community of more than 35,000 data scientists from across the world.

Bull described three example projects. The first project helped a school district approach budget benchmarking in the absence of structured data about school spending. DrivenData helped build an algorithm for the school district to generate predictions for spending as well as information about what part of the budget was being used. This automated process

[7] The website for DrivenData is https://www.drivendata.org/, accessed February 13, 2020.

replaced the approximately 300 staff hours per year that were spent analyzing spreadsheets with similar information. The second project helped a community improve its strategy for capturing water from coastal fog with mesh nets. DrivenData used data from weather stations located next to these mesh nets to try to predict their yield. This work prompted the community to prioritize the placement of new fog nets. The third project helped to prioritize health inspections for Boston restaurants using data from 4 years of health code violations combined with Yelp reviews and ratings. With this new method in place, inspectors were able to find 25 percent more violations and thus better protect citizens.

Bull explained that achieving the highest accuracy is not always the desired outcome when building a model. Instead, the desired outcome is how the accuracy works in concert with other goals and metrics for success. With this in mind, DrivenData hosted a new type of competition, Concept to Clinic, in which contributors earn points and achieve visibility by submitting their work to an open source repository. This adds an element of collaboration to the competition and promotes sharing throughout the process instead of upon completion, Bull continued. He described DrivenData's other open source projects, including Cookiecutter Data Science,[8] a standardized project structure for doing data science work, and Deon,[9] an ethics-checklist generator for projects. DrivenData also engages directly with organizations to solve data science problems. In closing, Bull shared a Data Impact Field Guide, with concrete challenges to consider before engaging in a project:

- *Finding a project*. Bull said that this is the most difficult part of the process and where the greatest need exists. Ninety-five percent of the time, organizations want help measuring impact. However, data scientists may not be the best equipped to do this in a short amount of time. If one thinks about impact measurement as early as during the data collection stage, the majority of the work will be done by a domain expert, whereas if one thinks about impact measurement during data analysis, the majority of the work will be done by a data scientist.
- *Launching a project*. Bull noted that because social-sector organizations exist for the public good, they demand higher attention to data ethics. For example, questions about security, explainability, and responsibility arise during the data collection, modeling, and

[8] The website for Cookiecutter Data Science is https://drivendata.github.io/cookiecutter-data-science/, accessed February 13, 2020.

[9] The website for Deon is http://deon.drivendata.org/, accessed February 13, 2020.

deployment phases, respectively. He asserted that better ethics develop through increased practice.

- *Running a project*. A project should build trust and empathy between the user and the technologies by embedding ideas from human-centered design thinking into the data science process. A human-centered data scientist will go to the field and observe data being generated; design plans with the user by iterating jointly on prototypes; assess outcomes both quantitatively and qualitatively; and be honest about and learn from failures.
- *Wrapping up a project*. The capacity gap between the social sector and either industry or academia is wide and can jeopardize solution hand-offs. There is also a shortage of more than 140,000 data scientists in industry, a problem felt heavily in the social sector.

## TEACHING PEOPLE TO THINK WITH DATA

*James Hodson, AI for Good Foundation*

Hodson explained that the AI for Good Foundation[10] was established in 2014 after a series of workshops at Stanford University about the status of AI and future innovation. Participants discussed core problems, breakthrough methodologies, and social impacts. After the workshops, he continued, it became clear that a bridge between research laboratories and government, industry, and nonprofit stakeholders was needed. Questions emerged about how AI aligns with the notion of social good as well as how communities could be built to enable long-term change. In response, the AI for Good Foundation adopted the United Nations' 17 Sustainable Development Goals[11] as its framework. Although these goals are unlikely to be attained in the near term, Hodson noted, they raise questions about how to solve this generation's challenges. The AI for Good Foundation continues to build the capacity to reach these goals through partnerships with academic laboratories. He said that cross-departmental initiatives at academic institutions, in combination with engagement from actors on the ground, are promising.

[10] The website for the AI for Good Foundation is https://ai4good.org/, accessed February 13, 2020.

[11] The Sustainable Development Goals are as follows: no poverty; zero hunger; good health and well-being; quality education; gender equality; clear water and sanitation; affordable and clean energy; decent work and economic growth; industry, innovation, and infrastructure; reduced inequalities; sustainable cities and communities; responsible production and consumption; climate action; life below water; life on land; peace, justice, and strong institutions; and partnerships for the goals. For more information about these goals, see https://www.un.org/sustainabledevelopment/sustainable-development-goals/, accessed February 13, 2020.

He presented a potential definition of data science: a set of algorithmic methods and engineering practices that need a channel for development and adoption within empirical research. He added that industry and society need data literacy to harness the value of data and to aid in solutions to a wide variety of problems. Hodson said that it is important for students to understand the realities of the challenges people are facing in the real world. He emphasized that data science does not need to be housed in a stand-alone department because it should not be viewed as a different field. He explained that each academic discipline has its own long-established tradition of working with data, and, although it would require additional faculty training, each discipline could teach important aspects of data science within its department. Academic institutions have a responsibility to train people to go into industry and government to solve hard problems with data, rather than training everyone to be a data scientist, he continued.

Hodson said that society should embrace data-driven science; data literacy across campus; cross-disciplinary research and teaching resources; open infrastructure, data, and methods; data innovation hubs; data science for social good; and diversity. The main barriers to achieving these goals are that the methods are often taught independently from the research process; students are seldom taught how to evaluate, clean, and merge data; and the teaching of applied data science in a laboratory setting is too short, too stylized, and has no impact. Hodson noted that discussions about ethics should not be motivated only by regulatory purposes. To truly bring social impact into the data science classroom, one semester of instruction is insufficient, he continued. Sahami asked to what extent students should be engaging in projects with real social impact and measuring results versus understanding the issues and methodology. He noted that, in academia, faculty are often constrained by time, expertise, and resources. Hodson agreed that merging best educational practices with social impact is challenging. While he acknowledged that there is an opportunity to use projects as gateways for continuing interaction, he said that they are not necessary to teach the fundamental principles of data science.

## CAN AI REDUCE GANG VIOLENCE OR CAUSE MORE HARM?

*Desmond Patton, Columbia University*

Patton's current work uses qualitative methods, machine learning, and community expertise to better understand how social media provides a window into gang violence. SafeLab's[12] interdisciplinary team of social

[12] The website for SafeLab is https://safelab.socialwork.columbia.edu/, accessed February 13, 2020.

scientists, computer scientists, and domain experts develops technology tools to support the prevention of gang violence. Patton was motivated to study this area by the rise of crime in Chicago—764 homicides occurred in 2016, most of which involved guns, public spaces, and prior altercations, many of which were described in social media posts.

SafeLab studied how a now-deceased gang member, Gakirah, narrated her life on social media and how other people responded to her posts. Many of the posts were difficult to understand in terms of language, context, and nuance, so a methodological approach was needed to understand the data. During the first stage of the contextual analysis, the research question and study population were clarified, the social media corpus was created, and domain experts (i.e., gang members and other youth in the community) were identified. After annotators received training, they began to code the data and to develop a baseline interpretation. Annotators then created descriptions informed by the context of the social media post, and machine learning was used to label data as "loss," "aggression," or "other." Domain experts would then review the labels and help reconcile the interpretations by providing additional context. The labeled data sets were fed into natural language processing algorithms, which developed additional tags and labels to translate the social media data into standard English.

Patton's team is developing greater accuracy and leveraging more context; in 2018, it developed a new labeled data set, six times larger than the previous, and integrated neural net approaches. The team has also established new partnerships with computer vision specialists so that information can also be collected from images, which tend to better identify aggression and substance abuse, according to Patton. He explained that ethics is especially important in this line of work: the team is careful in how it uses information about aggression in young men and women of color, has its annotators sign nondisclosure agreements, and refrains from sharing publicly any images from the data set. This work provokes a conversation about the importance of data in context—Patton's team is developing a conceptual framework to theorize how social media policing can negatively impact communities of color and is creating digital interventions for youth. Ultimately, Patton's goal is to build empathy and to drive behavioral change, because young people may not understand the consequences of their digital footprints.

Eric Kolaczyk, Boston University, asked about the challenges and lessons learned during annotator training. Patton responded that a main challenge was trying to figure out how to best support the diverse annotators. He noted that he did not anticipate the way that "life would get in the way" for the young people serving as domain experts and added that

challenges exist in maintaining relationships with them. The social work students had to learn how to treat the user as a whole person and how to interpret more accurately. Patton also cited a need to be aware of the triggers that can happen for annotators confronted with disturbing posts—for example, about violence toward women. In response to McKeown, a member of Patton's research team, Patton commented that the social work and computer science students worked well together and pushed each other toward the best solutions. The social work students taught the computer science students strategies to confront real-world problems and challenges, while the computer science students taught the social work students to develop data literacy and to ask the right research questions. In response to a question from Ullman about law enforcement's use of electronic footprint monitoring, Patton suggested that people should challenge and critique the methodology as well as understand the context. He emphasized the importance of using these techniques equitably and applying them across demographic groups uniformly. Louis Gross, University of Tennessee, described a workshop he will host in May 2019 on the mathematics of gun violence and potential impacts for alternative interventions, and Patton encouraged him to include social scientists in the conversation to think about data patterns.

## OPEN DISCUSSION

### Data Science Education

Ullman asked the speakers to outline the technical differences between "data science for social good" and "data science." Bull responded that it is important to give all data scientists a concrete process to ask the right questions in order to understand the domain they are working in, especially when it comes time to hand off a solution to an organization. Hodson said that there is great opportunity to make social impact through data science, but that is not the most important part of rethinking data science education. He reiterated that data science is a set of processes and methodologies in which all departments should partake as opposed to a separate discipline. He encouraged cross-departmental collaboration instead of teaching data science as an isolated subject. Bull said that data science may face similar challenges to the field of software engineering in finding a disciplinary home that has both a particular set of skills and domain-specific research questions. Hero asked how data science will scale to meet high student demand if it is not housed in a separate department. Hodson said that courses that are not necessarily department-specific will have to be created.

### Collaboration and Management

Kolaczyk asked how people in academia could best interact with Bull's and Hodson's organizations. Hodson said that the AI for Good Foundation has done many programs jointly with universities and public research institutes (e.g., workshop series, co-teaching). He noted that the foundation tries to unite researchers, students, and community stakeholder groups; it helps external organizations understand where they need advice and interaction and helps researchers understand how theoretical research can be applied. Bull suggested that academic institutions avoid partnering with DrivenData because doing so could create a bottleneck; however, the lessons learned from working with organizations could be used as resources for educators who wish to set up their own projects. Educators could set up long-term partnerships with other organizations across multiple years and multiple cohorts of students, Bull said. Treisman noted that the relational trust needed when working in political environments is more complex than when trying to help a business optimize sales, for example. He highlighted cycles of interaction between data users and data owners, in which trust has to be built around everyone knowing and following the same rules. He wondered whether anyone has written descriptions of these processes as well as how we might make it easier for people to learn how to do this work. Bull agreed that working with social-sector organizations is often more difficult because their metrics of success are undefined and that building trust is critical. He suggested that the data science community think carefully about the best way to engage with these organizations. Hodson noted that organizational behavior research may provide insight into these areas. He said that the structure of the institutions that people are working within have to change to allow for these new types of interactions. Treisman added that the role of design expertise is often underestimated when organizations attempt to improve. He explained that data management/optimization techniques, institutional mechanisms for knowledge management, and clever design are essential, most of which does not come from technical, mathematical tools. Bull appreciated Treisman's description and agreed that DrivenData faces a challenge of balancing creativity and knowledge management with technical know-how. Hodson agreed but added that educators have a responsibility to teach people how to develop architectures that will lead to better outcomes.

## SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Following the presentations and open discussions, roundtable participants divided into three groups to discuss specific themes from the day.

The first group discussed integrating data context into students' coursework. On behalf of her group, Levy explained that each discipline has a different way of facilitating communication between domain experts and technical experts. It is important for students to develop an appreciation for what each person contributes to such a conversation. While it may be possible to teach students how to have those conversations, Levy continued, it is a skill that needs to be practiced and developed over time. She said that students should explore and experience misunderstandings of language, culture, biases, assumptions, and constraints in order to be better practitioners in context. Levy's group also questioned the use of the phrase "social good," as its meaning may vary by context. Her group said that conversations about what "social good" means, who defines it, and who benefits from it should be included in data science curricula.

The second group discussed the benefits and drawbacks of increased training around data science for social good. On behalf of his group, Ullman acknowledged that some students and faculty are interested only in the theory of a subject rather than its practical application. He used mathematics as an example of a discipline that has been driven by theory, successfully, for 3,000 years. In data science, he continued, people who are interested in developing new machine learning models without paying attention to what they will be used for could create problems. He suggested that it may be ineffective to orient data science education programs toward people who are uninterested in how their ideas will be applied. When people are forced to work in diverse teams (e.g., data scientists and domain experts), people step outside of their comfort zones and explore broader issues. Ullman's group advocated for a curriculum with a solid mix of theory and practice and noted that a flipped classroom is one way to facilitate such a curriculum.

The third group discussed how to incorporate ethics in a responsible and informed manner across the curriculum. On behalf of his group, Sahami explained that definitions of "social good" and "ethics" remain unclear. He suggested integrating ethics into data science instead of discussing it as a separate entity so as to better develop ethical behavior. Although there are many layers in the technology stack—for example, who is responsible for how technology is used—issues of ethics, social justice, and societal good are often combined and thus not considered adequately. Sahami noted that data science education and practice could benefit from the best ethical practices of other more established communities and that alternative models could be embedded across multiple disciplines. Sahami's group also discussed the potential for those in leadership to speak more openly about issues of ethics so as to make the concept more accessible to young people. Sahami pointed out that data science does not yet view itself as a profession like medicine, which has a clear

code of ethics. Utts added that faculty are trained with integrity in their disciplines and should pass those principles along to their students in every course, which Howe connected to Gee's earlier discussion of "professional sovereignty." Kolaczyk wondered whether society has reached a point where the potential to do good or harm is at a completely different scale than ever before, forcing practitioners and educators to wrestle with larger issues. Treisman noted that the data science community can influence the infrastructures that currently stipulate ethical behavior.

# 11

# Meeting #10: Improving Coordination Between Academia and Industry

The 10th Roundtable on Data Science Postsecondary Education was held on March 29, 2019, at the Arnold and Mabel Beckman Center of the National Academies of Sciences, Engineering, and Medicine in Irvine, California. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss common challenges in establishing, maintaining, and evolving partnerships in data science between academia and industry, and to learn about ongoing programs at academic institutions and research groups around the United States. 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.

Eric Kolaczyk, Boston University, welcomed roundtable participants and noted that although partnerships between industry and academia have existed for years, such collaborations are now occurring at a different scale and with a new intensity, owing in part to the emergence of data science. Academia–industry partnerships enable students to integrate data science skills to address real-world problems. Students also gain insight into the industry workforce and potential career opportunities. And members of industry can experiment with minimal investment, tapping into new developments from academia and identifying prospective hires. Challenges to developing successful partnerships include initiating interactions, maintaining support from institutions, aligning expectations, and navigating issues of data sharing and intellectual property (IP).

Roundtable speakers and participants discussed best practices to create effective academia–industry collaborations around data science research and education.

## OVERVIEW OF ACADEMIA–INDUSTRY PARTNERSHIPS

*Lise Getoor, University of California, Santa Cruz*

Getoor commented that data science presents a unique opportunity for new models of engagement to address challenges in academia and industry. Existing models of academia–industry collaboration include sponsored research, summer internships, capstone projects, visiting researcher status, and formal industrial membership programs.

She explained that there is no one-size-fits-all model for academia–industry partnerships; it is important to develop a shared vision around building a thriving data science education and research community that spans academia and industry, with students at the center. The industry "ecosystem" includes "heavy-hitters" in data science (e.g., Google, Amazon, Microsoft, IBM, Facebook), start-ups, and new adopters, each with different needs and opportunities. Styles of collaboration (e.g., to educate, share expertise, or collaborate on research), expectations, and timelines differ both between industry and academia and across companies. The needs of and opportunities within the academic ecosystem vary based on the institution's ranking, location, and major disciplines. Cultural differences among data science domains can also be a consideration—for example, the tradition of project-based work that can align well with industry expectations is more common in statistics than in computer science and mathematics, in her experience. Getoor provided a brief overview of the Data Science $D^3$ (Data, Discovery, and Decisions) Research Center at the University of California, Santa Cruz.[1] It focuses on academia–industry collaborations around richly structured sociobehavioral data and uses probabilistic programming language to develop templates for sociotechnical systems. The research center follows the National Science Foundation's (NSF's) Industry–University Cooperative Research model, which provides a template for addressing IP issues. Industry benefits from the fresh perspectives and research that emerge from partnerships like these, and students benefit from opportunities to work in teams and conduct research with data for real-world problems.

[1] The website for the Data Science D3 Research Center is https://d3.ucsc.edu/, accessed February 13, 2020.

## PANEL ON MECHANISMS FOR ENGAGING AND FOSTERING INDUSTRY PARTNERSHIPS

*Adam Causgrove and Rebecca Nugent, Carnegie Mellon University*

Causgrove is a corporate relations officer at Carnegie Mellon University (CMU), where he advocates specifically on behalf of the departments in CMU's Dietrich College of Humanities and Social Sciences. He and Nugent discussed the value of corporate relations officers, particularly for taking a holistic approach to supporting and sustaining academia–industry partnerships. Corporate relations officers highlight the diverse opportunities available to potential industry collaborators as well as the diverse students at CMU in the hopes that companies will choose to engage in long-term partnerships with any and all of CMU's colleges.

Causgrove described seven channels through which industry can engage with CMU: student engagement, sponsored research, faculty engagement, professional education, licensing and technology transfer, start-ups, and co-location. Student-centric interactions are particularly popular with industry partners, and engagement is tailored to remain mutually beneficial for CMU and for the companies over time. He mentioned that more than 200 institutions are members of the Network of Academic Corporate Relations Officers,[2] which performs benchmarking, develops best practices for building relationships with industry, and offers resources for institutions that wish to establish corporate partnerships.

Nugent explained that CMU is formalizing an institution-wide Corporate Affiliated Projects (CAP) program. In the CAP program, local, national, and global industry partners work with faculty to scope real-world problems for collaborations with top-tier undergraduate-, masters-, and Ph.D.-level students and advising faculty. In particular, Dietrich College hosts a Statistics and Data Science Corporate Capstone program, which is focused on experiential learning and tied to a semester-long elective course. This program arose in response to two trends: the recent job market strongly pulled students toward industry careers, and summer internship opportunities are too competitive and restrictive for students (particularly those with summer visa constraints). Meetings occur both in-person and virtually, the experience concludes with student presentations, and both students and faculty receive financial incentives to participate.

