Roundtable

February 2017

HIGHLIGHTS

Roundtable on Data Science Post-Secondary Education

Meeting #1 - December 14, 2016

The Roundtable on Data Science Post-Secondary 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. Watch meeting videos or download presentations at **nas.edu/data-science-education-roundtable-1**.

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 just at the undergraduate level, but also at the K-12 levels and in community colleges. 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 distinguish-

ing 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:

- Basic foundations (e.g., understanding data through algorithms, machine learning, curation, visualization/modeling, and computational systems),
- Advanced foundations (e.g., understanding large-scale data through high-performance computing and advanced machine learning), and
- Practicum (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 at 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 science (e.g., computer science, engineering, statistics) or domain science (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, analy-

sis, 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 if 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, 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 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 the National Institutes of Health (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 if 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 if NIH targets any of its programs to undergraduates so 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 if 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 impor-

tance 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 lifecycle 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 if "business acumen" is different 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's: domain experts teach data-intensive courses focused on problems, not methods. Christopher Malone, Winona State University, asked if 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 is 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.

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

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

ROUNDTABLE MEMBERS ABSENT: Kathleen McKeown, Columbia University; Brian Caffo, Johns Hopkins University; Ronald Coifman, Yale University; Emily Fox, University of Washington; Johannes Gehrke, Microsoft; Deborah Nolan, University of California, Berkeley; Alex Pentland, Massachusetts Institute of Technology; and Claudia Perlich, Dstillery

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; 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; Duncan Temple-Lang, University of California, Davis; William Velez, University of Arizona; Elena Zheleva, National Science Foundation; and Andrew Zieffler, University of Minnesota, Twin Cities

STAFF PRESENT: Lida Beninson, Program Officer, BOSE; Linda Casola, Research Associate, BMSA; Janel Dear, Senior Program Assistant, CSTB; Jon Eisenberg, Board Director, CSTB; Michelle Schwalbe, Board Director, BMSA; Scott Weidman, Deputy Executive Director, Division on Engineering and Physical Sciences; Ben Wender, Program Officer, BMSA

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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