Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report

DETAILS

68 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-46502-1 | DOI: 10.17226/24886

CONTRIBUTORS

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective; Computer Science and Telecommunications Board; Board on Mathematical Sciences and Analytics; Committee on Applied and Theoretical Statistics; Division on Engineering and Physical Sciences; Board on Science Education; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

ENVISIONING THE DATA SCIENCE DISCIPLINE: THE UNDERGRADUATE PERSPECTIVE

Interim Report

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective

Computer Science and Telecommunications Board
Board on Mathematical Sciences and Analytics
Committee on Applied and Theoretical Statistics
Division on Engineering and Physical Sciences

Board on Science Education
Division of Behavioral and Social Sciences and Education

A Consensus Study Report of

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. 1626983 from the National Science Foundation. 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-XXXXX-X
International Standard Book Number-10: 0-309-XXXXX-X
Digital Object Identifier: https://doi.org/10.17226/24886

Additional copies of this publication are available for sale 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. 2017. *Envisioning The Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24886.

PREPUBLICATION COPY – SUBJECT TO FURTHER EDITORIAL CORRECTION

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. C. D. Mote, Jr., 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**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON ENVISIONING THE DATA SCIENCE DISCIPLINE: THE UNDERGRADUATE PERSPECTIVE

LAURA HAAS, NAE,[1] University of Massachusetts Amherst, *Co-Chair*
ALFRED O. HERO III, University of Michigan, *Co-Chair*
ANI ADHIKARI, University of California, Berkeley
DAVID CULLER, NAE, University of California, Berkeley
DAVID DONOHO, NAS,[2] Stanford University
E. THOMAS EWING, Virginia Tech
LOUIS J. GROSS, The University of Tennessee, Knoxville
NICHOLAS HORTON, Amherst College
JULIA LANE, New York University
ANDREW MCCALLUM, University of Massachusetts Amherst
RICHARD MCCULLOUGH, Harvard University
REBECCA NUGENT, Carnegie Mellon University
LEE RAINIE, Pew Research Center
ROB RUTENBAR, University of Illinois, Urbana-Champaign
KRISTIN TOLLE, Microsoft Research
TALITHIA WILLIAMS, Harvey Mudd College
ANDREW ZIEFFLER, University of Minnesota

Staff

JON EISENBERG, Board Director, Computer Science and Telecommunications Board (CSTB), *Study Co-Director*
MICHELLE K. SCHWALBE, Board Director, Board on Mathematical Sciences and Analytics (BMSA), *Study Co-Director*
LINDA CASOLA, Associate Program Officer and Editor, BMSA
JANEL DEAR, Senior Program Assistant, CSTB (until May 2017)
RENEE HAWKINS, Financial Manager, CSTB
AMY STEPHENS, Program Officer, Board on Science Education
BEN WENDER, Program Officer, BMSA

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

COMMITTEE ON APPLIED AND THEORETICAL STATISTICS

ALFRED O. HERO III, University of Michigan, *Chair*
ALICIA CARRIQUIRY, Iowa State University
MICHAEL J. DANIELS, University of Texas, Austin
KATHERINE BENNETT ENSOR, Rice University
AMY HERRING, University of North Carolina, Chapel Hill
NICHOLAS HORTON, Amherst College
DAVID MADIGAN, Columbia University
JOSÉ M.F. MOURA, NAE, Carnegie Mellon University
NANCY REID, University of Toronto
CYNTHIA RUDIN, Duke University
AARTI SINGH, Carnegie Mellon University

Staff

BEN WENDER, Director
LINDA CASOLA, Associate Program Officer and Editor
BETH DOLAN, Financial Manager
RODNEY N. HOWARD, Administrative Assistant

Acknowledgments

This Consensus Study Report 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 published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We thank the following individuals for their review of this report:

Richard De Veaux, Williams College,
W. Eric L. Grimson, Massachusetts Institute of Technology,
C.K. Gunsalus, University of Illinois, Urbana-Champaign,
Iain M. Johnstone, NAS,[1] Stanford University,
Brian Kotz, Montgomery College,
Peter Norvig, Google, Inc., and
Renata Rawlings-Goss, Georgia Institute of Technology.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by Raghu Ramakrishnan, Microsoft Corporation. He responsible for making certain that an independent examination of this report 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 authoring committee and the National Academies.

The committee would like to thank Andy Burnett, Knowinnovation, for facilitating the committee's May workshop as well as the following staff members from the National Science Foundation for their input, assistance, and support of this study: Stephanie August, Chaitan Baru, Eva Campo, Erwin Gianchandani, Nandini Kannan, Sara Kiesler, Gabriel Perez-Giz, Earnestine Psalmonds-Easter, and Elena Zheleva.

The committee would also like to thank the following individuals for providing input to this study:

John Abowd, U.S. Census Bureau,
Deb Agarwal, Lawrence Berkeley National Laboratory,
Jon Ahlquist, Florida State University,
Barbara Alvin, Eastern Washington University,
Barbara Anthony, Southwestern University,
David Austin, North Carolina State University,
Maria Aysa-Lastra, Winthrop University,
Tom Barr, American Mathematical Society,
Laura Bartley, University of Oklahoma,
Nina Bijedic, University "Džemal Bijedić"of Mostar,

[1] Member, National Academy of Sciences.

Sally Blake, Flagler College,
Roselie Bright, U.S. Food and Drug Administration,
Quincy Brown, American Association for the Advancement of Science,
Andy Burnett, Knowinnovation,
Dave Campbell, Simon Fraser University,
Robert Campbell, Brown University,
Robert Carver, Stonehill College,
Amy Chang, American Society for Microbiology,
Lei Cheng, Olivet Nazarene University,
Hongmei Chi, Florida A&M University,
Alok Choudhary, Northwestern University,
William Coberly, University of Tulsa,
Peyton Cook, University of Tulsa,
Bill Corey, University of Virginia,
Catherine Cramer, New York Hall of Science,
James Curry, University of Colorado, Boulder,
Nicole Dalzell, Duke University,
Juliana DeCastro, Núcleo de Planejamento Estratégico de Transporte e Turismo,
Sam Donovan, University of Pittsburgh,
Renee Dopplick, Association for Computing Machinery,
Maureen Doyle, Northern Kentucky University,
Ruth Duerr, Ronin Institute,
Arturo Duran, IVA Ventures,
Stephen Edwards, ACM Administrative Centre,
Sandra Ellis, Texas A&M University, Corpus Christi,
Paula Faulkner, North Carolina Agricultural and Technical State University,
Raya Feldman, University of California, Santa Barbara,
Dilberto Ferraren, Visayas State University,
William Finzer, Concord Consortium,
Julia Fisher, Coker College,
Roger French, Case Western Reserve University,
Kimberly Gardner, Kennesaw State University,
Sommer Gentry, U. S. Naval Academy,
Tara Ghazi, University of California, Berkeley,
Richard Gill, Brigham Young University,
Shana Gillette, U.S. Agency for International Development,
Juan Godoy, Universidad Nacional de Córdoba, Consejo Nacional de Investigaciones Cientificas y Técnicas,
Greg Goins, North Carolina A&T State University,
Robert Gould, University of California, Los Angeles,
C. K. Gunsalus, University of Illinois, Urbana-Champaign,
Mirsad Hadzikadic, University of North Carolina, Charlotte,
Jim Hammerman, TERC,
Michael Harris, Bunker Hill Community College,
John Hathaway, Brigham Young University-Idaho,
Kristin Hunter-Thomson, Rutgers University,
Ambra Hyskaj, National Association of Public Health Albania,
Charles Isbell, Georgia Institute of Technology,
Mark Jack, Florida A&M University,
Bob Jecklin, University of Wisconsin, La Crosse,
Xia Jing, Ohio University,