Nugent noted the value of collaborating across disciplines, with attention to aligning logistics, project goals, and educational project agreements.

[2] The website for the Network of Academic Corporate Relations Officers is https://nacrocon.org/, accessed February 13, 2020.

The Statistics and Data Science Corporate Capstone program is governed by CMU's Educational Project Agreement, which includes language to define the relationship, nondisclosure terms, policies for data sharing, and the project cost and scope (CMU, 2017). This agreement protects the IP of the students and faculty. To begin building a network with industry, she suggested that faculty engage with their institutions' career centers to organize annual flagship events that draw potential partners to campus at low stakes.

### Mehran Sahami, Stanford University

Sahami explained that Stanford's academia–industry research collaborations in computer science often focus on innovations in artificial intelligence (AI), data science, human–computer interaction, computer science theory, security, graphics, systems, and biocomputation. He provided an overview of data science and AI collaborations at Stanford including the Stanford AI Laboratory,[3] which is a research laboratory and university-wide affiliated program (e.g., statistics, bioengineering, medicine) focused on machine learning, vision, natural language processing, and genomics. Common features of effective academia–industry engagement include formal and informal interactions among the company, faculty, and students; continuous two-way communication; facilitated access to research; and recruitment.

Many of Stanford's collaborations are housed in the Computer Forum,[4] which is the university's industrial liaison program. The Computer Forum brings together industry (more than 100 affiliate companies who each pay an annual membership fee of $21,000) and computer science and electrical engineering faculty and students for both research and recruiting purposes. The Computer Forum also hosts conferences, workshops, and symposia and gives financial support to the computer science and electrical engineering departments. Once a faculty liaison is assigned to a member company, mutual talks and visits occur, potential research collaborations are identified, and the company decides whether it would like to participate in a visiting scholar program to embed one of its researchers in a Stanford research laboratory. Stanford's Recruiting Program,[5] which is part of the Computer Forum, hosts information ses-

[3] The website for the Stanford AI Laboratory is http://ai.stanford.edu/, accessed February 13, 2020.

[4] The website for the Computer Forum is https://forum.stanford.edu/, accessed February 13, 2020.

[5] The website for the Recruiting Program is https://forum.stanford.edu/careers/recruiting.php, accessed February 13, 2020.

sions, on-campus interviews, career fairs, career workshops, company tours, office hours, and networking events.

Sahami also described a Stanford course with corporate engagement. Companies present a high-level problem for which they need a solution, and participating students do a two-quarter project to explore that area. The cost for each company to participate is $75,000, and there are more companies that want to participate than there are student teams available each year.

### Michael Franklin, University of Chicago and Formerly University of California, Berkeley

Franklin highlighted the University of California, Berkeley, success in creating multifaculty projects that engage industry. For example, the Berkeley Algorithms, Machines, and People Laboratory (AMPLab),[6] a big data research center, built the open source Berkeley Data Analytics Stack. AMPLab, a collaborative project, began in 2011 and concluded in 2016, resulting in 34 new faculty, several products, and four start-ups. A true public–private partnership, 50 percent of the funding for AMPLab came from NSF, the Defense Advanced Research Projects Agency, the Department of Energy, and the Department of Homeland Security, and 50 percent came from 40 industry partners. AMPLab nurtured its relationship with industry collaborators through twice-yearly retreats, during which faculty received feedback on project directions and students received feedback on research ideas. As part of its outreach and training initiatives, AMPLab also hosted AMPCamp,[7] a big data boot camp.

Franklin explained that building open source software is a valuable way for academia to collaborate with industry. However, a system cannot simply be built and passed on; a community has to be constructed and remain engaged (see Patterson, 2014). For example, AMPLab students created a meet-up group for Apache Spark, which now has more than 500,000 members across multiple meet-ups. He believes that AMPLab's approach was successful because its commitment to producing open source software and publishing vigorously nearly eliminated IP issues and fostered benefits for both industry and academia. Industry secured early access to ideas and plans, recruiting opportunities, and membership in a neutral community. Students accessed early adopters (and sometimes data), advice and mentorship, and internship and job opportunities, and practiced communicating their ideas. Faculty participated in

[6] The website for the AMPLab is https://amplab.cs.berkeley.edu/, accessed February 13, 2020.

[7] The website for AMPCamp is http://ampcamp.berkeley.edu/, accessed February 13, 2020.

a collaborative, flexible, diverse, and impactful platform; gained novel feedback; and received industry funding to augment federal grants.

The University of Chicago, however, is only newly involved in industry partnerships. Challenges to establishing these relationships include companies' limited perspectives about the value of academic research, companies' lawyers becoming involved too early in the process, and increased university competition for the attention of "enlightened" companies (e.g., Amazon, Google, Microsoft). Additionally, Franklin continued, administrators at some universities maintain outdated perspectives about IP and real-world engagement and fail to reward their faculty for industry collaborations. And some faculty underestimate the value of collaboration. To overcome these challenges, he suggested that institutions exploit local campus strengths and reach beyond a single department, as well as identify and exploit regional advantages where there is a concentration of universities, industrial strengths, and unique research assets (e.g., national laboratories). He wondered whether NSF could play a role in convening academia-industry partnerships, because its Computer and Information Science and Engineering division has already facilitated successful programs with several industry partners.

## PANEL DISCUSSION

### Establishing Partnerships

Nugent suggested that academic institutions dedicate time to develop a framework and educate industry about the potential benefits of partnership. Causgrove added that Dietrich College has coordinated with the other six colleges at CMU to ensure that all industry partners receive the same educational agreement—an especially important feature for faculty and companies new to partnerships. Victoria Stodden, University of Illinois, Urbana-Champaign, observed that because academic research is distinct from industrial research (in terms of problems and incentives), it is crucial to understand how the two can reinforce one another. She agreed that NSF could prompt such conversations and promote resource sharing. Franklin noted that although many complexities need to be addressed before partnerships can be established, a spectrum of research exists (as opposed to there being a distinction between academic and industrial research). Sahami added that academia–industry collaborations are responsible for much of the progress in deep learning; furthermore, more faculty could be inclined to leave academia for industry if silos between academic and industrial research persist. Mark Tygert, Facebook Artificial Intelligence Research, suggested that participants read the work of Yann LeCun as evidence of productive exchanges between academia and industry.

### Increasing Incentives

Charles Isbell, Georgia Institute of Technology, wondered how to change the culture of academia so that faculty are rewarded for engaging in partnerships. Sahami suggested that junior faculty structure partnerships around potential publications but noted that they sometimes avoid industry collaboration for fear that their Ph.D. students will leave academia for industry jobs. Franklin commented that faculty have to broaden their perspectives of promotion and reward systems (and then educate administrators)—especially in the evolving areas of computer science and data science, in which many definitions of success exist. Nugent said that CMU faculty receive summer research funding as a reward for helping with partnerships.

### Tracking and Replicating Success

Nicholas Horton, Amherst College, asked how the panelists' institutions have tracked their students' progress and wondered whether alumni serve as allies for these industry partnerships. Nugent said that CMU's Corporate Capstone program is not yet mature enough to assess the feedback loop, but, anecdotally, students are talking about the program at career fairs and recent alumni are promoting the program to their supervisors. Causgrove added that a number of senior-level alumni relationships have also been leveraged. Sahami reiterated that the key to successful partnerships is maintaining relationships over time. Kathleen McKeown, Columbia University, asked how to replicate these programs at scale, especially given the substantial amount of money companies contribute to participate. Franklin replied that although replicating AMPLab has proven more difficult than anticipated, he still believes that it is possible. He wondered whether industry could peruse NSF's pipeline of research proposals to prompt partnerships, and Nugent suggested that universities focus on engaging local companies.

## PANEL ON NATIONAL PERSPECTIVES ON ACADEMIA–INDUSTRY COORDINATION

*Ben Zorn, Microsoft, and*
*Leader of the Computing Community Consortium (CCC) Interim Report on "Evolving Academia/Industry Relations in Computing Research"*

Zorn described the mission of CCC (a standing committee of the Computing Research Association [CRA]) as to "catalyze the computing research community and enable the pursuit of innovative, high-impact research." A 2017 CRA survey showed that computer science enrollment

at the undergraduate level has more than quadrupled during the past 10 years, which makes it difficult for faculty to teach and maintain close relationships with students in large classes. He added that computing technology influences nearly all aspects of humans' lives; thus, interesting research challenges and rich opportunities for collaboration between computer science and other disciplines (e.g., transportation, health sciences, and biology) exist.

The CCC Industry Working Group was established in 2018 to better understand academia–industry relations. Its interim report (CCC, 2019) builds on the CCC's 2015 report *The Future of Computing Research: Industry–Academic Collaborations*. Anecdotal evidence in the interim report revealed a significant increase in faculty joint appointments in certain research areas, which could affect a university's culture and mission negatively (e.g., impact on research agenda, conflicts of interest and IP issues, decreased faculty participation on committees for admission and hiring, and decreased mentoring and face time with students). Because some joint appointments could have an indefinite duration, academic institutions might have to develop novel arrangements to cover 50 percent of each participating faculty member's time (or, in some cases, 80 percent), Zorn explained. He suggested the implementation of contracts as one way to ensure that students remain the priority of the faculty. Many positive outcomes of this type of engagement also exist. These experiences meet industry's increased demand for talent in an era ripe with access to data and computing capabilities. Faculty and graduate students have the opportunity to participate in ambitious and impactful research and to access increased resources and salary.

CCC's goal is to preserve the positive aspects of these academia–industry partnerships while understanding and mitigating risks. CCC hopes to expand data gathering, understand best practices of current faculty–student arrangements, and document novel company approaches to deepening academic engagement.

### Chaitan Baru, University California, San Diego

Baru observed that computer science and data science are optimal areas for collaboration with industry. During the past few years, NSF has facilitated a number of such interactions—for example, NSF BIGDATA,[8] NSF/Intel Partnership on Foundational Microarchitecture Research,[9] NSF

[8] The website for NSF BIGDATA is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767, accessed February 13, 2020.

[9] The website for the NSF/Intel Partnership on Foundational Microarchitecture Research is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505450, accessed February 13, 2020.

Campus Cyberinfrastructure,[10] and NSF Program on Fairness in AI.[11] He asserted that hands-on experience is essential for future data scientists and cited programs at North Carolina State University, Harvey Mudd College, and University California, San Diego (UCSD) as exemplars. He suggested that all data science curricula and faculty competencies should align with this vision of "clinical practice" to remain competitive.

Baru described the value of completing an "industry postdoc"—an immersive, practical experience with government agencies, nonprofits, or large or small companies (from Internet giants to start-ups). This experience could occur immediately after the Ph.D. is completed (in order to become better qualified for data science faculty jobs) or after the receipt of a job offer. A variety of modalities exists to fund such an experience (e.g., two-way between the agency and industry or three-way among the agency, industry, and university), and it should be governed by a mentorship plan that includes standards for compliance.

An example of an implementation vehicle for academia–industry collaboration is NSF's Grant Opportunities for Academic Liaison with Industry (GOALI).[12] There are currently 300 GOALI awards, only 2 percent of which are in computer science. In the future, Baru hopes that an NSF GOALI program will be created with net new funds and with programs for industry postdocs and industry sabbaticals. Baru concluded by noting that many opportunities exist for academia to collaborate with industry on technological innovation if the right engagement mechanisms are identified.

### Rachel Levy, Mathematical Association of America

Levy described the mission of the Mathematical Association of America (MAA) as "to advance the understanding of mathematics and its impact on the world." MAA provides guidelines for departmental reviews and experiential learning-based instruction, and it strives for mathematics to cross disciplines so that all people view themselves as mathematics "doers."

Levy shared examples of three MAA programs that relate to data science: (1) StatPREP,[13] which provides resources, workshops, and webinars for faculty on how to bring the modern tools and methods of data

[10] The website for Campus Cyberinfrastructure is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748, accessed February 13, 2020.

[11] The website for the NSF Program on Fairness in AI is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505651&org=NSF, accessed February 13, 2020.

[12] The website for NSF's Grant Opportunities for Academic Liaison with Industry is https://nsf.gov/pubs/2016/nsf16099/nsf16099.jsp, accessed February 13, 2020.

[13] The website for StatPREP is http://statprep.org/, accessed February 13, 2020.

science to elementary statistics courses; (2) PICMath,[14] which prepares mathematical sciences students for industry careers through a semester-long course on industry research problems as well as provides training, resources, and support for faculty teaching that course; and (3) Big Math Network,[15] which helps mathematics faculty, via the *Big Jobs Guide* (Levy et al., 2018), advise students interested in industry careers.

While tracking students who earned degrees in mathematics to better understand their job placement, MAA found that their job titles are rarely "mathematician." Levy observed that mathematicians do not always have a presence in academia–industry partnerships, despite their high level of interest. She suggested that industry could help MAA understand how to create meaningful experiences—building more partnerships, staying connected with mathematics graduates who accept jobs in industry, and creating challenges and competitions with broad participation that integrate data science—that would build competencies for future hires.

## PANEL DISCUSSION

### Balancing Faculty Responsibilities with Industry Experiences

McKeown reiterated the benefits of faculty joint appointments: faculty have the opportunity to work with interesting industry data and problems, to understand what students will experience when they enter industry, and to establish relationships that could lead to funding opportunities. Nugent noted that faculty who remain on campus and train the Ph.D. students whose advisors are unavailable need to be supported. Industry could sponsor faculty lines at universities to help alleviate this burden, she continued. Mark Green, University of California, Los Angeles, observed that no standards exist to protect students who have invested in the expertise of advisors who become unavailable, and it is unclear what body would have the credibility to suggest them. He appreciated the value of joint appointments but wondered whether there is a better process.

Emily Fox, University of Washington, said that her institution recently conducted a survey of Ph.D. students' perspectives on advising relationships: some students found it beneficial to have their advisors on leave and working in industry (e.g., increased access to resources), while more students thought that the advisors' decreased availability had a negative impact on the cohesiveness of their Ph.D. cohorts. Fox noted that while

[14] The website for PICMath is https://www.maa.org/programs-and-communities/professional-development/pic-math, accessed February 13, 2020.

[15] The website for the Big Math Network is https://bigmathnetwork.org/, accessed February 13, 2020.

the long-term benefit to Ph.D. students is immeasurable, faculty joint appointments take a substantial toll on Ph.D. students working on their dissertations. Green suggested that, to begin to address some of the concerns about faculty joint appointments and academia–industry partnerships, the mathematics community could compile a list of interesting problems that came from industry and led to important research. He also emphasized the need to use data to understand the capacity of the economy to absorb students being trained in Ph.D. programs. Levy noted that this conversation should be expanded to include VITAL faculty (an acronym for visitors, instructors, teaching assistants, adjuncts, and lecturers) as well as industry partnerships with faculty and students at 2-year colleges.

### Expanding Opportunities for Students

Alfred Hero III, University of Michigan, mentioned that the thriving economy in southeast Michigan has enabled the Michigan Institute for Data Science[16] to be successful in securing industry partnerships. However, because the competition is intense and industry partners often require exclusive nondisclosure agreements, universities run the risk of being limited to partnering with only one company. He suggested that universities engage more with national laboratories, which provide experiential learning on interesting problems without the competition. Deb Agarwal, Lawrence Berkeley National Laboratory, said that the data science community is doing a disservice to students if it continues to focus only on partnerships with companies instead of including government agencies, nongovernmental organizations, and national laboratories. She emphasized that national laboratories have an abundance of opportunities for students to work for the common good on unclassified research related to problems of national interest, generally without IP issues.

Catherine Brooks, University of Arizona, mentioned that her institution has developed a taskforce to identify synergies across the university to better present itself as a unified whole to industry partners. She explained that universities need to be more nimble and less siloed. Kolaczyk added that it is important to propagate lessons from experiential learning at the Ph.D. level across degree levels and across industries.

[16] The website for the Michigan Institute for Data Science is https://midas.umich.edu/, accessed February 13, 2020.

## A NEW MODEL FOR ACADEMIA–INDUSTRY PARTNERSHIPS

*Gary King, Harvard University (via webcast)*

King described a political science innovation that addresses the problem of data access for university researchers, motivated by the mission of social science to understand and solve problems that affect human society. King observed that the social sciences have access to more data than ever before, but these data are still a smaller fraction of the data that exist in the world. One goal of King's research is to understand how to incentivize private companies to release data for research that creates public good, without harming themselves.

King is working with Facebook to facilitate studies of the effect of social media on elections and democracy. This data-intensive research is funded by eight ideologically diverse charitable foundations that agreed to pool their funds and let one group of academics decide how to allocate grants. He asked Facebook for full access to its people, products, data, and platforms as well as freedom to publish without prepublication approval. Because Facebook would not agree to both of these terms for any one researcher, King created two groups of researchers: (1) a commission of distinguished academics at Social Science One,[17] an organization he created with Nate Persily at Stanford, who have signed nondisclosure agreements, have complete access to Facebook data, and have agreed not to publish; and (2) a group of outside academics who apply for limited data access and have complete academic freedom (i.e., no prepublication approval) to publish. Facebook, the foundations, and Social Science One agreed on the scope of the project, the commission identified relevant Facebook data sets and issued a request for proposals, and the outside academics applied for access to those data. There are three data sets to which access is now being provided: CrowdTangle, a collection of Facebook's political advertisements, and all of the public URLs shared on Facebook. The project will create its own surveys and will make arrangements with the American National Election Survey, the British National Election Survey, and other large academic surveys to include a question that asks respondents to share their Facebook data with the researchers. The outside academics will follow institutional review board processes and engage in a merit peer review and an ethical peer review, and the final decisions will be made by the commission. Facebook is building a privacy-preserving computer infrastructure.

[17] The website for Social Science One is https://socialscience.one/, accessed February 13, 2020.

The timeline for this innovative project has been extended and has included challenges such as dozens of legal agreements. When researchers receive data access (as opposed to data), the academic research model changes from one of individual responsibility to one of collective responsibility. King's goal is to convey to companies and the public that data are an asset to create social good and solve the world's problems, while preserving privacy.

## PANEL ON INDUSTRY ACTIVITIES AND EXPERIENCES FROM ACADEMIC PARTNERSHIPS

### Mike Willardson, Facebook

Willardson described Facebook's mission as to "give people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what is going on in the world, and to share and express what matters to them." As of 2018, Facebook had 35,000 employees and 2.32 billion monthly active users. Willardson provided an overview of the research activities within Facebook's Research[18] Operations and Academic Relations division. The research activities vary by subject matter; for example, IP is an important concern for research in augmented reality/virtual reality because it is used in commercial products. Facebook believes strongly in building community through open source technology, and investing in open source increases employee retention and recruitment.