Jeremiah Johnson, University of New Hampshire,
John Johnstone, University of Alabama, Birmingham,
Ryan Jones, Middle Tennessee State University,
Sungkyu Jung, University of Pittsburgh,
Michael Kangas, Doane University,
Roxanne Kapikian, GlaxoSmithKline,
Danny Kaplan, Macalester College,
Casey Kennington, Boise State University,
Deepak Khatry, MedImmune,
Brian Kotz, Montgomery College,
Vldimir Krylov, Crimean Engineering and Pedagogical University,
Kristin Kuter, Saint Mary's College, Notre Dame, Indiana,
Jay Labov, National Academies of Sciences, Engineering, and Medicine,
Paula Lackie, Carleton College,
Sharon Lane-Getaz, St. Olaf College,
Duncan Temple Lang, University of California, Davis,
Jeff Leek, Johns Hopkins University,
Matthew Liberatore, University of Toledo,
Haralambos Marmanis, Marmanis Group,
Pat Marsteller, Emory University,
Abhinav Maurya, Carnegie Mellon University,
Victoria McGovern, Burroughs Wellcome Fund,
Daniel Angel Ferreira Mena, DAF-Engineering,
Chris Mentzel, Gordon and Betty Moore Foundation,
Antoni Miklewski, Polish Academy of Sciences,
Ashlea Milburn, University of Arkansas,
Alex Montilla, U.S. Environmental Protection Agency,
Sheri Morgan, Mental Health Association of Franklin and Fulton Counties,
Richard Morris, MGI-RamCo,
Mary Kehoe Moynihan, Cape Cod Community College,
Bhramar Mukherjee, University of Michigan,
Sherman Mumford, University of North Carolina, Charlotte,
Ivo Neitzel, Faculdade de Ciências e Tecnologia de Birigui,
Richard Nelesen, University of California, San Diego,
Joseph Nelson, George Washington University,
Claudia Neuhauser, University of Minnesota,
Deborah Nolan, University of California, Berkeley,
Kofi Nyamekye, Integrated Activity-Based Simulation Research, Inc.,
Monika Oli, University of Florida,
Fred Oswald, Rice University,
Dennis Pearl, Pennsylvania State University,
Joan Peckham, University of Rhode Island,
Vikas Pejaver, University of Washington,
Gabriel Perez-Giz, National Science Foundation,
Patrick Perry, New York University,
Steve Pierson, American Statistical Association,
Hridesh Rajan, Iowa State University,
Louise Raphael, Howard University,
Renata Rawlings-Goss, Georgia Institute of Technology,
Peggy Rejto, Normandale Community College,
Loren Rhodes, Juniata College,

Patrick Riley, Google, Inc.,
Martina Rosenberg, University of New Mexico,
Kim Roth, Juniata College,
Bill Roweton, Chadron State College,
Andee Rubin, TERC,
Maya Sapiurka, Society for Neuroscience,
Karl Schmitt, Valparaiso University,
Kala Seal, Loyola Marymount University,
Arun Sharma, Wagner College,
Lauren Showalter, National Academies of Science, Engineering, and Medicine,
Christine Smith, University of New Mexico,
S. Srinivasan, Texas Southern University,
Anil Srivastava, Open Health Systems Laboratory,
Natalya St. Clair, Concord Consortium,
Victoria Stodden, University of Illinois, Urbana-Champaign,
Martin Storksdieck, Oregon State University,
George Strawn, National Academies of Science, Engineering, and Medicine,
Ralph Stuart, Keene State College,
Kalum Udagepola, Scientific Research Development Institute of Technology Australia,
Mel van Drunen, HAS University of Applied Sciences,
William Yslas Velez, University of Arizona,
Ron Wasserstein, American Statistical Association,
Cheryl Welsch, State University of New York, Sullivan,
Mary Whelan, Arizona State University,
Nekesha Williams, Louisiana State University,
Emerald Wilson, Prince George's Community College,
Brian Wingenroth, National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland,
William Winter, State University of New York College of Environmental Science and Forestry,
Mary Wright, Brown University, and
Paul Zachos, Association for the Cooperative Advancement of Science and Education.

Contents

Summary

The need to manage, analyze, and extract knowledge from data is pervasive across industry, government, and academia. Scientists, engineers, and executives routinely encounter enormous volumes of data, and new techniques and tools are emerging to create knowledge out of these data, some of them capable of working with real-time streams of data. The nation's ability to make use of these data depends on the availability of an educated workforce with necessary expertise. With these new capabilities have come novel ethical challenges regarding the effectiveness and appropriateness of broad applications of data analyses.

The future of data science education is impacted by the continuing evolution of computing technology, analytical approaches, and tools; the corresponding demand from employers for new knowledge and skills; and new models for delivering education. Educational institutions may need to revise the content of their curricula and embrace multiple models of educational delivery (e.g., online, self-paced, team teaching both in and out of the classroom) to better appeal to a broad population of students and to better prepare students to enter the workforce.

The field of data science has emerged to address the proliferation of data and the need to manage and understand it. Data science is a hybrid of multiple disciplines and skill sets, draws on diverse fields (including computer science, statistics, and mathematics), encompasses topics in ethics and privacy, and depends on specifics of the domains to which it is applied. Fueled by the explosion of data, jobs that involve data science have proliferated and an array of data science programs at the undergraduate and graduate levels have been established. Nevertheless, data science is still in its infancy, which suggests the importance of envisioning what the field might look like in the future and what key steps can be taken now to move data science education in that direction. Future data science programs will need to incorporate a variety of skills. Strong analytic skills are needed to work with large, complex data sets. Oral and written communication skills are also necessary to engage with diverse audiences about real-world problems, to work in teams, and to participate in effective problem solving for both technical and ethical dilemmas encountered in uses of data science.

The committee has also identified several apparent hallmarks of effective data science education. Using real data will expose students to the messiness they will confront when solving real-world problems. Selecting applications with broad impact will make instruction more compelling, helping to attract and retain students. Teaching commonly used current methods will prepare them for the workplace, as will exposure to working in teams. Critical curricular topics include mathematical foundations, computational thinking, statistical thinking, principles of effective data management, techniques for data description and curation, data modeling approaches, effective communication skills, reproducibility challenges and current best practices, exposure to ethical dilemmas and problem-solving skills, and a range of domain-specific topics. Maturity in these and other areas results in what this committee defines as "data acumen," which enables data scientists to make good judgments and decisions with data. The process of starting students down the path toward data acumen is a chief objective of data science education.

Because data science is inherently concerned with understanding and addressing real-world problems and challenges, the new and expanding field of data science may appeal to a wider variety of students. The field of data science encompasses multiple disciplines and varied skill sets and has the potential to attract students with diverse academic backgrounds and interests. The opportunity to build in

broad participation, diversity, and inclusion from the onset is a notable advantage, as compared to other related fields of study. The field presents new opportunities to attract and engage underrepresented student populations. Such potential opportunities can be realized through innovative cross-disciplinary pedagogical approaches led by highly trained and flexible faculty. To further increase participation, 4-year institutions can partner with 2-year institutions that have flexible programs as a way to offer more entry points into data science for advanced high school students, current members of the workforce, and future transfer students. Assessment and evaluation are especially valuable when building these new programs in part because it encourages consideration of how well curricular objectives are being met. The very tools of experimental design and analysis common in the field of data science will likely prove valuable in evaluating the success of data science programs.

This interim report from the Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective begins to address the statement of task presented in Box S.1. Specifically, this report lays out some of the information and comments that the committee has gathered and heard during the first half of its study, offers perspectives on the current state of data science education, and poses some questions that may shape the way data science education evolves in the future. This National Academies of Sciences, Engineering, and Medicine study, sponsored by the National Science Foundation, will conclude in early 2018 with a final report that lays out a vision for future data science education. What follows in this interim report are initial observations and findings concerning the state of data science education, a discussion of forward-looking opportunities, and key questions on which the committee seeks broad input.