Willardson described several innovative Facebook partnerships. The Open Compute Project[19] democratizes hardware by bringing industry and universities together to build products. This mechanism works well for multiple industry partners because there are no exclusive rights, and everyone benefits equally. The Telecom Infra Project[20] is a collaborative effort to build and deploy telecommunications network infrastructure. Facebook also engages with faculty and students through fellowship programs, emerging scholar awards, research awards, research collaborations, and visiting researcher and postdoctoral positions. He added that Facebook establishes broad master agreements with universities to cultivate long-term relationships.

[18] The website for Facebook Research is https://research.fb.com/, accessed February 13, 2020.

[19] The website for the Open Compute Project is https://www.opencompute.org/, accessed February 13, 2020.

[20] The website for the Telecom Infra Project is https://telecominfraproject.com/, accessed February 13, 2020.

### Mary Ellen Sullivan, MassMutual

Sullivan explained that MassMutual operates for the benefit of its members and participating policy holders by helping people secure their futures and protect their loved ones. MassMutual has 7,500 employees and 9,000 nationwide advisors. MassMutual employs 100 data scientists in four data science domains—risk and product, operations, finance investments, and marketing and sales—to enable data-driven decision making throughout the enterprise.

Sullivan explained that academia–industry partnership is essential at MassMutual. The company supports science, technology, engineering, and mathematics curricula and programs; engages with local faculty; co-sponsors community education and events; engages with student groups; invests in training and development programs; and collaborates on research initiatives with university partners. Smith College, Mount Holyoke, and the University of Massachusetts, Amherst, each have partnerships with MassMutual, and the University of Vermont will be the company's next collaborator. MassMutual works with university administration, faculty, and student groups to ensure that programs are working effectively and offering mutual benefits. In 2014, MassMutual launched the Data Science Development program,[21] and it will launch a Data Engineering Development program[22] in summer 2019. Each cohort of the Data Science Development Program has four to eight participants, 80 percent of whom are women. Both programs offer hands-on training and mentorship, full-time employment on an innovative and fast-paced team, and tuition sponsorship for either a master's degree or a certificate from a local university. In January 2018, MassMutual hosted a Women in Data Science Conference,[23] and it hosts monthly data science meet-ups in Boston, Data Days for Good,[24] and hackathons.

### Peter Norvig, Google

Norvig said that one of Google's most significant responsibilities is to help grow the field of data science, starting at the K-12 level by developing curriculum and educating teachers. Google supports Girls Who

[21] The website for the Data Science Development program is https://datascience.massmutual.com/dsdp, accessed February 13, 2020.

[22] The website for the Data Engineering Development program is https://datascience.massmutual.com/dedp, accessed February 13, 2020.

[23] The website for the Women in Data Science Conference is http://www.science.smith.edu/wdsboston/, accessed February 13, 2020.

[24] The website for Data Days for Good is https://blog.massmutual.com/post/data-days-good, accessed February 13, 2020.

Code,[25] as well as groups within historically black colleges and universities to develop and co-teach classes. Google's own educational materials (some are co-developed with Coursera or Kaggle) are available through massive open online courses. Google is reviewing its guidelines for data sharing and is promoting academia–industry collaborations that will develop responsible and productive researchers by hiring interns, welcoming visiting faculty, and offering faculty joint appointments. Norvig emphasized that when a faculty member decides to leave academia for a career in industry, that move should be viewed as a new opportunity (not a failure). Likewise, Google staff are encouraged and supported to co-advise students and to teach in the classroom or online.

**Daniel Marcu, Amazon**

Marcu noted that members of industry and academia alike should be making efforts to enhance their communication and collaboration. Amazon has a variety of collaborative engagement models and a significant research breadth (e.g., hardware, economics, sustainability, logistics, avionics, robotics). Students can participate in 3-month internships or full-time postdoctoral opportunities as well as apply for research grants and Amazon Web Services credits. Faculty can apply for academic grants, secure Amazon Web Services resources and data, and attend Tech Talk Series and academic conferences. The Amazon Scholars program[26] offers deeper levels of engagement by enabling professors to work on Amazon's large-scale, high-impact technical challenges without leaving their academic institutions. Amazon Community Programs include a graduate research symposium (which pairs student researchers with Amazon's scientists to exchange new innovations and research concepts), scientific meeting sponsorships, and an internal academic advisory council.

When developing partnerships with industry, Marcu suggested that faculty need to understand the potential partner, consider the best-suited model of engagement, and formulate interesting proposals. Administrators could aid in the process by simplifying engagement models. Inhibitors to success include faculty members who dictate terms to the partners and write ineffective proposals as well as administrators who treat industry engagements as one-off activities. Marcu believes standardized agreements could accelerate collaboration.

[25] The website for Girls Who Code is https://girlswhocode.com/, accessed February 13, 2020.

[26] The website for the Amazon Scholars program is https://www.amazon.jobs/en/landing_pages/scholars, accessed February 13, 2020.

## PANEL DISCUSSION

### Engaging Students

Chris Mentzel, Gordon and Betty Moore Foundation, asked the panelists how often their companies engage with disciplines that intersect with data science. Norvig acknowledged that the majority of Google's interactions are with computer scientists and said that it can be difficult to advertise for and evaluate proposals from other fields without appropriate expertise on staff. Marcu said that Amazon engages frequently with economists, computer scientists, data scientists, and machine learning experts. Willardson noted that Facebook engages often with data scientists who have expertise in artificial intelligence, machine learning, connectivity research, and natural language processing, and Sullivan commented that MassMutual's engagement extends beyond the discipline of data science.

Duncan Temple Lang, University of California, Davis, asked the panelists what skills students need to be prepared for industry careers. Levy suggested Kaggle as a useful tool for mathematics Ph.D.s who want to move to industry. Sullivan said that MassMutual emphasizes skills that are essential for business but rarely developed at the undergraduate level, such as leading, giving and getting feedback, and tailoring presentations to different audiences. MassMutual began a partnership with EdX and is establishing requirements around a series of self-paced online courses to help reinforce these skills. Norvig agreed that these skills are crucial, especially the ability to work effectively in teams and to give meticulous attention to detail. Marcu noted that it would be beneficial for students to understand that academic research is not inherently superior to industry research. Nugent and Kolaczyk suggested that members of academia and industry avoid referring to these skills as "soft skills." Not only is it offensive to the fields that teach these skills, but also such language causes students to drastically underestimate how important those skills are and how difficult they are to learn.

### Navigating Two Cultures

In response to a question from Causgrove about formalizing academia–industry partnerships, Marcu said that although many conversations are happening at different levels across academia and industry, it can be difficult to bridge the communication gap and begin to move forward with effective partnerships. Baru noted that it is easier to partner with companies that understand the culture of academia, and he suggested that those companies help others in industry to better understand the research ethos. Willardson agreed that sharing best practices throughout industry would improve consistency. Setting the context and determining

the value proposition before entering into partnership is also effective, he continued. Norvig mentioned that there are different measures of successful partnerships—professors need to publish, while industry teams are recognized for research even if it leads to failure. Tygert mentioned an agreement between Facebook and the University of California, Berkeley, to share students. Instead of negotiating separate agreements, Google, Amazon, and others signed on to this agreement. Hero said that the flow of students and faculty has moved away from academia and toward industry during the past 5 years; he wondered how to reinforce positive relationships between industry and academia and reverse this imbalance. Levy wondered what mechanisms would motivate industry employees to embrace teaching or training opportunities. Antonio Ortega, University of Southern California, asked about strategies to attract junior faculty to partnerships. Sullivan said that MassMutual's Data Engineering Development program has an academic advisory board that includes junior-level faculty, and Marcu noted that the number of opportunities in general for junior faculty has increased.

Sahami asked whether companies have policies for the length of visiting faculty terms. Sullivan replied that the faculty going to MassMutual are only joining an academic advisory board or teaching one-off in-house workshops and that academic institutions welcome that level of cross-pollination. Willardson said that Facebook defines a limited term for visiting faculty, and Norvig said that although Google supports freedom of choice, it recognizes that there can be negative repercussions from extended faculty appointments and tries to maintain good relationships with partnering departments.

James Frew, University of California, Santa Barbara, asked about impediments (beyond IP issues) to these partnerships. Willardson said that both partners must be willing to accept some level of calculated risk in order for the partnership to be successful. At MassMutual, the issue is less about risk and more about workforce: because MassMutual is building pipelines for people to enter its organization, it can be challenging to keep pace with changing skills and relevant curricula. Marcu said that the biggest hindrance is the lack of well-established models of collaboration.

### Sharing Data in Partnership

Noting the growing trend to provide artifacts alongside publications (e.g., the data and code that support a paper's claims), Stodden inquired about policies for sharing artifacts that emerge from collaborative work. Marcu said that this trend presents an opportunity for industry, not a barrier to participation in partnerships. Willardson explained that the subject matter will determine whether Facebook pursues sponsored research

agreements (e.g., user data can be shared only in a controlled environment and with prepublication review, so such research is unlikely to be part of these agreements). Facebook is not trying to control outcomes in this case; rather, it is trying to prevent the inadvertent dissemination of confidential information. Norvig suggested that industry provide funding for open source journals and noted that increased partnership among nonprofits, academia, and industry is needed to address issues of data ownership and proprietary publishers. The University of California, for example, recently stopped paying for use of Elsevier. Mark Krzysko, U.S. Department of Defense, emphasized that sharing and consuming data are complex in part because of challenges with access and dissemination and a lack of clear policies. Norvig said that Google employees have access to internal data, while grant recipients do not. To establish mutually beneficial partnerships, industry needs to make more relevant nonproprietary data sets available and help pose more germane problems. Sullivan said that MassMutual will never share clients' confidential data. Other publicly available data, however, are used for research (e.g., for health and longevity studies, which can be used to provide information to customers).

Baru noted the success of NSF's Computer Science for All initiative[27]— however, part of the curriculum has languished because teachers did not have access to data sets. It would be helpful if industry partners would contribute data (real or synthetic) for teachers to use. Zorn said that it is important to find the right technology that will empower companies to share data by preventing unauthorized access and highlighting mutually beneficial opportunities of data sharing. Sahami suggested a new model for data sharing in which third-party public institutions are leveraged to socialize the associated risk instead of having either the company or the researcher assume the risk.

## BREAKOUT GROUP DISCUSSIONS

Following the presentations and open discussions, roundtable participants divided into three groups to create sketches of Ten Simple Rules[28] for Creating a Successful Academia–Industry Collaboration at the levels of undergraduate, master's, and Ph.D. education. These sketches represent collections of diverse ideas and are not meant to be read as consensus viewpoints. A representative from each group summarized the discussions among the breakout group members as follows:

[27] The Computer Science for All website is https://www.nsf.gov/news/special_reports/csed/csforall.jsp, accessed February 13, 2020.

[28] Inspired by PLOS's Ten Simple Rules series (see https://collections.plos.org/ten-simple-rules).

Hunter Glanz, California Polytechnic State University, presented the following suggestions for effective collaborations between industry and undergraduate students: (1) keep curriculum current and exploit curricular flexibility; (2) offer experiential learning opportunities early and often; (3) ensure that both parties continually benefit from the interactions; (4) offer capstone experiences; (5) promote early comprehensive experiences (starting with an open-ended problem and working through to the communication of findings) in which students have to make choices; (6) provide multiple points of inclusive entry for data science learners; (7) educate both parties on data ethics; (8) develop a mutual understanding of unique cultures and environments; (9) provide genuine and varied data sources in a consistent manner; (10) create a reproducible, transferrable data science best practices kit; and (11) promote classroom and company visits.

Kolaczyk highlighted the following suggestions for successful partnerships between industry and master's-level students: (1) take a holistic approach to training, rather than teaching topics in separate silos; (2) build skills in communication and team interaction; (3) create opportunities for repeated practice; (4) expose students to industry in multiple ways and at many levels; (5) encourage humility and reduce anxiety among faculty and students; (6) become an active listener and learn to use vocabulary that is conducive to collaboration; (7) nurture academia–industry relationships; (8) define collaborative projects through an iterative process, with both parties vested; (9) own the collaboration on both sides; and (10) lay the intellectual groundwork before involving lawyers.

Nina Mishra, Amazon, shared her breakout group's discussion of considerations for fruitful collaborations between industry and Ph.D. students: (1) consider creating a Ph.D. in data science; (2) encourage students to do multiple data science internships; (3) create a consortium of industry collaborators who contribute data and problems; (4) ensure that all parties agree on a project and its duration before it begins; (5) create prolonged internship opportunities; (6) encourage open source and open science; (7) identify potential conflicts of interest ahead of time; (8) formally include internship work in the thesis; (9) avoid letting industry drive what happens to students; and (10) maintain a high bar for dissertation work and graduation.

# 12

# Meeting #11: Data Science Education at Two-Year Colleges

The 11th Roundtable on Data Science Postsecondary Education convened virtually on June 12, 2019. Stakeholders from data science education programs, government agencies, nonprofit organizations, professional societies, research organizations, foundations, and industry discussed current efforts in developing data science curricula and programs at 2-year colleges, opportunities for professional development in data science education, strategies for building partnerships with nearby 4-year and master's-granting institutions, and techniques for understanding the needs of local employers. 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.

Welcoming roundtable members and participants to the meeting, co-chair Kathleen McKeown, Columbia University, noted that as tuition increases at 4-year institutions across the United States, enrollment at more affordable 2-year colleges continues to grow. At the same time, demand for employees with data science skills is expanding across industries. In light of these trends, participants explored emerging approaches for integrating data science into 2-year curricula as well as strategies to enable connections between 2-year colleges and other postsecondary institutions.

## SETTING THE LANDSCAPE: TWO-YEAR COLLEGES AND DATA SCIENCE EDUCATION

*Nicholas Horton, Amherst College*

Horton emphasized the important role that 2-year colleges play in the education system and in the development of a diverse and inclusive workforce. More than 6 million students are enrolled at 2-year colleges, representing approximately one-third of the total undergraduate student population in the United States. A large proportion of Pell Grant recipients, U.S. veterans, and historically underrepresented students are enrolled in 2-year colleges, where "open-door policies" provide accessible, affordable pathways—average annual tuition is approximately $4,000. The committed educators and administrators who are focused on student success and community engagement in 2-year colleges face distinct structural and organizational challenges, Horton explained. He referenced the December 2018 roundtable meeting in which speaker D.J. Patil, head of technology at Devoted Health and former Chief Data Scientist in the White House Office of Science and Technology Policy, highlighted the value of 2-year colleges, especially in light of the interdisciplinary nature of data science. Patil explained that his experience at a 2-year college gave him "three gifts": a love of mathematics, an understanding of how to write in various genres, and confidence to succeed at the postsecondary level. He considered this experience to be a crucial "on-ramp" to his future success.

Horton provided an overview of the 2018 National Academies' consensus study report *Data Science for Undergraduates: Opportunities and Options*, which identified 10 components of "data acumen": (1) mathematical foundations, (2) computational foundations, (3) statistical foundations, (4) data management and curation, (5) data description and visualization, (6) data modeling and assessment, (7) workflow and reproducibility, (8) communication and teamwork, (9) domain-specific considerations, and (10) ethical problem solving (NASEM, 2018b). He noted the need for students at 2-year colleges to develop the appropriate depth of understanding in each of these areas and for faculty to have the time and resources to support such learning opportunities. Select recommendations from the report include ensuring that 2- and 4-year institutions work together on issues related to data science education, training, and workforce development; attracting and retaining students who have varied backgrounds and levels of preparation to data science programs; and remaining flexible and developing incentives as programs evolve. The National Science Foundation (NSF) funded the Two-Year College Data Science Summit to highlight innovative and effective programs at 2-year colleges, delineate three pathways to serve students' unique needs (i.e.,

certificate, associate-to-transfer, and associate-to-workforce programs), and identify next steps. Recommendations that emerged from that summit included (1) creating courses with modern and compelling introductions to statistics, (2) ensuring opportunities to engage with realistic problems and real data, (3) reducing barriers to entry, (4) ensuring depth in algorithmic thinking, (5) requiring fluency in computational language, (6) infusing ethics, and (7) fostering active learning (Gould et al., 2018).

In closing, Horton reiterated that data science is not just for doctoral-, master's-, or bachelor's-level students and that a 2-year education offers the "only affordable game in town." Major changes in pedagogy and course content are under way in science, technology, engineering, and mathematics (STEM) pathways to ensure student success. With best practices in flux, he continued, the need for professional development and continuing education is increasing—faculty need incentives, time, and resources to prepare to teach data science. He provided a series of framing questions for the remaining sessions of the meeting:

- How do we ensure that data science programs attract and retain students with varied backgrounds?
- How do we ensure that faculty development programs are robust and effective?
- How do we develop curricula that instill data acumen and are responsive to workforce needs?
- How can we build/maintain/grow a 2-year college data science community?
- How do we build effective connections between 2- and 4-year institutions?
- How do we build effective connections between 2-year colleges and industry?

## PANEL ON INTEGRATING DATA LITERACY INTO COURSEWORK AND DEVELOPING DATA SCIENCE PROGRAMS

*Randy Kochevar, Oceans of Data Institute, EDC*

Kochevar explained that a dramatic change has occurred during the past several years: many people now learn about the world through streams of data from remote sensors. This presents a challenge for educators who are introducing students to data. Instead of working with personally collected data sets (i.e., dozens of measurements), students can now work with much larger data sets (on the scale of many megabytes). At the same time, visualization skills have become more sophisticated. Kochevar proposed that all K-16 institutions have a responsibility to help

students develop the required skills to work with complex data sets. Because little research has been conducted on how to best cultivate these skills and limited awareness exists as to the value of data science educator training, he said that new strategies to "educate the educators" are essential. The Education Development Center's Oceans of Data Institute (ODI) promotes the data literacy of K-16 students by building a research-based learning progression, developing and testing curricula and tools, and acting as a hub to convene diverse stakeholders.

Using the acronym CLIP, Kochevar described big data as Complex (i.e., different types of data collected in different ways), Large (i.e., more data than would be needed to answer a specific question), Interactive (i.e., data visualization tools can be used to compare different data sets), and Professionally collected (i.e., not by students). He explained that by studying how people use data in the real world, it becomes possible to understand the foundational skills that students need to develop. To aid in this effort, ODI has created expert worker profiles, including the Profile of a Big-Data-Enabled Specialist (ODI, 2014) and the Profile of the Data Practitioner (ODI, 2016). In closing, Kochevar offered a definition of data literacy in the age of big data: "The data literate individual understands, explains, and documents the utility and limitations of data by becoming a critical consumer of data, controlling his/her personal data trail, finding meaning, and taking action based on data. S/he can identify, collect, evaluate, analyze, interpret, present, and protect data."