BOX S.1
Statement of Task

A National Academies of Sciences, Engineering, and Medicine study will set forth a vision for the emerging discipline of data science at the undergraduate level. It will emphasize core underlying principles, intellectual content, and pedagogical issues specific to data science, including core concepts that distinguish it from neighboring disciplines. It will not consider the practicalities of creating materials, courses, or programs. It will develop this vision considering applications of and careers in data science. It will focus on the undergraduate level, addressing related issues at the middle and high school as well as community colleges as appropriate, and will draw on experiences in creating master's-level programs. It will also consider opportunities created by the emergence of a new STEM [science, technology, engineering, and mathematics] field to engage underrepresented student populations and consider ways to reduce the "leakage" seen in existing STEM pathways. Information gathering will center around two workshops, the first likely focused on principles and intellectual content, and the second likely focused on pedagogy and implications for middle and high schools and community colleges. To get material on the record quickly and spark community feedback, an interim report will be issued following the first workshop. The interim report will not include recommendations, but may include findings or conclusions if the evidence warrants. A final report will be issued following both workshops and committee deliberations setting forth a vision for undergraduate education in data science.

The preliminary findings from the committee are described throughout this interim report and recapped in Chapter 5 where they are accompanied by a set of open questions developed by the committee about the future of data science education (also summarized in Box S.2).

Public input is sought on the following topics:

- Additional content for this study, including but not limited to case studies from institutions providing data science education, innovative ways to bring researchers together, best practices for program evaluation, and ideas for future topical webinars;
- A proposed Data Science Oath, outlined at the beginning of Chapter 5; and
- Questions about how data science can evolve in the future.

BOX S.2
Questions for Public Input

The committee has identified the following themes, which are discussed throughout this interim report, and is soliciting input on the open questions. Please visit the following webpage to provide input: http://www.nas.edu/EnvisioningDS.

Building data acumen in a data science curriculum

- *Which key components should be included in data science curriculum, both now and in the future?*
- *How could these components be prioritized or best conveyed for differing types of data science programs?*
- *How can opportunities to enhance data acumen (i.e., the ability to make good judgments and decisions with data) be integrated into data science educational programs?*
- *How can data acumen be measured or evaluated?*

Real-world applications

- *How can partnerships between industry and educational programs be encouraged?*
- *Could a focus on real problems serve as a means to attracting more diverse students?*
- *How can students gain access to real-world data sets?*

Ethics

- *How can ethical considerations be best incorporated throughout the data science curriculum?*
- *How can students be taught to apply ethical decision making throughout the problem-solving process?*

Oral and written communication skills and teamwork

- *How can communication and teamwork be fostered in data science programs?*
- *What type of multidisciplinary teams serve as effective models for the real world? Will these groupings be different in the future?*

Pedagogical approaches and inter-departmental collaboration

- *What are known good practices for fostering collaboration between departments and existing programs?*
- *What new directions and opportunities exist for new curricular initiatives?*
- *What pedagogical approaches are particularly relevant to data science, both now and in the future?*

Faculty and curriculum development

- *What types of training would be beneficial to faculty?*
- *How could incentives be restructured to encourage more faculty development in data science?*

Organizational structure and institutional infrastructure

- *What are current infrastructure obstacles and how can they be rethought going forward?*
- *How could organizational structures be modified and/or incentives added to encourage data science collaboration and innovation?*

Educational approaches

- *How can data science programs build in flexibility and adaptability so they can be most responsive to changes in the field?*
- *How can flexibility encourage more diverse students?*

Diversity, inclusion, and broad participation

- *How can broad participation, diversity, and inclusion be ingrained in data science programs?*
- *What strategies to recruit and retain diverse students can data science programs deploy, and what examples can inform these efforts?*

Partnerships between institutions

- *How can partnerships between 2- and 4-year institutions be facilitated?*
- *How do the skills and concepts taught at a 2-year institution vary based on students' goals?*
- *What aspects of data science education are appropriate and feasible to develop at 2-year institutions?*

Assessment and evaluation

- *What evaluation and assessment objectives are currently being used in data science programs, and how will these differ in the future?*
- *What best practices in evaluation and assessment can inform data science programs?*
- *What data are available to evaluate the effectiveness of different data science approaches?*
- *What standard evaluation approaches should be adopted?*

Please visit the following webpage to provide input: www.nas.edu/EnvisioningDS.

1

Introduction

ENVISIONING DATA SCIENCE FROM AN UNDERGRADUATE PERSPECTIVE

The emergence of a novel science of data highlights the need for new principles for data collection, storage, integration, and analysis. These new scientific principles are leading to new tools that uniquely respond to the challenges of big data. However, the main concepts, skills, and ethics powering this emerging discipline of data science still need to be identified. A new generation of tool developers and tool users will require the ability to make good judgments and decisions with data and use tools responsibly and effectively (referred to as "data acumen" throughout this report). Some of these developers and users will draw from computing, mathematics, and statistics fields, but many will come from other fields and application domains. Educators and administrators are beginning to reimagine data science course content, delivery, and enrollment at the undergraduate level to best prepare students to operate in this new paradigm.

New and greater volumes of information compound long-standing challenges of data analysis—and raise new ones. The ability to measure, understand, and react to data can affect scientific discovery, social interaction, political tradition, economic practice, public health, and many other areas. Some data science applications are low risk, such as recommender systems that suggest purchases within an online shopping platform or select advertisements for website visitors. Although provider sales may be affected if undesirable products are recommended and users may be dissatisfied with their purchases, the overall impact of poor recommender systems to individuals and society is low. Of greater impact, census data are used to redraw political boundaries, allocate funds, and inform other critical public policy decisions. While new volumes and types of information can make analyses more accurate than past methods that relied on sparse surveys with lower than desired response rates, people can be negatively affected if the interpretation of the data does not account for all relevant factors. A program that a family depends on may not have sufficient funding, or a policy might be enacted that has unintended consequences for large segments of the population if weak data analysis is used. Thus, it is important that data are collected and analyzed appropriately, especially as new demands are placed on data collection and evaluation and as new technologies emerge. It is equally important that there are clear principles guiding the use of data for human good. Further, the complexity of the analyses and the increasing dependency on data across all the fields of human endeavor drive demand for "smarter" tools and best practices for data science that will minimize mistakes in interpretation.

Academic institutions and industry recognize these shifts and are rapidly embracing the idea that there is an emerging discipline of data science that is unique yet builds on knowledge from existing disciplines (NRC, 2014). Traditional statistical methods are well established and clearly understood but often do not scale to handle the vast volumes of data that must be analyzed for today's data science. Computing is unparalleled in its capacity to handle vast volumes or fast-flowing streams of information, but often without statistical and inferential guarantees, which can result in unreliable results and biased or unfair interpretations of the data (Jordan, 2013). Domain areas (e.g., business, medicine, natural science, social sciences, or engineering) are developing and adapting techniques to solve specific research

questions, which can be more effective than using general methods. However, these approaches may suffer from insufficient mathematical or statistical rigor or lack computational scalability.

Although the definition of data science is evolving, it centers around the notion of multidisciplinary and interdisciplinary approaches to extracting knowledge or insights from data for use in a broad range of applications. It is the field of science that relies on processes and systems (mathematical, computational, and social) to derive information or insights from data. It is about synthesizing the most relevant parts of the foundational disciplines to solve particular classes of problems or applications while also creating novel techniques to address the "cracks" between those disciplines where no approaches may yet exist. This flexibility is an essential component of data science and is equally important in data science education.

Data scientists have the potential to help address critical real-world challenges. The following list includes just a few illustrative examples:

- *Enabling more accurate diagnosis of melanomas through better analysis of images.* Within the clinical field, deep learning techniques[1] have been applied to detect melanoma, the most deadly form of skin cancer. These methods improve the analysis of tissue images, enabling a more accurate diagnosis than traditional techniques (Codella et al., 2017).
- *Enhancing business decisions.* Business analytics can assist entrepreneurs and company executives in making timely decisions based on market trends. This can be coupled with online social media information to respond directly to consumer demands or create a more personalized advertising experience (Chen et al., 2012).
- *Helping aid organizations respond quickly.* Data science and analytics are used to assist aid organizations to respond more quickly in times of need, such as when the Swedish Migration Board used data science to make predictions about and determine national implications for immigration trends (Pratt, 2016).
- *Developing "smart cities."* Cities around the world such as London, Rio de Janeiro, and New York City collect real-time data from a variety of sources, such as public transportation smart cards and traffic cameras, environmental sensors for parameters such as temperature and humidity, and social media interactions regarding local issues. The data can then be processed, analyzed, and utilized to improve city efficiency and cost-effectiveness as well as resident well-being (Kitchin, 2014).