### Joyce Malyn-Smith, Oceans of Data Institute, EDC

Malyn-Smith reiterated that ODI works with educators to incorporate data skills into curricula and develops tools for big data career pathways, based on input from industry. ODI's tool kit includes (1) expert worker profiles, (2) rubrics to guide assessment, (3) a gap analysis tool for assessing industry value and school capability, (4) a curriculum analysis tool, (5) a course planning tool, and (6) a stackable credentials model (EDC, 2017). Most recently, ODI established partnerships with four 2-year colleges (Normandale Community College, Bunker Hill Community College, Johnson County Community College, and Sinclair Community College) through an NSF-Advanced Technological Education (ATE) project titled Creating Pathways for Big Data Careers.[1]

Malyn-Smith described her work at ODI trying to identify and articulate what data-related skills are used in the workplace and how local

[1] The website for Creating Pathways for Big Data Careers is https://www.nsf.gov/awardsearch/showAward?AWD_ID=1501927&HistoricalAwards=false, accessed February 13, 2020.

2-year colleges could incorporate them into their curricula. By interviewing workers and convening focus groups, ODI developed expert worker profiles (e.g., the data practitioner, the data scientist) to capture the broad range of skills, knowledge, and behaviors that are required to be successful in different roles in the workplace. Educators could use these profiles to identify where and how in their curricula these skills are covered, often leading to curricular modifications. Similarly, students could use the content and vocabulary in these profiles to sharpen their résumés, and employers could use them to evaluate employee performance and create balanced teams. Another method Malyn-Smith described that can help 2-year colleges align their curricula with employer needs is to conduct a gap analysis—asking industry partners about their expectations for employees to complete specific tasks and asking educators how their curricula prepare students to complete these tasks. Comparing responses could help to identify gaps in student training, she added.

### Paul Hansford, Sinclair Community College

Recalling Sinclair Community College's 1887 motto—"find the need and endeavor to meet it"—Hansford discussed the institution's innovative approaches to supporting its students. He noted that "closing the skills gap is part of [Sinclair's] DNA." Sinclair serves 30,000 students annually, offers more than 270 programs of study, has the lowest tuition in the state of Ohio, and is the largest workforce education provider in its region. It is among the top 5 percent of the 1,100 2-year colleges in the United States in terms of enrollment size, physical plant, and the variety and complexity of educational programs of study, according to Hansford. He noted that new programs at Sinclair have arisen in response to both market demands and the desire to embed data literacy, analytics, and science into decision making that benefits communities. He added that data should act as the foundation for decision making, not as a substitute for human judgment. He also observed that publicly available data tools are surfacing—for example, a 70-question data literacy exam[2] and a resource on 17 character traits of a data-literate person.[3]

Sinclair offers three data programs via its Department of Computer Information Systems. A 1-year technical certificate in data analytics[4] has

[2] The website for the data literacy exam is https://thedataliteracyproject.org/assessment, accessed February 13, 2020.

[3] The website for this resource is https://dataliteracy.com/resources/, accessed February 13, 2020.

[4] The website for the technical certificate in data analytics is https://www.sinclair.edu/program/params/programCode/DA-S-CRT/, accessed February 13, 2020.

been available since Fall 2012; the Data Analytics Associate of Applied Science degree[5] has been available since Fall 2018; and Data Fundamentals,[6] a short-term technical certificate, will be offered in Fall 2019. Similar degree programs and/or certificates are also available across the disciplines of business information systems, geography, allied health, and marketing. He described Sinclair's partnerships with ODI and NSF, which have helped to increase student interest as well as prompted inquiries from other institutions about replicating programs and acquiring resources. Hansford related that Sinclair's future goals include learning from students' field experiences, adjusting courses to meet the needs of the local market, working closely in mentorship with other institutions, and spreading domain-specific certifications across disciplines.

**Michael Harris, Bunker Hill Community College**

Harris described Bunker Hill Community College as a diverse campus, with a student population that is 25 percent African American, 25 percent Hispanic, 25 percent Caucasian, and 25 percent other. He explained that Bunker Hill has a three-phase data analytics program, which was devised based on ODI's stackable credentials model (mentioned above). Bunker Hill also used ODI's Profile of the Data Practitioner and its associated heat map to understand core competencies for data practitioners and to determine student learning outcomes.

Harris explained that the first phase of the data analytics program is a data management certificate,[7] available since Fall 2015. Students receive an introduction to data science and data management, learn to work in groups, and solve real-world problems. This curriculum includes the following five courses: IT Problem Solving, Introduction to Big Data, Statistics, SQL Programming, and Advanced Excel. The data analytics certificate,[8] first offered in Fall 2017, is the second phase of the data analytics program. Students who have completed the data management certificate only have to take four additional courses—Data Analytics and Predictive Analytics, Python Programming, Database Programming, and Operating Systems—to earn the data analytics certificate. Students who start in this second phase of the program would have to

[5] The website for the Data Analytics Associate of Applied Science degree is https://www.sinclair.edu/program/params/programCode/DATA-S-AAS/, accessed February 13, 2020.

[6] The website for Data Fundamentals is https://www.sinclair.edu/program/params/programCode/DF-S-STC/, accessed February 13, 2020.

[7] The website for the data management certificate is https://www.bhcc.edu/programsofstudy/programs/datamanagementfast-trackcertificateprogram/, accessed February 13, 2020.

[8] The website for the data analytics certificate is https://www.bhcc.edu/programsofstudy/programs/dataanalyticscertificateprogram/, accessed February 13, 2020.

complete all nine courses, Harris explained. The third and final phase of the data analytics program, an associate's degree in data analytics, will be offered for the first time in Fall 2019. To attain the degree, students must take a total of 10 core data courses (i.e., the nine previously listed courses and one in data visualization), two general education science courses (which enable students to transfer to a STEM program in certain 4-year institutions in Massachusetts),[9] four additional general education courses (including two English courses), and three to four mathematics courses (e.g., pre-calculus, calculus, statistics, and linear algebra).

## TEACHING DATA LITERACY IN THE CONTEXT OF ADVANCING WORKPLACE TECHNOLOGY

*Ann-Claire Anderson, Center for Occupational Research and Development*

Anderson described Preparing Technicians for the Future of Work,[10] a project funded by NSF to enhance STEM programs in advanced technology fields for 2-year colleges. The project was developed in response to several issues: the nature of work is changing rapidly; advanced technologies are eliminating some jobs and creating others; NSF's 2016 10 Big Ideas[11] emphasize a new research agenda, including the "Future of Work at the Human–Technology Frontier"; and technicians are at the center of much of this "disruption." The mission of the project is to "enable the NSF-ATE community to collaborate regionally with industry partners, within and across disciplines, on the transformation of associate's degree programs to prepare U.S. technicians for the future of work," she continued. As the project team tries to make predictions about the workforce in 2030, it considers the following industry-agreed-upon interconnected technologies: big data, autonomous robots, simulation, system integration, Internet of Things, cybersecurity, cloud computing, additive manufacturing, and augmented reality. The project is based on five suppositions: (1) technology will continue to evolve in a cross-disciplinary way, (2) technicians will need a multidisciplinary skill set, (3) some new skills will emerge that are common across multiple technologies, (4) the core knowledge that all technicians must possess

[9] The website for the MassTransfer program is https://www.mass.edu/masstransfer/, accessed February 13, 2020.

[10] The website for Preparing Technicians for the Future of Work is https://www.preparingtechnicians.org/, accessed February 13, 2020.

[11] The website for NSF's 10 Big Ideas is https://www.nsf.gov/news/special_reports/big_ideas/, accessed February 13, 2020.

will need to be augmented, and (5) 2-year technical programs will need to adapt their curricula.

Anderson explained that the project team conducts employee interviews, visits industry sites, and convenes employers and educators on both a national and regional level to better understand issues that postsecondary institutions face relative to the future of work. From its conversations with educators and industry representatives, the project team has identified three sets of cross-cutting skills: data knowledge and analysis, business knowledge and processes, and advanced digital literacy. She described a number of pathways to develop these skills: 2-year college transfer programs, 1-year certificate programs, degree programs aligned with field specialization, stand-alone courses, microcredentials, advanced coursework for returning professionals, and bootcamps/continuing education. Anderson noted that because many supervisors want to hire people with industry-relevant experience, technicians often go to work immediately after attending a 2-year college instead of pursuing a 4-year degree. Because much of the data science work being performed today will be completed by people with 2-year degrees, she suggested that data analysis be integrated into technical programs and taught in the context of real work—technicians need to be able to manipulate, interpret, compare, contrast, merge, and operate on data to resolve problems, while using Excel and other common software. This requires institutions to revise mathematics prerequisite courses to reflect the changing demands of the skilled workforce.

She concluded that the project continues to examine what the future holds for STEM education at the associate's level. Next steps include interviewing skilled technical workers about new technologies and needed skills, convening educators and chief executive officers who represent a range of technical disciplines, adopting existing competency frameworks (from ODI and the U.S. Department of Labor) to identify specific skills required by industries of the future, developing recommendations for associate's degree programs in advanced technology, collaborating with 2-year colleges and companies to facilitate the implementation of recommendations, and facilitating the ongoing work of regional networks dedicated to training technicians for the future.

## ROUNDTABLE DISCUSSION

Victoria Stodden, University of Illinois, Urbana-Champaign, wondered how 2-year colleges bridge the needs of different student populations—for example, those who transfer to 4-year colleges and those who enter the workforce. Anderson emphasized the distinct differences between the pathways for these populations. She noted that institutional

structure and/or state requirements influence a 2-year college's ability to serve both populations. Malyn-Smith pointed out that when analyzing courses across four 2-year colleges to create the stackable credentials model, the only difference between data science pathways for students who planned to transfer and those who planned to enter the workforce was one mathematics course—this demonstrates that 2-year colleges likely can provide appropriate trajectories for a variety of students. Horton agreed and pointed to California's system of higher education, in which courses are clearly mapped for students to transfer from a 2- to a 4-year institution. This transition is even smoother for dual enrollees (i.e., high school students taking community college courses), in his view.

Uri Treisman, University of Texas, Austin, described data science programs as "powerful resources for students seeking upward mobility." He wondered about student enrollment in Sinclair's and Bunker Hill's data programs. Harris said that, historically, approximately 70 percent of Bunker Hill students have been students with undergraduate or graduate degrees who were seeking data science skills for industry jobs, and 30 percent have been people seeking first-time degrees. Similarly, Hansford said that 80 percent of Sinclair's certificate students are people who are retooling. Jeffrey Ullman, Stanford University, and Treisman asked about the role of traditional foundational coursework (e.g., mathematics, statistics, business, and/or computer science) versus more applied courses in these data science programs. Specifically, Ullman wondered whether emerging data science programs deemphasize the study of methods and foundations. Harris explained that after consulting with representatives from industry, Bunker Hill decided to add data science projects to fundamental courses so that students would receive a balanced education. Hansford noted that Sinclair's data curriculum includes several classes that emphasize fundamental content (e.g., programming, operating systems, mathematics, statistics) as well as additional courses to align with state requirements. Treisman pointed out that employers play an important role in the survival of institutions; he asked how to manage programs so that they best serve students and meet the demands of both the institutions and industry. Hansford said that Sinclair conducts annual reviews with industry to discuss its curriculum and plans to seek feedback from alumni in the field. Harris said that he meets each semester with a representative from industry and is currently setting up articulation agreements with 4-year institutions.

## BREAKOUT GROUP DISCUSSION: DATA SCIENCE CAREERS AND INDUSTRY PARTNERSHIPS

Horton (moderator) posed a question about the level of skill mastery expected with an associate's degree in data science. Anderson noted that data continue to be collected from companies and technicians to understand what skills are required by the workforce and what type of on-the-job training is available. For example, critical thinking might be more important for a technician in a particular role than a specific mathematics skill set. According to Malyn-Smith, ODI's profiles and rubrics are continually revised based on feedback from current practitioners. Horton then asked how to determine which foundational skill sets are better suited for an associate's degree than a bachelor's degree. Malyn-Smith noted that although the biotechnology industry originally sought individuals with bachelors' degrees, as 2-year colleges enhanced their programs, employers found that individuals with associates' degrees were well suited for many of their jobs.

Horton asked how students make a smooth transition from a 2-year college to the workforce. Shalita Giannini, Milwaukee Area Technical College, noted the importance of integrating hands-on projects and assessments that relate to the real world to best prepare students for the workplace. In response to a question from Mark Tygert, Facebook Artificial Intelligence Research, Horton said that while co-ops are popular at 4-year institutions, they are starting to emerge at 2-year colleges. He noted that capstone projects with realistic expectations also provide valuable training for students. Anderson suggested that students do apprenticeships or internships—strong partnerships are needed between employers and institutions in order for these to be worthwhile experiences. Malyn-Smith agreed and proposed that institutions consult employer advisory boards when designing and revising programs. Anderson added that early recruiting strategies and dual-enrollment opportunities also show promise. Asia Mieczkowska, University of North Carolina, Chapel Hill, said that aligning workforce needs with broader foundations is important. Malyn-Smith remarked that some simple strategies are being overlooked, such as inviting guest speakers to class or taking students to visit companies. She and Anderson added that having instructors visit employers could also be helpful. Tyler Kloefkorn, National Academies, asked how to foster collaboration between academia and industry. To begin a partnership, Anderson suggested engaging colleagues who have technical connections as well as designing multidisciplinary courses to help build bridges within a community. In response to a question from Jennifer Travis, Lone Star College, Horton said that while buy-in from multiple programs is important, the 2-year data science landscape is heterogeneous

and a clear set of best practices for creating these partnerships does not yet exist. Angelika Gulbis, Madison Area Technical College, added that her institution employs a liberal arts internship coordinator who maintains relationships with industry partners. Horton noted the value of scaling and replicating such models while maintaining flexibility.

## BREAKOUT GROUP DISCUSSION: DATA SCIENCE LITERACY, CURRICULA, CERTIFICATES, AND DEGREES

Kochevar (moderator) explained that this discussion would focus on how 2-year colleges decide whether to offer degrees or certificates. Jean Wilson, Carroll Community College, proposed consulting local employers—for example, would they hire a student who has a certificate instead of a degree? Nicki Kowalchuk, Milwaukee Community College, noted the difficulty in motivating employers to accept 2-year college graduates for data analyst positions and expressed a broader concern that a 2-year degree may not be sufficient for most employers. Kochevar shared his experience working with Columbia College and regional businesses to develop an internship program to help bridge this gap between 2-year colleges and local employers. He described this as an effective way to evaluate how students fit into the work environment when leaving their degree or certificate programs. Kowalchuk responded that although Milwaukee Community College has established a partnership with Northwestern Mutual and is seeing increases in the employment of 2-year graduates, a master's degree is still preferred by many employers. In response to a question from Kelley Engle, Harrisburg Community College, Hansford responded that businesses are receptive to Sinclair's 1-year certificate program, which primarily serves students with 4-year degrees who are seeking to add a specific skill set. Treisman noted that Indian River Community College, Alamo College, and Austin Community College have long-term relationships with employers and might have best practices to share (e.g., colocation of facilities at community colleges).

Kochevar asserted that data literacy will eventually be part of every job. He said that students need to develop skills, starting in elementary school, that will allow them to move in and out of the world of mathematics gracefully through quantitative thinking. He asked how best to build data literacy into 2-year college curricula. Linda Grisham, Massachusetts Bay Community College, noted that NSF has promoted data literacy (e.g., through its BioQUEST and QUBES programs), but disciplines remain siloed. She added that faculty need professional development to change their approaches. Hansford proposed that traditional literacy (i.e., reading, writing, and mathematics) be reconfigured to

include courses in data visualization, Python, and R. Treisman explained that local employers seek employees with general data savviness and that quantitative literacy and data acumen are becoming increasingly important at 4-year institutions. Harris said that, to prepare students who plan to transfer to 4-year institutions, Bunker Hill will offer a data visualization course as an elective. Treisman noted that while 4-year institutions are using R, 2-year colleges often have fewer resources to allocate to software modernization. He added that a systems approach, as well as a governing authority, is needed to facilitate the transitions between 2- and 4-year institutions.

## CASE STUDIES: OPPORTUNITIES AND CHALLENGES

### Adopting Data 8 at a Two-Year College

*Ava Meredith, Seattle Central College*

Meredith stated that Seattle Central College surveyed 100 of its students and discovered that approximately 80 percent had heard of data science/data analytics, and approximately 60 percent were interested in taking a data science/data analytics course. Based on student interest and industry needs, the mathematics and information technology faculty at Seattle Central identified the need for a data science curriculum and decided to adopt a version of Data 8—a popular introductory data science course at the University of California, Berkeley,[12] that combines inferential thinking, computational thinking, and consideration for social issues in data analysis. The course is designed to be accessible to a broad range of students because it does not require prerequisites beyond high school algebra. Meredith explained that Seattle Central will adopt six goals of the Data 8 course: diversity, equity, pedagogical clarity, scalability, depth, and barrier-free entry. Before implementing any new program, she explained that the curriculum should be aligned to students' backgrounds and needs; administrative constraints should be addressed; and the decision to offer an associate's degree, a certificate, or a single class (for transfer or workforce education) should be evaluated.

Core concepts from Data 8 will be included in the Seattle Central curriculum, course content will be managed with Jupyter Notebooks, and the course language will be Python3, she continued. However, there are a number of areas in which Seattle Central's approach differs. Instead of offering Data 8 in its original integrated format, Seattle Central will offer the program as a set of linked courses: Introduction to Data Analytics and

[12] The course website for Data 8 is http://data8.org/, accessed February 13, 2020.

Introduction to Statistics. Students will register for both courses concurrently, and faculty will coordinate the coursework. Meredith explained that Seattle Central opted to focus the course on "data analytics" instead of "data science" after research indicated that unlike data science jobs, data analytics jobs do not require a master's degree or a Ph.D. Software installation will be part of the curriculum. Instead of working with clean data, Seattle Central students will work with imperfect data sets and real Python libraries and will use GitHub as a code repository and for assignment submissions. Last, the curricula will be offered in flexible modalities (e.g., hybrid and eventually online). Students who choose to pursue a certificate in data analytics will take two additional courses: Python and Database and Data Visualization. Meredith described next steps to include piloting both this new data analytics course and a certificate in data analytics in Spring 2020, developing a plan to advertise and attract a diverse student body, collaborating with the social sciences department to create connector modules and to work with its data sets, partnering with other institutions, and identifying faculty training opportunities.