However, there are also many instances of high-impact and high-profile data science research resulting in flawed or inaccurate findings, as well as ethical and legal quandaries. The following list includes a few illustrative examples:

- *Inaccurate predictions of flu trends.* In 2013, Google Flu Trends over-predicted true influenza-related doctors' visits as determined by the Centers for Disease Control and Prevention. This was primarily the result of overreliance on outdated models (Butler, 2013).
- *Use of personally identifiable data.* The abundance of data available on individuals from companies and social media can present ethical dilemmas to researchers in terms of privacy, scalability of results, and subject participation agreement. For instance, a 2013 study linked numerous Twitter users to sensitive information from their financial institutions, which contributed to discussions of if and when researchers should be required to obtain written consent when using nominally publicly accessible information (Danyllo et al., 2013).
- *Predictive policing.* There is much debate over the use and appropriateness of predictive policing—the use of data science by law enforcement to predict crime before it occurs. There

[1] Deep learning is a powerful class of a machine learning methods that explore data representations using supervised, semi-supervised, or unsupervised learning.

is no consensus yet on the effectiveness of this methodology, and civil liberties groups argue that the data used to develop (i.e., train) the models are inherently biased (Hvistendahl, 2016).

Data science is currently being practiced in hundreds of organizations within industry, academia, and government, often by self-taught practitioners. There are indications of strong demand in a variety of domains for graduates with data science skills. A recent study by IBM found more than 2.3 million data science and analytics job listings in 2015, and both job openings and job demand are projected to grow significantly by 2020. Three-fifths of the data science and analytics jobs are in the finance and insurance, professional services, and information technology sectors, but the manufacturing, health care, and retail sectors also are hiring significant numbers of data scientists (Miller and Hughes, 2017; Columbus, 2017). The IBM study also shows that it takes significant time to find and hire staff with the right mix of skills and experience. Since many employers are themselves new to the use of data science, they may not be able to provide training and therefore may seek individuals who have appropriate classwork and hands-on experience.

Current data science courses, programs, and degrees are highly variable in part because emerging educational approaches start from different institutional contexts, aim to reach students in different communities, address different challenges, and achieve different goals. This variation makes it challenging to lay out a single vision for data science education in the future that would apply to all institutions of higher learning, but it also allows data science to be customized and to reach broader populations than other similar fields have done in the past. Data science educational programs are emerging within many existing fields such as statistics, computer science, business, and social sciences. These field-specific approaches bring about unique distinctions in how data science is taught, which skills are emphasized, which students are served, and which career paths graduates pursue. Other data science educational programs are taking a cross-disciplinary approach—for example, integrating statistics and computer science concepts into the undergraduate data science degree program. (Several example programs are discussed in Chapter 3 of this report.)

This report highlights some of the important common threads that can be woven throughout much of data science education. Chapter 2 discusses the foundational, translational, ethical, and professional skills that help students acquire data science skills and knowledge. Chapter 3 explores the role of innovative curriculum development and provides some considerations for institutions. Chapter 4 examines ways to ensure broad participation in data science, including recruitment and retention strategies, institutional partnerships, K-12 objectives, public outreach, and the role of evaluation and assessment. Chapter 5 provides some reflections on the committee's findings throughout the report and proposes questions for public input. That chapter also lays out a draft Hippocratic Oath for data science, which is also open for public input.

The themes described in this report underlay data science education, but they are not necessarily novel challenges or even unique to data science (as demonstrated by the historical case study in Box 1.1). The lessons learned from other disciplines can help pave the way to ensuring the success of data science education.

This report aims to lay out some key questions to help advance conversations around data science education and to provide institutions and participants with a clearer picture of paths forward.

STUDY ORIGIN AND APPROACH

This National Academies' interim report begins to address the statement of task for the committee, presented in Box 1.2.

BOX 1.1
Data Science Education: An Expedition

Envisioning a vibrant and robust approach to undergraduate data science education is an important mission that may require considerable deliberation and effort to achieve. Such missions of exploration feature challenges and opportunities.

A useful historical analogy can be made with the Lewis and Clark expedition (1804–1806), which laid the foundation for great discoveries about a newly expanded nation through systematic data collection and analysis. In his instructions to Meriwether Lewis, U.S. President Thomas Jefferson described the data-gathering equipment made accessible to the party: "Instruments for ascertaining by celestial observations the geography of the country thro' which you will pass, have been already provided."

Data collection and aspects of what we might think of as reproducible data were important to Jefferson (1803):

> Your observations are to be taken with great pains & accuracy to be entered distinctly, & intelligibly for others as well as yourself, to comprehend all the elements necessary, with the aid of the usual tables to fix the latitude & longitude of the places at which they were taken, & are to be rendered to the war office, for the purpose of having the calculations made concurrently by proper persons within the U.S. Several copies of these as well as of your other notes, should be made at leisure times, & put into the care of the most trustworthy of your attendants, to guard by multiplying them against the accidental losses to which they will be exposed. A further guard would be that one of these copies be written on the paper of the birch, as less liable to injury from damp than common paper.

Aspects of non-technical (political considerations) were also acknowledged in the instructions (Jefferson, 1803):

> Your mission has been communicated to the Ministers here from France, Spain, & Great Britain, and through them to their governments: and such assurances given them as to its objects as we trust will satisfy them. The country of Louisiana having been ceded by Spain to France, the passport you have from the Minister of France, the representative of the present sovereign of the country, will be a protection with all its subjects: and that from the Minister of England will entitle you to the friendly aid of any traders of that allegiance with whom you may happen to meet.

The Lewis and Clark expedition made important contributions to science (most notably geography, ecology, and biology) while spurring growth and migration into the American frontier. Many discoveries were serendipitous, with unanticipated positive and negative outcomes. A similar ambitious exploration of how to formulate effective undergraduate data science education will not be easy but has the potential to have a significant impact on higher education and society.

COMMITTEE ACTIVITIES TO DATE

At the time of this writing, the study committee held two meetings and one webinar to collect information, engage a diverse community, and deliberate. The open session presentations given during these meetings are listed in Appendix B. Additional information-gathering activities are planned for the remainder of 2017, after which the committee will release a final report.

During the first meeting of the committee on December 12-13, 2016, participants discussed possible future directions based on progress with current data science programs; societal implications of the evolving field of data science; ways to expand diversity and inclusion in data science among students, employees, and even topic areas; and perspectives on envisioning the future of data science specifically for undergraduates.

BOX 1.2
Statement of Task

A National Academies of Sciences, Engineering, and Medicine study will set forth a vision for the emerging discipline of data science at the undergraduate level. It will emphasize core underlying principles, intellectual content, and pedagogical issues specific to data science, including core concepts that distinguish it from neighboring disciplines. It will not consider the practicalities of creating materials, courses, or programs. It will develop this vision considering applications of and careers in data science. It will focus on the undergraduate level, addressing related issues at the middle and high school as well as community colleges as appropriate, and will draw on experiences in creating master's-level programs. It will also consider opportunities created by the emergence of a new STEM [science, technology, engineering, and mathematics] field to engage underrepresented student populations and consider ways to reduce the "leakage" seen in existing STEM pathways. Information gathering will center around two workshops, the first likely focused on principles and intellectual content, and the second likely focused on pedagogy and implications for middle and high schools and community colleges. To get material on the record quickly and spark community feedback, an interim report will be issued following the first workshop. The interim report will not include recommendations, but may include findings or conclusions if the evidence warrants. A final report will be issued following both workshops and committee deliberations setting forth a vision for undergraduate education in data science.

The committee hosted a public webinar on April 25, 2017, and gathered public input and outside perspectives on the topics the committee should discuss throughout the remainder of the study.

The committee also held a workshop on May 2-3, 2017, where participants discussed (1) educational models to build relevant foundational, translational, and professional skills and knowledge for data scientists in various roles; (2) use of high-impact educational practices in the future delivery of data science education; and (3) strategies for broad participation in data science education that revolve around formal modes of evaluation and assessment. Other topics emerged as well, including the role of teacher education, the need to consult research on learning styles and teaching methods, the relationship between data science and popular culture, better methods for assessment of student and program success, and the ways in which students, institutions, and programs might change over the next 10 years and how these changes may affect plans for the future of data science education.

2

Acquiring Data Science Skills and Knowledge

Expanding data science training has the potential to transform scientific discovery, other academic research, many professions, and the broader society. With such an onset of new technologies, modes of thought, and means of communication, questions arise among industries, government agencies, and educational institutions: *What skills are needed to be successful in the workplace and in society? Is data science a fundamental skill that all students should have some exposure to? How can data literacy be improved? In what skills, methods, and technologies should future data scientists be trained, given the wide variety of potential applications?* Understanding the complexities of these questions is a first step in imagining the discipline for a diverse set of participants.

This chapter discusses some of the foundational, translational, ethical, and professional skills that make it possible for students to be effective data scientists.

FOUNDATIONAL SKILLS

What are the key ideas and principles to be included in the data science curriculum? One way to determine this is to consider the typical work cycle in which data scientists engage. For example, this cycle is often initiated with a domain-specific question that then leads to data collection. The data are typically curated, described, and modeled. Models are tested and deployed, then the results are put to use and communicated to stakeholders. There are potentially several phases of analysis within this work cycle that lead to other questions and deeper understandings. Additionally, data scientists will need to draw on computational, statistical, and mathematical knowledge, as well as domain-specific knowledge, to inform the analytic choices and interpretation made throughout the workflow.

A simplified description of a data scientist's work cycle helps illuminate the essential components of a data science curriculum. For example, De Veaux et al. (2017) describe the following six areas of focus for data science curricula that map well to this workflow: (1) data description and curation, (2) mathematical foundations, (3) computational thinking, (4) statistical thinking, (5) data modeling, and (6) communication, reproducibility, and ethics. Given different definitions of computational thinking, it may not be evident that De Veaux's six areas are intended to encompass not only basic computing concepts (such as abstraction and indirection) but also the array of computing skills required to manage data. Therefore, the committee suggests adding computing skills as a seventh important area.

Specific topics within these seven focus areas might include the following: software engineering, linear algebra, optimization, algorithms and data structures, information technology, basic statistics, uncertainty quantification, and tools for fitting models to data. Human–computer interaction research may also play a role in the foundational data science curriculum as it examines the design and use of computer technology focused on the interfaces between people and computers, including the range of ways in which humans are integrated with computational processes, data collection, dissemination, and analysis.

The path from research question to analysis has changed with the advent of data science. In the past, data were typically collected with a specific purpose and via a particular design to answer a priori research questions. These data, in turn, informed the statistical analyses. More analyses are utilizing

extant data and repurposing them to answer new questions and explore new hypotheses (Groves, 2011). As more of these existing data sets become accessible (e.g., via application programming interfaces), one core question is how to extract knowledge and insight from data that were collected for an entirely different purpose and, subsequently, with little forethought to the design necessary for answering those questions.

While it might be possible to take a piecemeal approach to the data science curriculum in which courses are selected from existing departments, and although these courses might look reasonable as a curricular whole, in reality, such a curriculum will almost certainly lack educational and cross-disciplinary cohesion unless there is some coordination across the departments and courses.

In developing an undergraduate data science curriculum, it is important to evaluate how particular topics and skill sets will both fulfill program requirements and prepare students to address data challenges they will face in their careers. Training in describing and documenting models, analyses, and value propositions effectively will benefit students preparing for a wide variety of data science careers. Although a deep theoretical foundation is less necessary for students pursuing data science positions after earning an undergraduate degree, an emphasis on developing sophisticated techniques and complex modeling skills is still valuable to help solve real-world data science problems. With more data and more complicated models available, interpretability of models is important in both data science education and practice, as is fairness in algorithms and computation-driven decisions. Many academic researchers, businesses, and government agencies that are hiring new employees value graduates with expertise in survey methodology and designed data, elements of statistical learning that serve as the framework for machine learning, Bayesian data analysis, and implementations of reproducible research.

Developing and applying these skill sets requires "data acumen"—the ability to make good judgments and decisions with data. This trait is increasingly important, especially given the large volume of data typically present in real-world problems, the relative ease of (mis)applying tools, and the vast ethical implications inherent in many data science analyses. With large volumes of data, it can be difficult to understand at first glance what is needed, what is possible, and what limitations exist. Still, questions remain as to how to most effectively build data acumen in students. Data acumen can be developed over time through research experience, industry partnerships, courses in creative data analysis, domain-specific data science courses or experiences, and extensions of capstone-like experiences throughout the curriculum. It can be enhanced through exposure to current key components of data science, including mathematical foundations, computational thinking, statistical thinking, data management, data description and curation, data modeling, ethical problem solving, communication and reproducibility, and domain-specific considerations. Similar to the concept of "mathematical maturity," which typically denotes a mixture of mathematical insights and experiences that mathematicians develop and strengthen with time, data acumen is not a final state to be reached but rather a skill that data scientists develop and refine over time.

Finding 2.1: A critical component of data science education is to guide students to develop data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts related to developing data acumen include the following:

- Mathematical foundations,
- Computational thinking,
- Statistical thinking,
- Data management,
- Data description and curation,
- Data modeling,
- Ethical problem solving,
- Communication and reproducibility, and
- Domain-specific considerations.

The necessary level of exposure to each area will vary based on the overall objectives and duration of the data science program as well as the goals for the students.

As is true with curriculum design in other academic settings, it is important to build a curriculum in which students can recognize when they do not know something so that they learn which questions to ask. It is also important for students to learn to question what they may already perceive as fact, to understand the risks associated with using data, and to recognize that data is a product of the specific context in which it is generated, collected, analyzed, and interpreted. Incorporating insights from the humanities and social sciences into data science curricula enables students to consider how behaviors, interactions, and attitudes shape data in an informed and grounded way.

Lessons learned and effective practices from other domains can help shape how data science is taught—including co-curricular activities (such as mentorship programs), individualized advising, supplemental opportunities for students to learn fundamentals (such as summer bridge programs), and introductory courses designed to appeal to a wide student audience. High-impact educational practices, such as those put forth by the Association of American Colleges and Universities, describe teaching and learning practices that have been shown to be beneficial for postsecondary students from many backgrounds. These practices take many different forms, depending on learner characteristics and on institutional priorities and contexts, but could be useful in ensuring that data science education is effective. Box 2.1 highlights the educational practices put forth by the Association of American Colleges and Universities and provides some examples of how they are currently being applied to data science education.

BOX 2.1
High-Impact Educational Practices[a]

A Brief Overview

The following teaching and learning practices have been widely tested and have been shown to be beneficial for college students from many backgrounds. These practices take many different forms, depending on learner characteristics and on institutional priorities and contexts.

On many campuses, assessment of student involvement in active learning practices such as these has made it possible to assess the practices' contribution to students' cumulative learning. However, on almost all campuses, utilization of active learning practices is unsystematic, to the detriment of student learning. Presented below are brief descriptions of high-impact practices that educational research suggests increase rates of student retention and student engagement.

First-Year Seminars and Experiences

Many schools now build into the curriculum first-year seminars or other programs that bring small groups of students together with faculty or staff on a regular basis. The highest-quality first-year experiences place a strong emphasis on critical inquiry, frequent writing, information literacy, collaborative learning, and other skills that develop students' intellectual and practical competencies. First-year seminars can also involve students with cutting-edge questions in scholarship and with faculty members' own research. An example of this first-year experience in data science has been explored by a professor at Duke University, who suggests the creation of a gateway course that would allow students to learn fundamental data science skills in a project-oriented curriculum (Cetinkaya-Rundel, 2017).