### DataUp: Increasing the Capacity for Data Science Education

*Renata Rawlings-Goss, South Big Data Regional Innovation Hub*

Rawlings-Goss described the objective of the South Big Data Hub: to connect industry, government, and academia around larger issues for societal and economic development, such as data science education and workforce. In 2016, the South Big Data Hub hosted a workshop—Bridging the Data Divide: Partnering with Diverse Schools to Broaden the Pipeline—in which more than 60 people from 2-year colleges, minority-serving institutions, 4-year liberal arts colleges, government, and industry participated. A consensus report, *Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher-Education Institutions*, emerged in 2018 from this workshop, detailing 13 challenges, 16 visions for the future, 10 tasks, and concrete next steps for data science education (Rawlings-Goss et al., 2018). Two of the challenges highlighted in this report centered on how to implement data science curricula at institutions without the necessary technology stack as well as how to design relevant faculty training.

She explained that DataUp,[13] launched in January 2018, addresses these challenges by providing hands-on training for instructor teams at minority-serving institutions, 2-year colleges, and 4-year liberal arts

[13] The website for DataUp is https://southbigdatahub.org/programs/dataup/, accessed February 13, 2020.

colleges. 2018-2019 DataUp awardees were Spelman College; the University of Puerto Rico, Rio Piedras; the University of the Virgin Islands; Texas A&M, Kingsville; Florida A&M University; Johnson C. Smith University; and Old Dominion University. Faculty (and students, in some cases) teams applied to participate in the year-long program that included a 2-day data science workshop and a train-the-trainers workshop. The train-the-trainers workshop included a partnership with Software Carpentry—upon completion, the teams are certified, supplied with resources, and expected to conduct data science training workshops in their regions. In its effort to democratize data tools, the South Big Data Hub also piloted a project to host a Jupyter Hub. Teams who participated in DataUp were able to use this software during the 2-day workshop to design their curricula.

Rawlings-Goss described possible improvements for the 2020 DataUp Program: (1) Because administrative pressure can constrain community college and tribal college participation, administrators should be included in the process prior to application. (2) Faculty time to participate in external training is limited, so the benefit to the college must be justified, and there must be a clear alignment between the training program and the institution's goals. (3) Decisions about course-level activity do not always reside with instructors, so it is important to identify course- and non-course-related activities that could be counted toward program completion (e.g., boot camps, meet-ups, or student groups). She encouraged roundtable participants to engage with the South Big Data Hub community by subscribing to its monthly newsletter, reading the HubBub blog, joining the South Hub Google group, watching the South Big Data Hub YouTube channel, and following @SouthBigDataHub on Twitter.

### Data Science: A Community College Approach

*Mary Rudis, Pennsylvania State University, Harrisburg*

Rudis described her presentation as a "story of hope for greater inclusiveness and diversity for tomorrow's coders, leaders, data practitioners, researchers, and innovators." She referenced a recent report from the Association for Computing Machinery, *Lighting the Path from Community College to Computing Careers*, which contains case studies about unique approaches to implementing computer science educational pathways across 2- and 4-year institutions in New Jersey, Kentucky, California, Oregon, and Hawaii (ACM, 2018). She also encouraged software developers to connect with 2-year colleges to offer support or host professional development.

Rudis noted that the Community College System of New Hampshire

was awarded a 2013 Innovation Fund Grant to create an undergraduate certificate in data science at Great Bay Community College and Manchester Community College. The objectives of the grant were to support the needs of private-sector companies in greater New England by developing a modern curriculum to create a data-literate workforce; providing a foundational set of coursework that students could apply immediately and transition into a 4-year (or higher) data science/analytics degree; and enhancing existing computer science/computing resources with modern data analytics and visualization tools. First offered in 2015, the Certificate in Practical Data Science[14] removes barriers to entry (i.e., only college-level composition and reading skills are required), offers a more modern approach to mathematics and models courses for liberal arts majors (e.g., the mathematics elective transfers to the University of New Hampshire), is marketed to high school mathematics students, and presents a schedule appropriate for students who rely on financial aid. The 1-year program includes Pre-Calculus, Elements of Data Science, Introduction to Python or Introduction to C++, Probability and Statistics for Scientists, Data Analysis, Visual Language, and a summer capstone project. Rudis clarified that this is not intended to be a "direct-to-workforce" certificate. Mathematics pathways were a barrier for students to complete the certificate program, so bridge courses (e.g., discrete mathematics) had to be developed to enable students from various tracks to move easily into a data science program. Rudis suggested that institutions take the process of implementing a data science program slowly, despite any external pressure that might exist, and carefully contemplate how courses will be taught and what professional development will be needed. Direct-to-workforce programs differ from transfer programs; course redesign will be necessary to meet the needs of the 21st century workforce, she concluded.

### Coordination and Collaboration Between Two- and Four-Year Institutions

*Lior Shamir, Lawrence Technological University and Kansas State University*

Shamir asserted that many 4-year institutions are well funded and suggested that 2- and 4-year institutions collaborate so that resources are allocated, shared, and used more effectively and equitably. Opportunities for collaboration include transfer programs; joint faculty training

[14] The website for the Certificate in Practical Data Science is http://greatbay.edu/courses/certificate-programs/data-practical-data-science, accessed February 13, 2020.

activities; research experiences; shared access to instructors, courses, and retention-driven resources; and integrative data science programs. Owing to the limited number of data science programs at both 2- and 4-year institutions, few data science transfer programs currently exist. However, he suggested that institutions think about the potential for transfer as they design their programs and begin to develop articulation agreements. He also emphasized the need to create a "soft-landing" for transfer students, who are entering a new environment—the need for institutional readiness to offer this support is often underestimated. For example, faculty training is especially important to alleviate stereotypes that 2-year colleges are not as rigorous as 4-year institutions. He also observed that 2-year colleges are often more diverse than 4-year institutions—teaching should be culturally responsive, embedding students' cultures in the learning process.

Shamir noted that, by definition, data science is a research job (i.e., making discoveries from data), yet research at 2-year colleges is underfunded. One approach to ensure that research is included in students' training is to incorporate Research Experiences for Undergraduates (REUs); however, some students will not be selected, others do not view themselves as researchers, and many do not have time for such a commitment. As a result, the REU model may not be the best solution for 2-year colleges. Instead, a course-based research experience (CRE) might be better suited to students' needs, he continued. Community college students can complete the CRE at a partner 4-year institution and transfer the credit toward their associates' degrees. The CRE includes the use of scientific practices, discovery, broadly relevant or important work, collaboration, and iteration. This type of experience serves a larger number of students and does not require any extra-curricular involvement.

## ROUNDTABLE DISCUSSION

Rachel Levy, Mathematical Association of America, wondered how specializations arise and progress as well as how they are categorized, especially in the midst of improving the feedback loop among workforce, industry, and academia. Brandeis Marshall, Spelman College, said that because careers are continuously evolving, industry and academia need to communicate about relevant skill sets and options for job titles. Levy commented on the interesting landscape of 2-year colleges, and Treisman remarked that new mathematics pathways allow students to take courses with a combination of computational, statistical, and mathematical thinking. He suggested that the data science community and mathematical societies capitalize on these reforms. Rudis said that 2-year colleges would welcome more leadership in this area, but she wondered whether this reform of mathematics teaching is happening throughout the educational

system. If not, students could encounter challenges when transferring from a 2- to a 4-year institution. In response to a question from McKeown about the proportion of the 2-year college population that could face this barrier, Shamir noted that 20 percent of 2-year graduates transfer to 4-year institutions.

Rudis said that much of what informs how courses are taught depends on the expertise and interests of the faculty. Marshall added that instructors matter, especially in terms of representation of marginalized groups. McKeown appreciated the strategies shared by Rudis and Shamir to remove barriers to entry and to embrace students' cultures and communities, respectively. She wondered how to attract students to mathematics who initially might not be interested in the discipline. Shamir highlighted Wright State University's approach in which engineers take mathematics that is relevant to their field. He noted that the K-12 system has a different mission than the higher education system, which can create knowledge gaps in certain academic areas that need to be closed. Rudis highlighted the importance of partnering with local K-12 institutions and beginning to target students as early as 5th grade. Students could attend mathematics camps hosted by community colleges; however, it is difficult to secure funding for such activities. An online participant asked whether best practices for engaging students transfer from one 2-year college to another. Shamir replied that each 2-year college is different, so it is important to understand and tailor approaches to each unique system. Treisman said that the demographics of 2-year colleges are changing. For example, 2-year colleges in many states are moving to joint programs with K-12 to remain fiscally viable, and the mathematical societies are considering how to integrate K-12 standards with postsecondary institution objectives. It is thus becoming easier to introduce ideas about data acumen into the K-12 curricula. He reiterated that the demand for students' data knowledge is increasing immensely at the 4-year level, and 2-year colleges will need to develop students' data savvy in a coherent way. Gulbis wondered whether it is possible to create a national standard for technician education. Anderson said that while it is possible, it is impractical. Two-year colleges prepare students for hundreds of different jobs, so while some essentials could be standardized, once they specialize in later years, there is not a one-size-fits-all approach. Shamir agreed with Anderson and said that much can be done through integrated data science programs.

## BREAKOUT GROUP DISCUSSION: COORDINATING WITH OTHER POSTSECONDARY INSTITUTIONS

Horton (moderator) asked about the typical barriers that a data science student encounters when transitioning from a 2- to a 4-year

institution and best practices to ease this transition. Shamir responded that computer programming can be a barrier; however, it is possible to work in data science without mastering computer programming. Horton added that it is important to think about meaningful pathways for students—allowing students to engage with data that are interesting to them can lead to the improvement of algorithmic thinking skills. He also cited considerations for restructuring courses—for example, it is impractical to require computer science before having students work with data, and students cannot be expected to complete an entire series of calculus before being introduced to statistics and modeling. Shamir responded that data science can start with data-driven thinking, and algorithmic thinking can follow later—if algorithmic thinking is a prerequisite, more barriers to entry will be created for students. Jessica Utts, University of California, Irvine, noted that California State University, East Bay, has a data science track for statistics majors[15] that does not require calculus and instead teaches using randomization-based methods, thus eliminating the barrier of calculus for transfer students. Horton pointed out that calculus is not included in the list of mathematical foundations for data acumen in Data Science for Undergraduates: Opportunities and Options. He added that useful levels of mathematical foundations and programming knowledge may differ depending on the type of program and the type of future job. He contrasted engineering programs, where traditional mathematics and computer science backgrounds are required, with business programs, which have fewer requirements in these areas.

Gulbis noted the importance of liberal arts and social sciences to the data science curricula and added that companies such as Apple hire individuals with backgrounds in both technology and liberal arts. Doris Dzameshie, AISCITE Institute, advised getting students involved with GitHub and company hackathons. Horton added that teaching data science across the curricula is important so as to develop capacity in all students. Shamir commented that it is essential to define what counts as a "foundation" of data science. David Bapst, Texas A&M University, agreed and noted that many STEM Ph.D.'s working in industry on data science problems may have little coursework in programming, mathematics, or statistics but have strong skills in using statistics and programming to seek an answer to a particular question. Horton said that many data science projects involve up to 90 percent of time wrangling data; this is equally true for undergraduate students. Bapst added that tools change quickly and unpredictably; data science curricula should be agnostic to the language or tools—which means that the coursework need not be tied

[15] For more information about this data science track, see http://catalog.csueastbay.edu/preview_program.php?catoid=19&poid=7726&returnto=12550, accessed February 13, 2020.

to a specific set of instructors—and updated regularly based on feedback from professionals.

John Hamman, Montgomery College, noted that it is challenging for 2-year colleges to align with multiple 4-year institutions. He noted that delaying programming and calculus coursework could make it difficult for students to transfer to a 4-year institution. Treisman emphasized the need for regional processes to negotiate transfer. Horton noted that in California, articulation agreements between 2- and 4-year institutions are structured with an online database of courses; in other states, they are arranged by state legislation. Hamman said that Montgomery College focused its efforts on aligning with programs at specific institutions, emphasizing that both administrators and faculty should be actively involved in developing these relationships. Treisman stated that data are needed to understand the magnitude of the equity problem that exists for students who transfer from a 2- to a 4-year institution. Shamir pointed out that administrators and faculty at 4-year institutions need to be prepared to work with transfer students from 2-year colleges, which requires training. Treisman agreed and noted that students from 2-year colleges can add much diversity to a 4-year institution. Kathryn Linehan, Montgomery College, described the challenge that arises in transferring course credits from a 2-year college to a 4-year institution. Treisman responded that student success is a necessity to maintain enrollment, and further work on fairer articulation agreements could help to address this equity problem. If a course will not transfer to a 4-year institution, it likely will not survive. John McKenzie, Babson College, noted that there is a Classification of Instructional Programs code for data science.

## BREAKOUT GROUP DISCUSSION: ENHANCING PROFESSIONAL DEVELOPMENT AND ADOPTING EXTERNAL CONTENT

Levy (moderator) asked what resources, programs, and activities exist to support 2-year college faculty in teaching data science. Meredith advised that industry be consulted for guidance on this topic. Grisham described the BioQUEST Curriculum Consortium,[16] which has 33 years of project work and resources as well as week-long workshops for high school and college life science faculty. She also cited QUBES, an NSF-supported project aimed at faculty professional development, which is comprised of a community of mathematics and biology educators. She elaborated that the community typically shares methods and resources

[16] The website for the BioQUEST Curriculum Consortium is https://bioquest.org, accessed February 13, 2020.

to help prepare students to use quantitative approaches to address real, complex biological problems. Levy added that QUBES hosts the mathematics modeling hub, and Grisham noted the importance of building a community, as these groups tend not to interact. Karen Coghlan, National Network of the Libraries of Medicine (NNLM), added that NNLM provides free webinars, classes, and materials for teaching and for research data management. In response to a question from JoEllen Green, Fresno City College, Rudis said that RStudio Cloud eliminates the need to install software and enables collaboration. Eric Simoneau, STATS4STEM.org, noted that RStudio Cloud is currently in alpha mode, which can result in dependability issues.

Rawlings-Goss inquired about institutions that have training programs from industry and wondered how those trainings are received, while Meredith considered the cost of training with certain companies as well as the cost to license technology to an institution. Tygert noted that industry is currently investing heavily in education and training because it has the funding that governments and nonprofit organizations typically do not have. He elaborated that while these efforts are focused on developing students' skills for future careers, there is also a focus on basic science and research and development. Shirley Usry, Hawkes Learning, asked how textbook and web content for students can keep pace with the evolving field of data science. A participant noted that this phenomenon is inevitable in such a dynamic field; it is important to focus on generalizable skills, knowledge, and behavior rather than focusing on specific nuances of a particular piece of software. The participant continued that while specific tools are useful for providing hands-on experience, it can be valuable to expose students to a variety of tools and then key in on underlying shared principles. Scott Tousley, Splunk, noted the similarly rapid pace of innovation in cybersecurity.

## GUIDED REFLECTION AND NEXT STEPS

*Brian Kotz, Montgomery College, and*
*Uri Treisman, University of Texas, Austin*

Kotz concluded that several organizations have expressed their desire to support or partner with 2-year colleges, thus increasing the visibility of 2-year data science education. Two-year colleges serve a wide range of students: the average Montgomery College student is over age 25, all are exclusively commuters, and some take only a course or two. While community colleges can offer nimble customization, funding and resource constraints make it difficult to implement new programs. He cited two key themes from the meeting: the value of high-quality collaboration

and independent customization to meet the unique needs of student populations.

Returning to Horton's framing questions for the meeting, Kotz offered the following commentary:

- *Advocating*—Demonstrate how important data science is and how it impacts all aspects of life.
- *Advertising*—Meet students face-to-face and raise awareness.
- *Managing expectations*—Success means better-informed students with marketable skills.
- *Showing what the students can do*—Share student capstone projects externally, such as with local government.
- *Assessing students and curricula*—Prepare students for larger goals beyond their next job.
- *Evolving*—Maintain flexibility, incentives, and resource sharing
- *Continuing to reflect and discuss*—Remain open to new perspectives and definitions.
- *Offering professional development*—Support educators so that they can support students.

Kotz also elaborated on topics that he would like the data science community to discuss in more depth in the future:

- *Storytelling*—Are students being trained to communicate about data efficiently and effectively?
- *Data analysts and data architects*—Does data science mean "playing in people's backyards" or "building and forming people's backyards"?
- *Distance education*—Do collaborative teams and open resources exist?
- *Local government*—Can students serve their communities through rewarding partnerships?
- *Ethics and privacy*—How are these topics being integrated in 2-year programs?

He hopes to see a platform in the data science education community that enables (1) frequent meetings, (2) systemic structural reforms, (3) an improved understanding of the capabilities of 2-year colleges and their data science students, (4) improved communication within and across institutions and between organizations, (5) a welcoming of others, (6) increased equity for students so that their circumstances do not affect their access and opportunity, and (7) the potential for the 2-year college

model to be embraced for data science. Doing so will empower students to change their lives and those around them, Kotz asserted.

Treisman thanked participants and noted that many of the practices discussed are worthy of attention. New structures will be needed to allow institutions to coordinate the development of their data-rich programs, and state governance and professional societies will need to play a role in helping to level the playing field for 2-year colleges. He reiterated that "transfer" is not just from 2- to 4-year institutions; it also involves students moving from 4- to 2-year institutions and from high school to community college. Administrators need to think about models for back-office functions to enable these transitions. It is also important to think about the role of traditional academic departments in the evolution of courses that develop data acumen. This discussion should be complemented by policy and additional information about the jobs for which people should prepare, he continued. Data science will continue to evolve quickly, and evidence-based modernization of curricula needs to be supported, he concluded.

# References

ACM (Association for Computing Machinery). 2018. *Lighting the Path from Community College to Computing Careers*. https://www.acm.org/binaries/content/assets/education/lighting-the-path-from-community-college-to-computing-careers.pdf.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. "Machine Bias." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baker, M. 2016. "1,500 Scientists Lift the Lid on Reproducibility." News Feature. *Nature*. May 25. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

Barone, L., J. Williams, and D. Micklos. 2017. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computational Biology* 13(10): e1005755. https://doi.org/10.1371/journal.pcbi.1005755.

Bloom, B.S. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. White Plains, NY: Longman.

Buckheit, J.B., and D.L. Donoho. "WaveLab and Reproducible Research." https://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf.

CCC (Computing Community Consortium). 2015. "The Future of Computing Research: Industry–Academic Collaborations." Volume 2. https://cra.org/ccc/wp-content/uploads/sites/2/2016/06/15125-CCC-Industry-Whitepaper-v4-1.pdf.

CCC. 2019. "Evolving Academia/Industry Relations in Computing Research: Interim Report." https://www.cccblog.org/wp-content/uploads/2019/03/Industry-Interim-Report-w-footnotes.pdf.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. 2000. "CRISP-DM 1.0: Step-By-Step Data Mining Guide." https://www.the-modeling-agency.com/crisp-dm.pdf.

CMU (Carnegie Mellon University). 2017. "Carnegie Mellon University Educational Project Agreement." https://www.ri.cmu.edu/wp-content/uploads/2017/01/Educational-Project-Agreement.pdf.

Conway, D. 2010. "The Data Science Venn Diagram." http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Cramer, C., M. Porter, H. Sayama, L. Sheetz, and S. Uzzo. 2015. "Network Literacy: Essential Concepts and Core Ideas." http://tinyurl.com/networkliteracy.

Dieterich, W., C. Mendoz, and T. Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." Northpointe, Inc., Research Department. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography* (S. Halevi and T. Rabin, eds.). TCC 2006. Lecture Notes in Computer Science 3876. Berlin, Heidelberg: Springer.

EDC (Education Development Center, Inc.). 2017. "Tools for Building a Big Data Career Pathway." http://oceansofdata.org/sites/oceansofdata.org/files/Tools%20for%20Building%20a%20Big%20Data%20Career%20Path.pdf.

FTC (Federal Trade Commission). 1998. *Privacy Online: A Report to Congress*. https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf.

Gould, R., R. Peck, J. Hanson, N. Horton, B. Kotz, K. Kubo, J. Malyn-Smith, M. Rudis, B. Thompson, M.D. Ward, and R. Wong. 2018. *The Two-Year College Data Science Summit: A Report on NSF DUE-1735199*. https://www.amstat.org/asa/files/pdfs/2018TYCDS-Final-Report.pdf.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.

Levy, R., R. Laugesen, and F. Santosa. 2018. *Big Jobs Guide.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Meyer, M., A. Cimpian, and S.J. Leslie. 2015. Women are underrepresented in fields where success is believed to require brilliance. *Frontiers in Psychology* 6:235.

Microsoft Azure. 2017. "What Is the Team Data Science Process?" https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview.

NAE (National Academy of Engineering). 2016. *Infusing Ethics into the Development of Engineers*. Washington, DC: The National Academies Press.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018a. *Graduate STEM Education for the 21st Century*. Washington, DC: The National Academies Press.

NASEM. 2018b. *Data Science for Undergraduates: Opportunities and Options.* Washington, DC: The National Academies Press.

NRC (National Research Council). 2011. *A Framework for K-12 Science Education*. Washington, DC: The National Academies Press.

NSF (National Science Foundation). 2015. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.* Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. May.

ODI (Oceans of Data Institute). 2014. *Profile of a Big-Data-Enabled Specialist*. Waltham, MA: Education Development Center, Inc.

ODI. 2016. *Profile of the Data Practitioner*. Waltham, MA: Education Development Center, Inc.

Patil, D.J., H. Mason, and M. Loukides. 2018. *Ethics and Data Science*. Sebastopol, CA: O'Reilly Media.

Patterson, D. 2014. How to build a bad research center. *Communications of the ACM* 57(3):33–36.

Rawlings-Goss, R., L. Cassel, M. Cragin, C. Cramer, A. Dingle, S. Friday-Stroud, A. Herron, et al. 2018. "Keeping Data Science Broad: Negotiating the Digital and Data Divide." https://drive.google.com/file/d/14l_PGq4AxOP9fhJbKqA2necsJZ-gdiKV/view.

Zweben, S., and B. Bizot. 2018. "2017 CRA Taulbee Survey." https://cra.org/crn/2018/05/2017-cra-taulbee-survey-another-year-of-record-undergrad-enrollment-doctoral-degree-production-steady-while-masters-production-rises-again/.

# Appendixes

# A

# Biographical Sketches of Roundtable Members

ERIC KOLACZYK (*Co-Chair*) is a professor of mathematics and statistics at Boston University. He obtained a B.S. degree in mathematics from the University of Chicago and M.S. and Ph.D. degrees in statistics from Stanford University. He has been on the faculty in the Department of Mathematics and Statistics at Boston University since 1998, and was faculty in the Department of Statistics at the University of Chicago before that. He also has been visiting faculty at Harvard University and Université Paris VII. He currently teaches an annual short course at École Nationale de la Statistique et de l'Administration Économique (ENSAE) in Paris. Professor Kolaczyk's main research interests currently revolve around the statistical analysis of network-indexed data, and include both the development of basic methodology and interdisciplinary work with collaborators in bioinformatics, computer science, geography, neuroscience, and sociology. Besides various research articles on these topics, he has also authored two books in this area: *Statistical Analysis of Network Data: Methods and Models* (2009) and, joint with Gabor Csardi, *Statistical Analysis of Network Data in R* (2014). Prior to his working in the area of networks, Professor Kolaczyk spent a decade working on statistical multiscale modeling. He is an elected fellow of the American Statistical Association (ASA), an elected senior member of the Institute for Electrical and Electronics Engineers (IEEE), an elected member of the International Statistical Institute (ISI), and a member of the Institute of Mathematical Statistics (IMS).

KATHLEEN R. M KEOWN (*Co-Chair*) is the Henry and Gertrude Rothschild Professor of Computer Science at Columbia University and she also serves as the director of the Institute for Data Sciences and Engineering. She served as department chair from 1998 to 2003 and as vice dean for research for the School of Engineering and Applied Science for 2 years. Dr. McKeown received a Ph.D. in computer science from the University of Pennsylvania in 1982 and has been at Columbia since then. Her research interests include text summarization, natural language generation, multi-media explanation, question-answering, and multilingual applications. In 1985 she received a National Science Foundation (NSF) Presidential Young Investigator Award, in 1991 she received a NSF Faculty Award for Women, in 1994 she was selected as a Association for the Advancement of Artificial Intelligence (AAAI) fellow, in 2003 she was elected as a fellow of the Association of Computing Machinery (ACM), and in 2012 she was selected as one of the founding fellows of the Association for Computational Linguistics. In 2010, she received the Anita Borg Women of Vision Award in Innovation for her work on text summarization. Dr. McKeown is also quite active nationally. She has served as president, vice president, and secretary-treasurer of the Association of Computational Linguistics. She has also served as a board member of the Computing Research Association and as secretary of the board.

JOHN M. ABOWD is the Edmund Ezra Day Professor of Economics and a professor of statistics and information science at Cornell University and the associate director for research and methodology and chief scientist at the U.S. Census Bureau. At the Census, he leads a directorate of research centers, each devoted to domains of investigation important to the future of social and economic statistics. At Cornell, his primary appointment remains in the Department of Economics in the ILR School. He is also research associate at the National Bureau of Economic Research, research affiliate at the Centre de Recherche en Économie et Statistique (CREST, Paris, France), research fellow at the Institute for Labor Economics (IZA, Bonn, Germany), and research fellow at IAB (Institut für Arbeitsmarkt-und Berufsforschung, Nürnberg, Germany). Dr. Abowd is the director of the Labor Dynamics Institute (LDI) at Cornell. He is the past president (2014-2015) and fellow of the Society of Labor Economists. He is past chair (2013) of the Business and Economic Statistics Section and fellow of the ASA. He is an elected member of the ISI. Dr. Abowd is also a fellow of the Econometric Society. He served as a Distinguished Senior Research Fellow at the Census Bureau (1998-2016). He served on the National Academies' Committee on National Statistics (2010-2016). He currently serves on the American Economic Association's Committee on Economic Statistics (2013-2018). He served as director of the Cornell

Institute for Social and Economic Research (CISER) from 1999 to 2007. Dr. Abowd has taught and done research at Cornell University since 1987, including 7 years on the faculty of the Johnson Graduate School of Management. His current research and many activities of the LDI focus on the creation, dissemination, privacy protection, and use of linked, longitudinal data on employees and employers. In his earlier work at the Census Bureau he provided scientific leadership for the Longitudinal Employer-Household Dynamics (LEHD) Program, which produces research and public-use data integrating censuses, demographic surveys, economic surveys, and administrative data. The LEHD Program's public use data products include the Quarterly Workforce Indicators, the most detailed time series data produced on the demographic characteristics of local American labor markets, and OnTheMap, a user-driven mapping tool for studying work-related commuting patterns. His original and ongoing research on integrated labor market data is done in collaboration with the Institut National de la Statistique et des Études Économiques (INSEE), the French national statistical institute. Dr. Abowd's other research interests include network models for integrated labor market data, statistical methods for confidentiality protection of micro-data, international comparisons of labor market outcomes, executive compensation with a focus on international comparisons, bargaining and other wage-setting institutions, and the econometric tools of labor market analysis. He served on the faculty at Princeton University, the University of Chicago, and the Massachusetts Institute of Technology (MIT) before joining Cornell.

DEB AGARWAL is a senior scientist at Lawrence Berkeley National Laboratory. Her research focuses on scientific tools that enable sharing of scientific experiments, advanced networking infrastructure to support sharing of scientific data, data analysis support infrastructure for eco-science, and cybersecurity infrastructure to secure collaborative environments. Dr. Agarwal is a senior fellow at the Berkeley Institute for Data science and an Inria International Chair, where she co-leads the DALHIS (Data Analysis on Large-scale Heterogeneous Infrastructures for Science) Inria Associated team. Dr. Agarwal also leads teams developing data server infrastructure to significantly enhance data browsing and analysis capabilities and enable eco-science synthesis at the watershed-scale to understand hydrologic and conservation questions and at the global-scale to understand carbon flux. Some of the projects Dr. Agarwal is working on include Genomes to Watersheds SFA2.0, AmeriFlux Management Project, FLUXNET, International Soil Carbon Network, and NGEE Tropics. Dr. Agarwal received her Ph.D. in electrical and computer engineering from University of California, Santa Barbara (UCSB) and a B.S. in mechanical engineering from Purdue University.

RON BRACHMAN is the director of the Jacobs Technion-Cornell Institute and a professor of computer science at Cornell University. He is responsible for the oversight of all institute activities and programs, continuing to develop its vision and strategy and grow it into a completely new role model of innovation for graduate education, while training new leaders who use deep science to change the world. Dr. Brachman received his B.S.E.E. from Princeton University (1971), from which he graduated Summa Cum Laude and Phi Beta Kappa. He received his S.M. (1972) and Ph.D. (1977) degrees in applied mathematics from Harvard University. His research specialization was artificial intelligence, specifically, knowledge representation and reasoning, an area in which he went on to become a world-renowned authority, authoring dozens of highly cited research papers, creating the new field of description logics, and co-authoring a leading textbook. Before joining Cornell Tech, Dr. Brachman had an outstanding career in research and research leadership at world-leading institutions such as Bell Labs, AT&T Labs, DARPA, and Yahoo Labs. At these institutions, he was responsible for recruiting world-class research teams and creating and leading innovative research and academic relationship programs. Dr. Brachman has served as president of AAAI and currently serves on the board of directors of the Computing Research Association. He is a fellow of ACM, IEEE, and AAAI.

JEFFREY BROCK is professor of mathematics and dean of science at Yale University. He focuses on low-dimensional geometry and topology, particularly on spaces with hyperbolic geometry (the most prevalent kind of non-Euclidean geometry). His joint work with R. Canary and Y. Minsky resulted in a solution to the "ending lamination conjecture" of W. Thurston, giving a kind of classification theorem for hyperbolic three-dimensional manifolds that are topologically finite in terms of certain pieces of "mathematical DNA" called laminations. He received his undergraduate degree in mathematics at Yale University and his Ph.D. in mathematics from the University of California, Berkeley, where he studied under Curtis McMullen. After holding postdoctoral positions at Stanford University and the University of Chicago, he joined Brown University as an associate professor. He was awarded the Donald D. Harrington Faculty Fellowship to visit the University of Texas and has had continuous National Science Foundation (NSF) support since receiving his Ph.D. In 2008, he was awarded a John S. Guggenheim Foundation Fellowship.

ALOK CHOUDHARY is the Henry and Isabel Dever Professor of Electrical Engineering and Computer Science and a professor at the Kellogg School of Management at Northwestern University. He is the founding director of the Center for Ultra-scale Computing and Information Security

(CUCIS), which involves several schools, national labs, and universities. Professor Choudhary is a fellow of the IEEE, fellow of the ACM, and a fellow of the American Academy of Advancement of Science (AAAS). Professor Choudhary is the founder, chair, and chief scientist of 4C, which is a big-data science and social media analytics company. 4C is formerly known as VoxSup, Inc., and Professor Choudhary served as its CEO from 2011 to 2013. He was a co-founder and vice president of technology of Accelchip, Inc., in 2000, which was eventually acquired by Xilinx. Professor Choudhary served as the chair of the Electrical Engineering and Computer Science Department from 2007 to 2011. From 1989 to 1996, he was on the faculty of the Electrical and Computer Engineering Department at Syracuse University. He is the recipient of the prestigious NSF Presidential Young Investigator Award in 1993. He has also received an IEEE Engineering Foundation award, an IBM Faculty Development award, and an Intel Research Council award. In 2006, he received the first award for "Excellence in Research, Teaching and Service" from the McCormick School of Engineering. Professor Choudhary received his Ph.D. in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 1989, an M.S. degree from the University of Massachusetts, Amherst, in 1986, and his B.E. (Hons.) degree from the Birla Institute of Technology and Science, Pilani, India, in 1982.

E. THOMAS EWING is an associate dean for graduate studies, research, and diversity in the College of Liberal Arts and Human Sciences and a professor in the Department of History at Virginia Tech. His education included a B.A. from Williams College and a Ph.D. in history from the University of Michigan. He teaches courses in digital humanities and created the course Data in Social Context. His publications include, as author, *Separate Schools: Gender, Policy, and Practice in the Postwar Soviet Union* (2010) and *The Teachers of Stalinism: Policy, Practice, and Power in Soviet Schools in the 1930s* (2002); as editor, *Revolution and Pedagogy: Transnational Perspectives on the Social Foundations of Education* (2005); and as co-editor, with David Hicks, *Education and the Great Depression: Lessons from a Global History* (2006). He has received funding from the National Endowment for the Humanities, the Spencer Foundation, and the National Council for Eurasian and East European Research.

EMILY FOX is an associate professor in the Paul G. Allen School of Computer Science and Engineering and Department of Statistics at the University of Washington, and is the Amazon Professor of Machine Learning. She received an S.B. in 2004 and Ph.D. in 2009 from the Department of Electrical Engineering and Computer Science at MIT. She has been awarded a Presidential Early Career Award for Scientists and Engineers

(2017), Sloan Research Fellowship (2015), Office of Naval Research Young Investigator award (2015), NSF CAREER award (2014), National Defense Science and Engineering Graduate Fellowship, NSF Graduate Research Fellowship, NSF Mathematical Sciences Postdoctoral Research Fellowship, Leonard J. Savage Thesis Award in Applied Methodology (2009), and MIT EECS Jin-Au Kong Outstanding Doctoral Thesis Prize (2009). Her research interests are in large-scale Bayesian dynamic modeling and computations.

JAMES FREW is an associate professor in the Bren School of Environmental Science and Management at UCSB and a principal investigator (PI) in UCSB's Earth Research Institute (ERI). His research interests lie in the emerging field of environmental informatics, a synthesis of computer, information, and Earth sciences. He is interested in information architectures that improve the discoverability, usability, and reliability of distributed environmental information. Trained as a geographer, he has worked in remote sensing, image processing, software architecture, massive distributed data systems, and digital libraries. His current research is focused on geospatial information provenance, science data curation, and applications of array databases, using remote sensing data products as operational test beds. He has affiliate appointments in UCSB's Departments of Geography and Computer Science. He received his Ph.D. in geography from UCSB in 1990. As part of his doctoral research, he developed the Image Processing Workbench, an open source set of software tools for remote sensing image processing. He served as both the manager and the acting director of UCSB's Computer Systems Laboratory (ERI's predecessor), and as the associate director of the Sequoia 2000 Project, a 3-year $14 million multicampus consortium formed to investigate large-scale data management aspects of global change problems. He was a co-PI on the Alexandria Project (part of NSF's Digital Libraries Initiative), where he directed the development of the Alexandria Digital Earth Prototype (ADEPT) testbed system. He also served on the National Academies Committee on Earth Science Data Utilization and as president (2009-2011) of the Federation of Earth Science Information Partners. During the 2005-2006 academic year, he was a visiting professor at the University of Edinburgh's Digital Curation Centre.

CONSTANTINE GATSONIS is the Henry Ledyard Goddard University Professor of Biostatistics at Brown University School of Public Health. He is the founding chair of the Department of Biostatistics and the founding director of the Center for Statistical Sciences at Brown. Dr. Gatsonis is a leading authority on the evaluation of diagnostic and screening tests and evidence synthesis for diagnostic accuracy studies. He has also made

major contributions to the development of methods for medical technology assessment and health services and outcomes research. Dr. Gatsonis is a cofounder of the American College of Radiology Imaging Network (ACRIN) and is now a group statistician for the ECOG-ACRIN collaborative group, a National Cancer Institute–funded collaborative group conducting multicenter studies across the spectrum of cancer care. Dr. Gatsonis chairs the National Academies Committee on Applied and Theoretical Statistics and is a member of the Committee on National Statistics and the Committee to Evaluate the Department of Veterans Affairs Mental Health Services. He has previously served on National Academies committees for a variety of scientific and health-related topics, including forensic science, comparative effectiveness research, immunization safety, aviation security, and modified-risk tobacco products. Dr. Gatsonis was the founding editor-in-chief of *Health Services and Outcomes Research Methodology* and currently serves as associate editor of the *Annals of Applied Statistics*. He was also elected fellow of the ASA and received the 2015 Long-Term Excellence Award from the Health Policy Statistics section of the ASA. He has a B.A. in mathematics from Princeton University, an M.A. in mathematics from Cornell University, and a Ph.D. in mathematical statistics from Cornell University.

LISE GETOOR is a professor in the Computer Science Department at University of California, Santa Cruz, and the director of its D3 Data Science Center. Her research areas include machine learning and reasoning under uncertainty; in addition, she works in data management, visual analytics, and social network analysis. She has more than 200 publications and is a fellow of the AAAI, and an elected board member of the International Machine Learning Society, serves on the board of the Computing Research Association, and has served as *Machine Learning Journal* action editor, associate editor for the *ACM Transactions of Knowledge Discovery from Data*, *JAIR* associate editor, and on the AAAI Council. She is a recipient of an NSF Career award and 11 best paper and best student paper awards. In 2014, she was recognized as one of the top 10 emerging researchers in data mining and data science based on citation and impact according to KD Nuggets. She is on the external advisory board of the San Diego Super Computer Center, and the scientific advisory board for the Max Planck Institute for Software Systems, and has served on the advisory board for companies including Sentient Technologies. She received her Ph.D. from Stanford University in 2001, her M.S. from UC Berkeley, and her B.S. from UCSB, and was a professor at the University of Maryland, College Park, from 2001 to 2013.