Common Intellectual Experiences

The older idea of a "core" curriculum has evolved into a variety of modern forms, such as a set of required common courses or a vertically organized general education program that includes advanced integrative studies and/or required participation in a learning community. These programs often combine broad themes—for example, technology and society, and global interdependence—with a variety of curricular and co-curricular

options for students. An example model of a course that could be utilized in a core curriculum is Data8, a foundational data science course at the University of California, Berkeley that analyzes the technical and social implications of data science in a hands-on manner. The course is intended for students without a background in statistics or computer science, allowing for students to explore these topics without pre-requisites (Culler, 2016).

Learning Communities

The key goals for learning communities are to encourage integration of learning across courses and to involve students with "big questions" that matter beyond the classroom. Students take two or more linked courses as a group and work closely with one another and with their professors. Many learning communities explore a common topic and/or common readings through the lenses of different disciplines. Some deliberately link "liberal arts" and "professional courses;" others feature service learning. An example of this within the data science undergraduate experience is the Statistics Living-Learning Community at Purdue University, which provides a small group of sophomore students with the ability to live together in the same dorm; take courses together in probability theory, statistical theory, and data analysis; and conduct a year-long research project together (Purdue University, 2013).

Writing-Intensive Courses

These courses emphasize writing at all levels of instruction and across the curriculum, including final-year projects. Students are encouraged to produce and revise various forms of writing for different audiences in different disciplines. The effectiveness of this repeated practice "across the curriculum" has led to parallel efforts in such areas as quantitative reasoning, oral communication, information literacy, and, on some campuses, ethical inquiry. This practice is currently incorporated into the core curriculum at Columbia University through the University Writing Course "Readings in Data Science," which aims to enhance students' ability to understand data science through enhancement of reading and writing techniques (Columbia University, 2013). This aligns with the curriculum practices recommended by the American Statistical Association to teach students how to strengthen communication skills within the data science field (ASA, 2014).

Collaborative Assignments and Projects

Collaborative learning combines the following two key goals: learning to work and solve problems in the company of others and sharpening one's own understanding by listening seriously to the insights of others, especially those with different backgrounds and life experiences. Approaches range from study groups within a course, to team-based assignments and writing, to cooperative projects and research. An example of this currently in place at a university level is CS169, "Software Engineering," at the University of California, Berkeley. This course is designed for students to work in groups of six with an outside client to develop a "software-as-a-service" deliverable by course completion, highlighting the importance of applying skills learned inside the classroom to a real-world application (University of California, Berkeley, 2017). Another example is the Michigan Data Science Team run out of the Michigan Institute for Data Science at the University of Michigan. This not-for-credit extracurricular activity is organized and run by undergraduate computer science and data science students with light faculty oversight. Students collect and analyze data in the context of an application with high impact on local community or society. For example, in 2016 they applied their skills to help the city of Flint, Michigan, cope with the lead contamination crisis through better data collection and analysis (Meisler, 2017).

Undergraduate Research

Many colleges and universities are now providing research experiences for students in all disciplines. Undergraduate research, however, has been most prominently used in science disciplines. With strong support from the National Science Foundation (NSF) and the research community, scientists are reshaping their courses to connect key concepts and questions with students' early and active involvement in systematic investigation and research. The goal is to involve students with actively contested questions, empirical observation, cutting-edge technologies, and the sense of excitement that comes from working to answer important questions. This is demonstrated through NSF's Research Experience for Undergraduates, which places undergraduate students at a research institution for the summer to conduct innovative research across disciplines, including data science (NSF, 2017). This emphasizes the importance in exposing students to real data and problems, which can be messier and more challenging to work with than data sets used within the classroom.

Diversity/Global Learning

Many colleges and universities now emphasize courses and programs that help students explore cultures, life experiences, and worldviews different from their own. These studies—which may address U.S. diversity, world cultures, or both—often explore "difficult differences," such as racial, ethnic, and gender inequality, or continuing struggles around the globe for human rights, freedom, and power. Frequently, intercultural studies are augmented by experiential learning in the community and/or by study abroad. For instance, the "Comparative Public Health: The U.S. and the World" study abroad program at St. Olaf College provides students a chance to explore public health facilities at the Centers for Disease Control and Prevention in Atlanta, Georgia, as well as the World Health Organization in Geneva, Switzerland. This interdisciplinary program allows for undergraduates to explore and compare international public health efforts while individually exploring a public health topic of interest (Legler, 2017). Additionally, the Undergraduate Fellowships for Community Engaged and Translational Research at Virginia Commonwealth University seeks to provide funding to a select few undergraduate research projects conducted with a community partner, with at least one project a year dedicated to human health (VCU, 2017).

Service Learning, Community-Based Learning

In these programs, field-based "experiential learning" with community partners is an instructional strategy—and often a required part of the course. The idea is to give students direct experience with issues they are studying in the curriculum and with ongoing efforts to analyze and solve problems in the community. A key element in these programs is the opportunity students have to both apply what they are learning in real-world settings and reflect in a classroom setting on their service experiences. These programs model the idea that giving back to the community is an important college outcome, and that working with community partners is good preparation for citizenship, work, and life. An example of this is the Center for Data Science and Public Policy at the University of Chicago, which has several opportunities to engage students in the intersection of public policy and data science. This includes undergraduate coursework opportunities such as "Data Analytics for Campaigns," "Machine Learning for Public Policy," and "Computation and Public Policy," and the Eric and Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago, which brings together undergraduate- and graduate-level data scientists from around the globe to solve real-world social problems in conjunction with government agencies and nonprofits (University of Chicago, 2017).

Internships

Internships are another increasingly common form of experiential learning. The idea is to provide students with direct experience in a work setting—usually related to their career interests—and to give them the benefit of supervision and coaching from professionals in the field. If the internship is taken for course credit, students complete a project or paper that is approved by a faculty member. The variety of internship opportunities available within the data science field provides students the opportunity to gain real exposure and explore all aspects of the field, from the more technical side to potential policy implications. There are a myriad of undergraduate data science internship examples, ranging from those housed at large technology companies to specialized programs such as the Atlanta Data Science for Social Good program.

Capstone Courses and Projects

Whether they are called "senior capstones" or some other name, these culminating experiences require students nearing the end of their college years to create a project of some sort that integrates and applies what they have learned. The project might be a research paper, a performance, a portfolio of "best work," or an exhibit of artwork. Capstones are offered both in departmental programs and, increasingly, in general education. For example, the Statistics Capstone Course at the University of Georgia provides students the opportunity to engage in a year-long data analytics project that enhances understanding of advanced statistical material while reinforcing oral and written communication skills (Lazar et al., 2012). Similarly, the computer science capstone courses at Virginia Tech requires undergraduates majoring in that field to pursue a design-intensive, team-based final project within the area of their choosing, ranging from "Issues in Scientific Computing" to "Human–Computer Interaction" (Virginia Polytechnic Institute and State University, 2007).

Undergraduate Teaching Assistantships

Undergraduate teaching assistantships are another form of experiential learning. In the State University of New York system, for example, "a student enrolled in a credit-bearing course with specific student learning outcomes to assist faculty in providing instructional support" can serve as an undergraduate teaching assistant. During such an experience, a teaching assistant for a data science course has the opportunity to learn about the

teaching and learning processes, acquire expertise in data science concepts, hone oral and written communication skills, develop leadership and teamwork skills, and practice good time management. Whether or not teaching assistants plan to join the teaching profession, this experience and the skills it fosters prepares these students for a wide range of data science career opportunities (SUNY, 2012).

[a] The committee adapted this list from Kuh (2008) and provided additional content relevant to this interim report.

TRANSLATIONAL SKILLS

In addition to developing foundational skills, it is valuable for data science students to apply techniques and technologies learned in the classroom or laboratory to specific and different situations in practice. In other words, educators would train students to *do* data science in real application contexts, incorporating real data, broad impact applications, and commonly deployed methods, as well as working in teams. Students benefit from experiences such as carrying out sentiment analysis of texts, generating interactive maps to explore spatial data, assessing relationships between links within social networks, drawing samples from a distribution, visualizing multidimensional data to draw conclusions, and make decisions using data from a variety of sources and domains.