MARK L. GREEN is a Distinguished Research Professor in the Department of Mathematics at the University of California, Los Angeles (UCLA).

He received his B.S. from MIT and his M.A. and Ph.D. from Princeton University. After teaching at the UC Berkeley and MIT, he joined UCLA as an assistant professor in 1975. He was a founding co-director and later director of the NSF-funded Institute for Pure and Applied Mathematics. Dr. Green's research has taken him into different areas of mathematics: several complex variables, differential geometry, commutative algebra, Hodge theory, and algebraic geometry. He received an Alfred P. Sloan fellowship, was an invited speaker at the International Congress of Mathematicians in Berlin in 1998 and gave the Chern Medal plenary laudation at the International Conference of Mathematicians in Seoul in 2014, and is a fellow of the American Academy of Arts and Sciences, of the AAAS, and of the American Mathematical Society. Professor Green served as vice chair of the high-profile Board on Mathematical Sciences and Analytics study on The Mathematical Sciences in 2025, and on the International Advisory Panel for the Canadian Long Range Planning Study for Mathematics. He was part of the U.S. Delegation to the General Assembly of the International Mathematical Union in Bangalore in 2010 and chair of the Committee of Visitors for the Division of Mathematical Sciences at NSF in 2013. He has served on the scientific boards of the Institute for Pure and Applied Mathematics, the Centre de Recherches Mathématiques, and the Banff International Research Station, and was a trustee of the American Mathematical Society. He served on the Mathematical Advisory Panel for the exhibition "Man Ray: Human Equations" at the Phillips Collection in Washington, D.C. He serves on the board of governors of the group Transforming Postsecondary Education in Math, the Advisory Committee of the Association for Women in Mathematics, and on the National Academies Board on Mathematical Sciences and Analytics.

ALFRED O. HERO III is the R. Jamison and Betty Williams Professor of Engineering at the University of Michigan. He received a B.S. (summa cum laude) from Boston University (1980) and a Ph.D. from Princeton University (1984), both in electrical engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science, and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. In 2008, he was awarded the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the École Supérieure d'Électricité, Gif-sur-Yvette, France. He is an IEEE fellow, and several of his research articles have received best paper awards. Professor Hero was awarded the University of Michigan Distinguished Faculty Achievement Award (2011). He received the IEEE Signal Processing Society Meritorious Service Award (1998) and the IEEE Third Millenium Medal (2000). He was president of the IEEE Signal Processing Society (2006-2008) and was on the board

of directors of the IEEE (2009-2011), where he served as director of Division IX (Signals and Applications). Dr. Hero's recent research interests have been in detection, classification, pattern analysis, and adaptive sampling for spatiotemporal data. Of particular interest are applications to network security, multimodal sensing and tracking, biomedical imaging, and genomic signal processing.

NICHOLAS J. HORTON is a professor of statistics at Amherst College. He has taught a variety of courses in statistics and related fields and is passionate about improving quantitative and computational literacy for students with a variety of backgrounds as well as engagement and mastery of higher-level concepts and capacities to undertake research. He is the chair of the Committee of Presidents of Statistical Societies and has served on the board of directors of the ASA and as chair of the Statistical Education Section of the ASA. He has published more than 150 papers in statistics and biomedical research and 4 books on statistical computing and data science. He has been the recipient of a number of national teaching awards. As an applied biostatistician, Dr. Horton's work is based squarely within the mathematical sciences, but spans other fields in order to ensure that research is conducted on a sound footing. The real-world research problems that these investigators face often require the use of novel solutions and approaches, because existing methodology is sometimes inadequate. Bridging the gap between theory and practice in interdisciplinary settings is often a challenge, and has been a particular focus of Dr. Horton's work. Dr. Horton earned his Sc.D. in biostatistics from the Harvard School of Public Health.

ERIC HORVITZ is a technical fellow and director at Microsoft Research. His interests include theoretical and practical challenges with developing computing systems that can learn from data and that can perceive, reason, and make decisions. His efforts and collaborations have led to fielded systems in the areas of online services, healthcare, transportation, ecommerce, operating systems, and aerospace. He has received the Feigenbaum Prize and the ACM-AAAI Allen Newell Prize for his contributions to artificial intelligence. He has been elected fellow of AAAI, ACM, and the National Academy of Engineering. He served as president of the AAAI and on advisory boards for the Allen Institute for Artificial Intelligence, NSF, the National Institutes of Health (NIH), the Defense Advanced Research Projects Agency (DARPA), the Computing Community Consortium (CCC), and on the National Academies Computer Science and Telecommunications Board. He is co-chair of the Partnership on AI to Benefit People and Society, recently announced by Amazon, Facebook, Google, IBM, and Microsoft. He did his doctoral work at Stanford University.

BILL HOWE is an associate professor in the Information School, adjunct associate professor in computer science and engineering, and associate director of the University of Washington (UW) eScience Institute. His research interests are in data management, curation, analytics, and visualization in the sciences. Dr. Howe played a leadership role in the Data Science Environment program at UW through a $32.8 million grant awarded jointly to UW, New York University, and UC Berkeley. With support from the MacArthur Foundation and Microsoft, Dr. Howe leads UW's participation in the national MetroLab Network focused on smart cities and data-intensive urban science. He also led the creation of the UW Data Science Master's Degree and serves as its inaugural program director and faculty chair. He has received two Jim Gray Seed Grant awards from Microsoft Research for work on managing environmental data, has had two papers selected for *Very Large Databases* "Best of Conference" issues (2004 and 2010), and co-authored what are currently the most-cited papers from both *Very Large Databases* (2010) and *Special Interest Group on Management of Data* (2012). Dr. Howe serves on the program and organizing committees for a number of conferences in the area of databases and scientific data management, developed a first massive open online course (MOOC) on data science that attracted more than 200,000 students across two offerings, and founded UW's Data Science for Social Good program. He has a Ph.D. in computer science from Portland State University and a bachelor's degree in industrial and systems engineering from Georgia Tech.

CHARLES ISBELL has been a leader in education efforts both at Georgia Tech's College of Computing, where he is senior associate dean for academic affairs, and nationally, where he has co-chaired the Computing Research Association's Subcommittee on Education and currently co-chairs the Coalition to Diversify Computing. At Georgia Tech, Dr. Isbell was one of the co-leaders of Threads. Threads is a successful, comprehensive restructuring of the computing curriculum that provided a cohesive, coordinated set of contexts or threads for teaching and learning computing skills, with a goal of making computing more inclusive, relevant, and exciting for a much broader audience. Dr. Isbell has won numerous teaching awards. Dr. Isbell received his Ph.D. from MIT. His research focuses on artificial intelligence and machine learning.

MARK E. KRZYSKO is deputy director of Enterprise Information for the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics Acquisition Resources and Analysis. In this role, Mr. Krzysko champions and facilitates innovative uses of information technologies to improve and streamline the acquisition process. Prior to this position, he served as the deputy director of Defense Procurement and

Acquisition Policy in the Electronic Business Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. He also served as the division director of Electronic Commerce Solutions for the Naval Air Systems Command, in various senior-level acquisition positions at the Naval Air Systems Command, and as program manager of partnering, the Acquisition Business Process Reengineering Effort, and as acquisition program manager for the Program Executive Office for Tactical Aircraft. Mr. Krzysko began his career in the private retail sector in various executive and managerial positions. He holds a B.S. degree in finance and a master of general administration in financial management from the University of Maryland, University College.

DUNCAN TEMPLE LANG is a professor in the Department of Statistics and director of the Data Sciences Initiative at the University of California, Davis. He joined UC Davis in January 2004. Prior to that, Dr. Temple Lang worked in the Statistics and Data Mining group at Bell Labs, the research arm of Lucent Technologies. He graduated from UC Berkeley with a Ph.D. in statistics, primarily in statistical computing systems. Although trained in statistics, the focus of his research is innovations in information technology and integrating computer science research concepts with the process of scientific and statistical research. An important aspect of his work is to facilitate the integration of software from different communities. Dr. Temple Lang returned to academia from industrial research with the purpose of introducing modern statistical computing to the statistics curriculum.

RACHEL LEVY joined the Mathematical Association of America (MAA) in Fall 2018 as deputy executive director. She came from Harvey Mudd College, where she was a professor of mathematics and associate dean for faculty development. As vice president for education for the Society for Industrial and Applied Mathematics, she advocated internationally for mathematical modeling education connecting K-16 to careers. She has co-authored a partial differential equations (PDE) text, and served as a lead author of the *Guidelines for Assessment and Instruction in Mathematical Modeling Education* and *The BIG Jobs Guide*. She was a recipient of the 2016 Harvey Mudd College Outstanding Faculty Member Award and was recognized by Princeton University Press in "The Best Writing on Mathematics" series. She received a 2016-2017 National Council of Teachers of Mathematics award for linking research and practice and received a 2013 MAA Alder Award for teaching.

BRANDEIS MARSHALL is an associate professor of computer science and chair of the Computer and Information Sciences Department at Spelman College. She received her Ph.D. and M.S. degrees in computer science

from Rensselaer Polytechnic Institute and her B.S. in computer science from the University of Rochester. Her research lies in the areas of information retrieval, data science, data mining, and social media. Dr. Marshall is the PI of an NSF HBCU-UP Targeted Infusion Project titled Data Science eXtension (http://dsxhub.org) that is integrating data science fundamentals into courses at Spelman and Morehouse Colleges. Dr. Marshall is also the director of the Data Analytics and Exploration (da+e) Laboratory, a research and education environment that aims to address real-world data issues, challenges, and solutions funded by federal and industry organizations. The da+e lab activities include timely data acquisition for aviation, BlackTwitter Project, and data/database security curricular development. She is active in mentoring the next generation of STEM professionals, particularly those from underrepresented groups. These engagements include but are not limited to serving on the program committees for the ACM Richard Tapia Diversity in Computing Conference and Grace Hopper Celebration of Women in Computing. From 2013 to 2016, she co-chaired the Broadening Participation in Data Mining Program (BPDM), co-located with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. BPDM fosters mentorship, guidance, and connections of minority and underrepresented groups in data mining, while also enriching technical aptitude and exposure.

CHRIS MENTZEL is director of the Data-Driven Discovery Initiative at the Gordon and Betty Moore Foundation. Previously, he led the grants administration department and also worked as senior network engineer for the foundation. He has also held positions as a systems engineer and integrator at UC Berkeley and at various Internet consulting firms in the Bay Area. An active member of the broader big data and open science communities, Mr. Mentzel serves on a number of advisory boards and program committees and speaks frequently at conferences and workshops on topics related to data-driven research. He received a B.A. in mathematics from the UC Santa Cruz and an M.Sc. in management science and engineering at Stanford University.

NINA MISHRA is a principal scientist at Amazon in Palo Alto, California. Her research interests are in data science, data mining, web search, machine learning, and privacy. Dr. Mishra has more than 16 years of experience leading projects in industry at Microsoft Research and HP Labs and more than 6 years of experience in academia as an associate professor at the University of Virginia and acting faculty at Stanford University. The projects that Dr. Mishra pursues encompass the design and evaluation of new data mining algorithms on real, colossal-sized data sets. She has authored ~50 publications in top venues including Web Search: WWW,

WSDM, SIGIR; Machine Learning: ICML, NIPS, AAAI, COLT; Databases: VLDB, PODS; Cryptography: CRYPTO, EUROCRYPT; and Theory: FOCS and SODA. She has been granted 13 patent applications with a dozen more still in the application stage. Dr. Mishra received her Ph.D. in computer science from the University of Illinois, Urbana-Champaign.

DEBORAH NOLAN is the chair of the Statistics Department and holds the Zaffaroni Family Chair at UC Berkeley. Her research has involved the empirical process, high-dimensional modeling, cross-validation, and most recently technology in education and reproducible research. Professor Nolan has been recognized at UC Berkeley for excellence in teaching and undergraduate student advising and is noted for working with and encouraging all students in their understanding of statistics. She co-directs the Cal Teach and Math for America, Berkeley, programs. Dr. Nolan also organizes Explorations in Statistics Research, a multicampus summer program to encourage undergraduates to pursue graduate studies in statistics. She is an elected fellow of the ASA and a fellow of the IMS. She is co-author of *Stat Labs* with Terry Speed, *Teaching Statistics* with Andrew Gelman, and *Data Science in R* with Duncan Temple Lang. Dr. Nolan received her A.B. from Vassar College and her Ph.D. in statistics from Yale University.

PETER NORVIG is a director of research at Google, Inc. He previously directed Google's core search algorithms group. He is co-author of *Artificial Intelligence: A Modern Approach*, the leading textbook in the field, and co-teacher of an artificial intelligence class that signed up 160,000 students, helping to kick off the current round of massive open online classes. He is a fellow of the AAAI, ACM, California Academy of Science, and American Academy of Arts and Sciences.

ANTONIO ORTEGA received a telecommunications engineering degree from the Universidad Politecnica de Madrid, Madrid, Spain, in 1989 and a Ph.D. in electrical engineering from Columbia University in 1994. In 1994, he joined the Electrical Engineering Department at the University of Southern California (USC), where he is currently a professor and has served as associate chair. He is also a visiting professor at National Institute of Informatics, Tokyo, Japan. He is a fellow of the IEEE and a member of ACM and Asia-Pacific Signal and Information Processing Association (APSIPA). He is currently a member of the board of governors of the IEEE Signal Processing Society, the inaugural editor-in-chief of the APSIPA *Transactions on Signal and Information Processing*, and a senior area editor for *IEEE Transactions on Image Processing*. He has received several paper awards, including most recently the 2016 IEEE Signal Processing

Magazine Award. His recent research work has focused on multiview coding, error tolerant compression, wavelet-based signal analysis, wireless sensor networks, and graph signal processing. Close to 40 Ph.D. students have completed their Ph.D. thesis under his supervision at USC, and his work has led to about 400 publications in international conferences and journals, as well as several patents.

CLAUDIA PERLICH is the chief scientist at Dstillery, leading the machine learning efforts that power Dstillery's digital intelligence for marketers and media companies. With more than 50 published scientific articles, she is a widely acclaimed expert on big data and machine learning applications, and an active speaker at data science and marketing conferences around the world. Dr. Perlich is the past winner of the Advertising Research Foundation's Grand Innovation Award and has been selected for Crain's New York's 40 Under 40 list, Wired Magazine's Smart List, and Fast Company's 100 Most Creative People. Dr. Perlich holds multiple patents in machine learning. She has won many data mining competitions and awards at Knowledge Discovery and Data Mining conferences, and served as the organization's general chair in 2014. Prior to joining Dstillery in 2010, Dr. Perlich worked at IBM's Watson Research Center, focusing on data analytics and machine learning. She holds a Ph.D. in information systems from New York University (where she continues to teach at the Stern School of Business) and an M.A. in computer science from the University of Colorado.

PATRICK O. PERRY is a statistician developing tools and methodology for nontraditional data, especially text and networks. He has worked on text summarization and scaling methods, dynamic network analysis, clustering methods for networks and other data, fitting methods for large-scale hierarchical models, and latent factor methods for high-dimensional data. His work has appeared in the *Journal of the Royal Statistical Society*, the *Annals of Applied Statistics*, and the *Journal of Machine Learning Research*, among other venues. Dr. Perry has developed and released open source implementations of his methods for the R software environment, and he has written a variety of other software packages for data analysis in the C and Haskell programming languages. Currently, Dr. Perry is an assistant professor of information, operations, and management sciences at the New York University Stern School of Business. He teaches courses in introductory statistics, forecasting time series data, and statistics for social data. Dr. Perry received a B.S. in mathematics, an M.S. in electrical engineering, and a Ph.D. in statistics from Stanford University, and he completed a postdoctoral fellowship at Harvard University.

MEHRAN SAHAMI is a professor (teaching) and associate chair for education in the Computer Science Department at Stanford University. He is also the Robert and Ruth Halperin University Fellow in Undergraduate Education. Dr. Sahami's research interests include computer science education, machine learning, and web search. Prior to joining the Stanford faculty in 2007, he was a senior research scientist at Google. He served as co-chair of the ACM/IEEE-CS joint task force on Computer Science Curricula 2013, which created curricular guidelines for college programs in computer science at an international level. Dr. Sahami is also immediate past co-chair of the ACM Education Board and was appointed by California Governor Jerry Brown to serve on the K12 California CS Strategic Implementation Advisory Panel. He co-founded the Symposium on Educational Advances in Artificial Intelligence (EAAI) and the ACM Conference on Learning at Scale (L@S).

VICTORIA STODDEN is an associate professor in the School of Information Sciences at the University of Illinois, Urbana-Champaign. She is a leading figure in the area of reproducibility in computational science, exploring how we can better ensure the reliability and usefulness of scientific results in the face of increasingly sophisticated computational approaches to research. Her work addresses a wide range of topics, including standards of openness for data and code sharing, legal and policy barriers to disseminating reproducible research, robustness in replicated findings, cyberinfrastructure to enable reproducibility, and scientific publishing practices. Dr. Stodden co-chairs the NSF Advisory Committee for CyberInfrastructure and is a member of the NSF Directorate for Computer and Information Science and Engineering (CISE) Advisory Committee. She also serves on the National Academies Committee on Responsible Science: Ensuring the Integrity of the Research Process. Previously an assistant professor of statistics at Columbia University, Dr. Stodden taught courses in data science, reproducible research, and statistical theory and was affiliated with the Institute for Data Sciences and Engineering. She co-edited two books released in 2014—*Privacy, Big Data, and the Public Good: Frameworks for Engagement* and *Implementing Reproducible Research*. Dr. Stodden earned both her Ph.D. in statistics and her law degree from Stanford University. She also holds a master's degree in economics from the University of British Columbia and a bachelor's degree in economics from the University of Ottawa.

URI TREISMAN is executive director of the Charles A. Dana Center for Mathematics and Science Education and a University Distinguished Teaching Professor of Mathematics and Public Affairs at the University of Texas, Austin. He is a Distinguished Senior Fellow at the Education

Commission of the States and chair of the Strong Start to Finish (SSTF) campaign's expert advisory board, a joint initiative of the Bill and Melinda Gates Foundation, the Kresge Foundation, and the Great Lakes Higher Education Guaranty Corporation. SSTF is focused on supporting innovation at scale in American higher education (strongstart.org). Dr. Treisman is active in the leadership of organizations working to improve American mathematics education. He is a founding member of Transforming Postsecondary Education in Mathematics (tpsemath.org) and serves as the representative of the American Mathematical Society to the AAAS (Education, Section Q). He leads the Dana Center Mathematics Pathways (dcmathpathways.org), an initiative that works to modernize entry-level college mathematics course sequences, and the Urban Mathematics Leadership Network, which supports mathematics leadership teams in America's largest urban school districts. He has served on the STEM working group of the President's Council of Advisors on Science and Technology, on the 21st-Century Commission on the Future of Community Colleges of the American Association of Community Colleges, and on the Carnegie/IAS Commission on Mathematics and Science Education. For his work in nurturing minority student high achievement in postsecondary mathematics, he was named a MacArthur Fellow in 1992 and the Harvard Foundation's Scientist of the Year in 2006.