It is useful for students to learn how to translate understandings across domains and think critically about assertions and also to appreciate the importance of reusing and sharing data. As an example, consider the insights that can be found from digitizing an entomological collection. While the images originally might have been thought to be important only for understanding insects, later study of the pollen on the legs of the insects could yield unique insights into the changing nature of ecological systems with local climate change over more than a century. There is great potential for unanticipated ancillary derivatives from the data generation and integration process, but only if the right foundational knowledge is instilled about how one can and should combine sources and share findings in a reproducible workflow for others to interrogate.

Students also benefit from experiences with integrating diverse data and accounting for outside factors. For example, randomized trials—where a treatment is applied to a randomly selected subgroup of a population and then the outcomes of the full group are tracked—are the gold standard of evidence-based practices. However, conducting and generalizing from randomized trials can be difficult in many settings. Since most trials have issues with adherence, compliance, and nonresponse, it is important to account for post-randomization factors. Similarly, survey methods that undergird large surveys and censuses, such as those undertaken by the Census Bureau and Bureau of Labor Statistics, have contributed greatly to understanding and decision making about our society and economy. Today, these surveys can be enhanced and complemented—but not replaced—by fusing data that do not arise from a well-characterized sampling frame and design with data that were carefully sampled and vetted. While this integration may not be straightforward, it has the potential to extend the reach of ongoing surveys and to answer new questions in different ways.

A major challenge relates to how results that are demonstrated using unstructured data are verified. This struggle is analogous to that of research environments adapting to the arrival of a new instrument. For example, the microscope, telescope, and genetic sequencing machines have all allowed researchers to resolve something previously unresolvable. Scientific cultures have to grapple with what to do when observations are unprecedented and when old methods are insufficient to determine whether the new instruments are accurate or not. The development of new data science methods to address data challenges and undertake new analyses will require similar adaption. This analogy may also apply to the expanding computational infrastructure and capacity, which will continue to provide the opportunities to carry out analyses that were previously infeasible.

It is also important to better understand what types of questions and information are amenable to data science approaches. For example, understanding and conveying what computational approaches have produced and why is an important skill. New frameworks and models for carefully constructed distillations to be communicated are needed, along with reports that can be validated, reproduced, and assessed. Similar to the "John Henry" folklore[1] that pitted man against machine, data science is furthering consideration of where the edges of human abilities are, what can be measured, and what can be analyzed.

Educational systems and structures need to prepare students to inhabit a world that will have different tools than those currently available. Students develop judgment through the practice of working through the entire data science cycle. They benefit from opportunities to gradually build large systems by composing smaller systems where the behavior of the smaller systems is better understood. While some curricula offer these opportunities in the form of a capstone course, internship or externship opportunities, or similar integrative experience, it is beneficial for additional complementary experiences to be provided earlier in the curriculum.

Finding 2.2: It is important for data science education to incorporate real data, broad impact applications, and commonly deployed methods.

ETHICAL SKILLS

In addition to the foundational and translational skills training that students receive, they would also benefit from a better understanding of ethics and social context of data (O'Neil, 2016; DeVeaux et al., 2017). Several ethical considerations and corresponding examples that could be discussed include the following:

- *Fairness.* This multifaceted consideration can be summarized as the ability for data science techniques to treat all people equitably and avoid bias that may be inherent in training data sets. This is an especially important factor for applications that directly affect individuals, such as in the design of criminal justice models that determine sentencing practices without introducing racial or socioeconomic bias (Berk et al., 2017).
- *Validity.* Before data science methodologies can be applied, it is vital to ensure that the data set contains valid (e.g., accurate and relevant) information. Use of data that are falsified, not current, incomplete, from an unbalanced sample, biased in terms of survivorship, or not measuring the appropriate factors could lead to faulty conclusions, such as inaccurate estimates for health care needs in a given area based on outdated survey information (IOM, 2009).
- *Data context.* Similar to validity, it is important for individuals to understand the context of data sets before they are processed and analyzed. Knowing where, when, and how data were collected could lead to important insights that aid in analysis, detect inherent biases, and mitigate risk to individuals whose information is contained within the data sets.
- *Data confidence.* Recognition of the limitations of data science is important for avoiding overconfidence and the inclination to draw stronger-than-appropriate conclusions. This "data hubris" could be detrimental, for example, if too much confidence is invested in a data science model that makes stock market predictions for investments without any consideration of model limitations (Zacharakis and Shepherd, 2001).
- *Stewardship.* In data science, stewardship refers to the supervision of a data set at all stages of existence, including collection, storage, and analysis. This facilitates protection of individuals

[1] As the legend goes, John Henry won but died from his efforts (see Johnstown Area Heritage Association, 2013).

whose information is within the data set, including considerations of intellectual property rights or cybersecurity risk.

- *Privacy.* Considerations regarding individuals' privacy with respect to how data are collected and analyzed arise in many disciplines. Further information on privacy concerns in terms of data science is discussed later in this section.

By developing a truly integrative curriculum, it is possible to explore how society is affected by and reflected in data. Students would learn the importance of asking the following sequence of questions frequently, consistently, and thoughtfully: Whose data is being collected, by whom, for what purposes, and with what possible implications? Data are often representations (and simplifications) of the lives of people; this point could be integrated into every phase of a data science curriculum. Ethics plays a central role as students learn to problem-solve with data. An important lesson for students to learn is that transparency, trust-building, and validation/replication are key concepts; reputable data scientists are able to show why they do their work, explain the benefits that will emerge from it, and characterize and communicate the limitations of that work.

The trade-offs related to privacy play a key role in this discussion as well. A question arises about whether people and their information are "public by default" (boyd, 2010), because levels of involvement are different and ever-changing for each individual. Moreover, data from multiple sources can be combined, enabling an even richer and more intimate understanding of subjects. For example, with widespread adoption of mobile devices, data scientists may have access to detailed information about a person's location over time, which, in combination with other data, may yield far more information than the individual intended them to know. Individuals appreciate the ability to make choices about and have control over their information; they want secure data sharing, clear disclosure mechanisms, and a process to gain reparation from damages due to data breaches. All of these real-world problems could provide robust content in a data science curriculum.

An example of a course that integrates a study of data with a study of social context is "Data in Social Context"[2] at Virginia Tech. This course promotes dual literacy (i.e., humanities skills for data analytics students and data analytics skills for humanities students), explores why people turn to data to explain historical phenomena, and shows students a different way to approach questions with accessible tools and data. It highlights how valuable social context is in data analytics; data are filled with narratives, and questions often arise about ethics, probability, and bias.

Finding 2.3: Incorporating ethics into an undergraduate data science program provides students with valuable skills that can be applied to complex, human-centered questions across disciplines.

PROFESSIONAL SKILLS

Broad professional skills are particularly critical in data science (BHEW, 2017; Hicks and Irizarry, 2017). Industry partners relate that desirable characteristics include the ability to state goals clearly, to validate solutions, and to communicate with both technical and nontechnical audiences. Communication, both written and verbal, plays a significant role in data science because of diverse application areas, interdisciplinary research groups, and the ubiquity of data spanning many fields and being produced by many people. Conveying information with diverse audiences, expressing nuance regarding evidence in the presence of uncertainty, communicating limitations of analyses, and ensuring that what is conveyed is a faithful and honest representation of the data are all essential to data science.

Communication skills can be strengthened through practice with communicating various types of information to diverse audiences, such as a course or experiences that emphasize public speaking as well as technical and nontechnical writing. Communication can also be strengthened by improved

[2] The website for "Data in Social Context" is http://ethomasewing.org/disc/, accessed August 21, 2017.

understanding of diverse audiences. For example, what aspects of the data science process and results would domain scientists need to know to further their research, versus what would managers need to know to make relevant business decisions, versus what would policy makers need to know to make sound policies? These courses could also include a section on effective data visualization and its benefits, especially when explaining best communication practices for an audience from a nontechnical background.

The ability to work well in multidisciplinary teams is also important to data science and highly valued by industry. Multidisciplinary teamwork offers students the opportunity to use creative problem solving and refine leadership skills, and allows for diverse perspectives when tackling data science problems.

> **Finding 2.4:** Strong oral and written communication skills and the ability to work well in multidisciplinary teams are critical to students' success in data science.