MARK TYGERT is a research scientist for Facebook Artificial Intelligence Research. Prior to this position, he was on the faculty at NYU's Courant Institute, UCLA, and Yale University. He received his B.A. in mathematics from Princeton University and his Ph.D. from Yale University. His research has focused on fast spherical harmonic transforms, randomized algorithms for linear algebra, and complements to chi-square tests. His recent honors include the 2010 William O. Baker Award from the U.S. National Academy of Sciences and the 2012 DARPA Young Faculty Award. His current research interests are in machine learning, statistics, and computational science and engineering, particularly numerical analysis.

JEFFREY D. ULLMAN is the S.W. Ascherman Professor of Engineering (Emeritus) at Stanford University, where he taught in the Department of Computer Science from 1979 to 2002. He worked at Bell Laboratories from 1966 to 1969 and taught at Princeton University (from which he also received his Ph.D. in 1966) between 1969 and 1979. He is the author or coauthor of widely read textbooks in compilers, databases, and algorithms, as well as the book in automata on which his automata course is based and the book on data mining on which his Mining of Massive Datasets course is based. He is a member of the National Academy of

Engineering and the American Academy of Arts and Sciences and winner of the ACM Karl V. Karlstrom Education award, the IEEE Von Neumann Medal, and the Knuth Prize.

JESSICA UTTS is a professor of statistics at UC Irvine, where she served as chair from 2010 to 2016. During her tenure as chair, the Statistics Department created an undergraduate major in data science. She was also the 2016 president of the ASA, and during her presidential year, the ASA board discussed and endorsed the report *Curriculum Guidelines for Undergraduate Programs in Data Science*. She received her B.A. in math and psychology at SUNY Binghamton and her M.A. and Ph.D. in statistics at Penn State University. She is the author of *Seeing Through Statistics* and the co-author with Robert Heckard of *Mind on Statistics* and *Statistical Ideas and Methods*. Dr. Utts has been active in the statistics education community at the high school and college levels. She served as a member and then chaired the Advanced Placement Statistics Development Committee for 6 years and currently serves as the chief reader for AP Statistics. She was a member of the ASA task force that produced the Guidelines for Assessment and Instruction in Statistics Education (GAISE) recommendations for elementary statistics courses. She was a founding member of the Statistics Department at UC Davis and spent many years on the faculty before moving to UC Irvine in 2008. She is the recipient of the Academic Senate Distinguished Teaching Award and the Magnar Ronning Award for Teaching Excellence, both at UC Davis. She is also a fellow of the ASA, the IMS, and the AAAS. Beyond statistics education, Dr. Utts's major contributions have been in applying statistics to a variety of disciplines, most notably to parapsychology, the laboratory study of psychic phenomena. She has appeared on numerous television shows, including *Larry King Live*, *ABC Nightline*, *CNN Morning News*, and *20/20*, and appears in a documentary included on the DVD with the movie *Suspect Zero*.

JANE YE has been a program officer at NIH for 15 years and currently manages a portfolio of advanced research projects in biomedical informatics with a special focus on bioinformatics and translational informatics. She has graduate degrees from Dartmouth College and Cornell University. Most of the research projects in her portfolio involve the application of computer and information sciences to improve the access, storage, retrieval, management, dissemination, and use of biomedical information. Before joining the NIH, she worked in the private sector as a senior bioinformatics scientist working on genomic data and gene discovery. She made contributions to the sequencing and publishing of the human genome.

# B

# Meeting Participants

## MEETING #1: THE FOUNDATIONS OF DATA SCIENCE FROM STATISTICS, COMPUTER SCIENCE, MATHEMATICS, AND ENGINEERING

*Roundtable members present*: Eric Kolaczyk, Boston University (via teleconference); John Abowd, U.S. Census Bureau; Ron Brachman, Cornell University; Alok Choudhary, Northwestern University; Michelle Dunn, National Institutes of Health; James Frew, University of California, Santa Barbara; Constantine Gatsonis, Brown University; Alfred Hero, University of Michigan; Nicholas Horton, 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; Antonio Ortega, University of Southern California; Patrick Perry, New York University; Victoria Stodden, University of Illinois, Urbana-Champaign; Mark Tygert, Facebook Artificial Intelligence Research (via teleconference); Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

*Guests present*: Stephanie August, National Science Foundation; Peter Bruce, Statistics.com; Isabel Cárdenas-Navia, Business-Higher Education Forum; David Culler, University of California, Berkeley; Tom Ewing, Virginia Tech (via teleconference); Lou Gross, University of Tennessee, Knoxville; Laura Haas, IBM; Linda Hyman, National Science Foundation; Sara Kiesler, National Science Foundation; Brian Kotz, Montgomery

College; Duncan Temple Lang, University of California, Davis; Natassja Linzau, U.S. Department of Commerce; Christopher Malone, Winona State University; Andrew McCallum, University of Massachusetts, Amherst; Richard McCullough, Harvard University; Steven Miller, IBM; Rebecca Nugent, Carnegie Mellon University; David Rabinowitz; Lee Rainie, Pew Research Center; Stephanie Rodriguez, National Science Foundation; Rob Rutenbar, University of Illinois, Urbana-Champaign; Daniel Siu, National Science Foundation; William Velez, University of Arizona; Elena Zheleva, National Science Foundation; and Andrew Zieffler, University of Minnesota, Twin Cities.

## MEETING #2: EXAMINING THE INTERSECTION OF DOMAIN EXPERTISE AND DATA SCIENCE

*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.

## MEETING #3: DATA SCIENCE EDUCATION IN THE WORKPLACE

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; John Abowd, U.S. Census Bureau; Ron Brachman (via webcast), Cornell University; Brian Caffo, Johns Hopkins University; Alok Choudhary, Northwestern University; Alfred Hero, University of Michigan; Nicholas Horton, Amherst College; Charles Isbell, Georgia Institute of Technology; Mark

Krzysko, U.S. Department of Defense; Chris Mentzel, Gordon and Betty Moore Foundation; Deborah Nolan, University of California, Berkeley; Antonio Ortega, University of Southern California; Claudia Perlich, Dstillery and New York University; Patrick Perry, New York University; Victoria Stodden, University of Illinois, Urbana-Champaign; Mark Tygert (via webcast), Facebook Artificial Intelligence Research; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

*Guests present*: Quincy Brown, American Association for the Advancement of Science; Ashley Campana, Booz Allen Hamilton; Catherine Cramer, New York Hall of Science; David Culler, University of California, Berkeley; Ying Ding, Indiana University; Renee Dopplick, Association for Computing Machinery; E. Thomas Ewing, Virginia Tech; William Finzer, Concord Consortium; Louis Gross, University of Tennessee, Knoxville; Laura Haas, IBM; Ryan Seth Jones, Middle Tennessee State University; Brian Kotz, Montgomery College; Ashley Lanier, Booz Allen Hamilton; David Levermore, University of Maryland, College Park; Andrew McCallum, University of Massachusetts, Amherst; Mary Moynihan, Cape Cod Community College; Rebecca Nugent, Carnegie Mellon University; Emily Plachy, IBM; Ron Prevost, U.S. Census Bureau; Lee Rainie, Pew Research Center; Hridesh Rajan, Iowa State University; Patrick Riley, Google; Rob Rutenbar, University of Illinois, Urbana-Champaign; Jordan Sellers, Howard University; Kristin Tolle, Microsoft; Ken Wilkins, National Institutes of Health; Drew Zachary, U.S. Department of Commerce; and Andrew Zieffler, University of Minnesota.

## MEETING #4: ALTERNATIVE MECHANISMS FOR DATA SCIENCE EDUCATION

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Ron Brachman, Cornell Tech; Alok Choudhary, Northwestern University; James Frew, University of California, Santa Barbara; Alfred Hero, University of Michigan; Nicholas Horton (via webcast), Amherst College; Mark Krzysko, U.S. Department of Defense; Chris Mentzel (via webcast), Gordon and Betty Moore Foundation; Nina Mishra, Amazon; Deborah Nolan, University of California, Berkeley; Antonio Ortega, University of Southern California; Victoria Stodden, University of Illinois, Urbana-Champaign; Mark Tygert (via webcast), Facebook Artificial Intelligence Research; and Jeffrey Ullman, Stanford University.

*Guests present*: Katy Börner, Indiana University; Andrew Bray, Reed College; Catherine Cramer, New York Hall of Science; Abhijith Gopakumar,

Northwestern University; Dan Nicolae, University of Chicago; Michelle Paulsen, Northwestern University; Hridesh Rajan, Iowa State University; Karl Schmitt, Valparaiso University; Stephen Uzzo, New York Hall of Science; Nicholas Wagner, Northwestern University; and David Ziganto, Metis.

## MEETING #5: INTEGRATING ETHICAL AND PRIVACY CONCERNS INTO DATA SCIENCE EDUCATION

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; John Abowd, U.S. Census Bureau; Constantine Gatsonis, Brown University; Alfred Hero, University of Michigan; Nicholas Horton, Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Patrick Perry, New York University; Victoria Stodden, University of Illinois, Urbana-Champaign; and Jeffrey Ullman, Stanford University.

*Guests present*: Aubra Anthony, USAID; Anna Arnando; Solon Barocas, Cornell University; David Culler, University of California, Berkeley; Michael Fountane, alumnus of Harvard University; Simson Garfinkel, U.S. Census Bureau; Anna Lauren Hoffmann, University of Washington; Brian Kotz, Montgomery College; Aaron Margolis, Federal Bureau of Investigation; Moses Namara, Clemson University; Kyle Novak, American Association for the Advancement of Science; Cathy O'Neil, mathbabe.org; Hridesh Rajan, Iowa State University; Aaron Roth, University of Pennsylvania; Mary Rudis, Bates College; Dhruv Sharma, Federal Deposit Insurance Corporation; Rochelle Tractenberg, Georgetown University; and Jevin West, University of Washington.

## MEETING #6: IMPROVING REPRODUCIBILITY BY TEACHING DATA SCIENCE AS A SCIENTIFIC PROCESS

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Alok Choudhary, Northwestern University; James Frew, University of California, Santa Barbara; Mark Green, University of California, Los Angeles; Alfred Hero, University of Michigan; Nicholas Horton (via webcast), Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Duncan Temple Lang, University of California, Davis; Brandeis Marshall, Spelman College; 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; 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*: Alison Gammie, National Institute of General Medical Sciences; Timothy Gardner, Riffyn; Charlotte Mazel-Cabasse, University of California, Berkeley; Mary Beth McLendon, Accel.AI; Laura Montoya, Accel.AI; Fernando Perez, University of California, Berkeley; Josh Quan, University of California, Berkeley; Anthony Suen, University of California, Berkeley; Tracy Teal, The Carpentries; Tom Treynor, Treynor Consulting; Eric Van Dusen, University of California, Berkeley; Adam Wolisz, Berlin University of Technology; and Buck Woody, Microsoft Research and AI.

## MEETING #7: PROGRAMS AND APPROACHES FOR DATA SCIENCE EDUCATION AT THE PH.D. LEVEL

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Jeffrey Brock, Brown University; Alok Choudhary, Northwestern University; James Frew, University of California, Santa Barbara; Lise Getoor, University of California, Santa Cruz; Alfred Hero III, University of Michigan; Nicholas Horton, Amherst College; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Duncan Temple Lang (via webcast), University of California, Davis; Uri Treisman, University of Texas, Austin; Mark Tygert, Facebook Artificial Intelligence Research; and Jeffrey Ullman, Stanford University.

*Guests present*: Bilikis Akindel, Duke Health Technology Solutions Analytics Center of Excellence; Magdalena Balazinska, University of Washington; Rocco Blais, National Intelligence University; Karim Boughida, University of Rhode Island; Philip Bourne, University of Virginia; Arlyn Burgess, University of Virginia; Patricia Cifuentes, Pan American Health Organization; Vasant Dhar, New York University; Jason Dunavant, Selbst; Lisa Federer, National Institutes of Health; Narryn Fisher, Technology Consulting; Michael Garris, National Institute of Standards and Technology; Robert Hershey, Robert L. Hershey, P.E.; Aditya Johri, George Mason University; Erick Jones, National Science Foundation; Yasir Khalid, Georgetown University; Elizabeth Linton, Mount Sinai Hospital; Bert Little, University of Louisville; Raghu

Machiraj, The Ohio State University; Burt Monroe, Pennsylvania State University; Ademola Okerinde, Kansas State University; George Onyullo, Department of Energy and Environment; Chuba Oraka, University of the Potomac; Abani Patra, University at Buffalo; Bryan Pijanowski, Purdue University; Hridesh Rajan, Iowa State University; Benjamin Ryan, Gallup, Inc.; Atma Sahu, Coppin State University; Steve Sawyer, Syracuse University; Eugenia Schenecker, George Washington University; Devavrat Shah, Massachusetts Institute of Technology; Sharad Sharma, Bowie State University; Surja Sharma, University of Maryland; Susan Singer, Rollins College; Martin Skarzynski, Foundation for Advanced Education in the Sciences; Andrew Sostek, Katanya; Daniel Spielman, Yale University; Ethan Steininger, Virtue Theory Inc; Michael Turner, University of North Carolina, Charlotte; Dila Udum, Bahcesehir University; Lucianne Walkowicz, Library of Congress; Bob Weinberg, Massachusetts Institute of Technology; Olivia Wong; Brian Wright, George Washington University; Bin Wu, Liberty Language Service, Inc; and Bing Xue, Consultant.

## MEETING #8: CHALLENGES AND OPPORTUNITIES TO BETTER ENGAGE WOMEN AND MINORITIES IN DATA SCIENCE EDUCATION

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Ron Brachman, Cornell Tech; Alok Choudhary, Northwestern University; Emily Fox, University of Washington; Lise Getoor, University of California, Santa Cruz; Nicholas Horton (via webcast), Amherst College; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Rachel Levy, Mathematical Association of America; Brandeis Marshall, Spelman College; Antonio Ortega, University of Southern California; Victoria Stodden, University of Illinois, Urbana-Champaign; Uri Treisman, University of Texas, Austin; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

*Guests present*: Sylvia Ankrah; Ginger Baxter, Emory University Goizueta Business School; Kamau Bobb (via webcast), Georgia Institute of Technology; Cheryl Brown, University of North Carolina, Charlotte; Cynthia Bryant, Georgia Institute of Technology; Gregory Chambers, Independent Contractor; Stephanie Espy, author; Blake Fleisher, Anidata; Lisa Gervin, LG Technical Devices; Ayanna Howard, Georgia Institute of Technology; Tremayne Jackson, Tesla; Vivian Lyon, Plaza Dynamics; Nancy Murray, Emory University; Rebecca Nugent, Carnegie Mellon University; Ami Radunskaya, Pomona College; Renata Rawlings-Goss, South Big Data Hub; Alicia Richhart, Georgia Institute of Technology;

Dalene Stangl, Carnegie Mellon University; Kendra Strickland, Georgia Institute of Technology; Lydia Tapia, University of New Mexico; and Charlie Wright, Georgia Institute of Technology.

## MEETING #9: MOTIVATING DATA SCIENCE EDUCATION THROUGH SOCIAL GOOD

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Lise Getoor, University of California, Santa Cruz; Alfred Hero III, University of Michigan; Nicholas Horton (via webcast), Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Rachel Levy, Mathematical Association of America; Nina Mishra, Amazon; Michael Pearson, Mathematical Association of America; Mehran Sahami, Stanford University; Uri Treisman, University of Texas, Austin; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

*Guests present*: Tensae Andargachew, New Jersey Institute of Technology; James Angelo, Leidos; Ted Avraham, The Jewish Student Satellite Initiative; Rahul Bhargava, Massachusetts Institute of Technology Media Lab; Cheri Borsky; Peter Bull, DrivenData; Michael P. Cohen, American Institutes for Research; Richard Esposito, Bureau of Labor Statistics; Adam Fagen, BioQUEST Curriculum Consortium; Matt Gee, University of Chicago and BrightHive; Lauri Goldkind, Fordham University; Louis Gross, University of Tennessee; Doug Hague, University of North Carolina, Charlotte; Robert Hershey, Robert L. Hershey, P.E.; James Hodson, AI for Good Foundation; Kristin Jenkins, BioQUEST; Benjamin Kallen, Lewis-Burke Associates; Brian Kotz, Montgomery College; Kathryn Kozak, American Mathematical Association of Two-Year Colleges; Zenobia Liendo, George Washington University/University of California, Berkeley; Elizabeth McDaniel, Institute for Defense Analyses; John McNutt, University of Delaware; Sharon McPherson, National Science Foundation; Peter Mecca, George Mason High School; D.J. Patil (via webcast), Devoted Health; Desmond Patton (via webcast), Columbia University; John Rowan, Coding Dojo; Frank Sanacory, The State University of New York College at Old Westbury; Yla Tausczik, University of Maryland iSchool; Jeremy Wojdak, Radford University/QUBES; Brian Wright, George Washington University; Li-chiung Yang; Maryam Zaringhalam, National Library of Medicine.

## MEETING #10: IMPROVING COORDINATION BETWEEN ACADEMIA AND INDUSTRY

*Roundtable members present*: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Jeffery Brock, Yale University; Emily Fox, University of Washington; James Frew, University of California, Santa Barbara; Lise Getoor, University of California, Santa Cruz; Mark Green, University of California, Los Angeles; Alfred Hero III, University of Michigan; Nicholas Horton, Amherst College; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Duncan Temple Lang, University of California, Davis; Rachel Levy, Mathematical Association of America; 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; Mehran Sahami, Stanford University; Victoria Stodden, University of Illinois, Urbana-Champaign; and Mark Tygert, Facebook Artificial Intelligence Research.

*Guests present*: Chaitan Baru, University of California, San Diego; Catherine Brooks, University of Arizona; Adam Causgrove, Carnegie Mellon University; Michael Franklin, University of Chicago; Hunter Glanz, California Polytechnic State University; Gary King, Harvard University; Daniel Marcu, Amazon; Rebecca Nugent, Carnegie Mellon University; Mary Ellen Sullivan, Mass Mutual; Mike Willardson, Facebook; and Ben Zorn, Microsoft.

## MEETING #11: DATA SCIENCE EDUCATION AT TWO-YEAR COLLEGES

This meeting was virtual, with more than 400 participants.