3

Data Science Education in the Future

What implications does the emergence of data science have for colleges and universities? This fundamental question leads to many other related questions, such as the following: How are tools and methods used in data science blended with other disciplines? How can students benefit from data science being delivered in unique ways? How does data science span siloed disciplines?

Before making curricular changes, an academic institution needs to take into account its infrastructure, its budget, and its business model, as well as the potential collateral benefits for the rest of the institution. Some universities start with curricular changes at the master's level because those programs are generally easier to develop than undergraduate programs; however, because professionals with undergraduate degrees will be using different skill sets to fill different workforce roles than those with graduate degrees, essential data science skills training needs to be included at all levels of postsecondary education.

This chapter describes alternatives and models for innovative curriculum development, provides suggestions for institutions, and features examples of both innovative data science curricular approaches (see Box 3.1) as well as innovative approaches to evaluation.

INNOVATIVE CURRICULUM DEVELOPMENT

New and modified courses may be necessary to teach in the emerging data science field. Curriculum development would benefit from drawing on the experiences of both new faculty with first-hand knowledge of emerging areas and more seasoned faculty with experience developing other curriculum initiatives. Co-curricular activities (activities that are connected to or mirror the academic curriculum) and interdisciplinary approaches also enhance educational experiences. Both established good practices and careful evaluations can help inform decisions among different approaches. There are a number of educational models that can be implemented in a data science curriculum depending on the curriculum goals that are identified. For example, a variety of educational pathways may each have their advantages in preparing students for their respective careers, whether these careers are vocational or professional. These pathways may include majors (and minors) centered on data science broadly, distinct data science concentrations in various fields (such as business, computer science, and statistics), or 1-year courses and certificate programs. Although full degree-granting programs in data science may not be available yet in many settings outside top-tier academic institutions, graduates will still need to come from community colleges, minority-serving institutions, and smaller colleges and universities in order to fill the pipeline of data talent. Professional master's programs provide another pathway, because they have been helpful in some other interdisciplinary fields in providing a flexible environment to explore new educational programs and preparing students for the kind of workforce positions that data science is already providing.

The process whereby the goals are achieved can be varied. In terms of delivering content, flipped courses, hybrid courses, independent studies, experiential learning, modular courses, hackathons,[1] data dives,[2] and just-in-time learning are all viable options for students, although it is important for course design to align with course objectives and that courses are taught by faculty with the right instructional skills. In terms of creating content, it can be beneficial to teach some courses with a disciplinary context so that students appreciate that data science is not an abstract set of approaches.

One of the limitations of course deployment is the ability of faculty to teach data science. Co-teaching allows faculty with diverse areas of expertise the opportunity to collaborate and offer a more well-rounded course to the students; however, additional administrative support is often necessary, given the often larger resource needs for co-taught courses. Faculty may need to participate in re-training and faculty development to keep pace with changing technologies. Portable courses, where a course developed for one institution is replicated at another, help supplement faculty knowledge gaps.

The differences among various programs, and the types of employees they create, are often unclear to both hiring organizations and human resource departments. Having industry involved in developing and/or retooling data science courses can help ensure that programs meet workplace needs and that students going through these data science programs have employment opportunities upon completion. Improved collaboration can also help shape and enhance career paths in industry with positions that can both utilize data science skill sets and provide interesting opportunities for growth. Better integration of industry teachers in academia could help foster this collaboration.

A first step in establishing a new curriculum is to consider relevant experiences from other disciplines (e.g., the digital humanities) that have recently emerged from a period of reorganization and innovation. However, it is important to note that no single curriculum will be appropriate for all institutions. It may be necessary for educators to consider possible new disciplines and alternative degree structures and student pathways. Curricular development for quantitative material typically involves prioritizing concepts and skills, using educational research to guide pedagogy, borrowing or adapting existing materials, sequencing coverage of concepts, enhancing high-priority concepts and skills through repetition, and assessing the results. Diverse pedagogical approaches are valuable; institutions could begin by encouraging collaboration between departments and existing programs in considering new curricular initiatives.

For data science curricula specifically, course assignments and exercises can help motivate and ground exploration if data sets, case studies, and examples are chosen thoughtfully. Current research problems may be a good source of compelling topics. Heterogeneity of examples from diverse disciplines is desirable for better, richer team experiences, and data science education will evolve as data science itself evolves.

The evaluation and assessment of different approaches benefit from being grounded in the same theory of change[3] that informs curriculum development, including specifying curriculum goals, identifying the right comparison group, and stating clearly the curricular intervention. It is helpful to disseminate the analytical results broadly so that the field can learn from both successes and failures.

---

[1] A "hackathon" is a competition in which programmers and analysts work together to complete a task, usually for a prescribed and limited time period.

[2] A "data dive" is an event in which organizations, often nonprofits, present a data-driven problem to a group with data science expertise to solve in a limited amount of time.

[3] The Theory of Change is "a comprehensive description and illustration of how and why a desired change is expected to happen in a particular context. It is focused in particular on mapping out or 'filling in' what has been described as the 'missing middle' between what a program or change initiative does (its activities or interventions) and how these lead to desired goals being achieved. It does this by first identifying the desired long-term goals and then works back from these to identify all the conditions (outcomes) that must be in place (and how these related to one another causally) for the goals to occur" (Center for Theory of Change, 2016).

Finding 3.1: Data science curricula are enhanced by bringing together faculty from different disciplines, utilizing diverse pedagogical approaches, and building upon existing educational programs.

BOX 3.1
Examples of Current Data Science Programs

Carnegie Mellon University

Carnegie Mellon University offers various options for statistics majors and tracks, emphasizing the interdisciplinary nature of data science education.[a] Majors include statistics, economics-statistics, and statistics-machine learning; and tracks include mathematical statistics and statistics-neuroscience. This interdisciplinary undergraduate statistics program incorporates a breadth of topics central to the study of data science—real problems, lessons in reproducibility, statistical computing, advanced data analysis, methodology courses, oral and written communication, and interdisciplinary projects.

University of Illinois, Urbana-Champaign

The data science discipline at the University of Illinois, Urbana-Champaign,[b] demonstrates the key theme of interdisciplinary data science education with a particular focus on the diversity of collaborative programs. Students choose from two broad bachelor's degree options: (1) a degree in statistics and computer science (with data science electives) or (2) a degree in discipline "X," which has a split curriculum consisting of half computer science and half discipline "X." Current discipline "X" options include anthropology, linguistics, chemistry, and astronomy. Crop science, advertising, philosophy, music, and geoscience have been approved as future "X" options. Data science efforts taking place through a massive open online course at the master's level, including topics in machine learning and analytics, probability and advanced methods, and lifecycle and curation, could migrate down to the undergraduate level.

University of California, Berkeley

The data science curriculum at the University of California, Berkeley[c] focuses on data science applications. It takes a modular approach to the undergraduate data science experience by fitting courses into an existing curriculum using a data science foundation course ("Data 8") and "connectors" (data-science-in-the-field courses). "Data 8" is designed for first- and second-year students of any major and offers hands-on experiences with real data while learning computing, statistical concepts, and programming. The course enrolls at least 500 students per semester. Examples of data science connectors include "Data Science for Smart Cities," "Data Science and the Mind," "Genomics and Data Science," and "Making Sense of Cultural Data." Current work is under way to develop a data science major.

University of Michigan

The summer programming for data science at the University of Michigan highlights an innovative educational pathway available to students of all backgrounds. The Big Data Summer Institute (funded by the National Institutes of Health and the Michigan Institute for Data Science at the University of Michigan)[d] is a 6-week-long interdisciplinary training and research program in biostatistics that introduces undergraduate students to the intersection of big data and human health. Drawing from the expertise and experience of biostatistics, statistics, and electrical engineering and computer science departments at the University of Michigan, the institute exposes undergraduate students to diverse experiences and techniques while also offering research experiences, professional development, journey lectures, and social and networking events. Students come from both traditional and nontraditional educational backgrounds. The Big Data Summer Institute shows how formal opportunities for teaching data science can extend beyond the classroom to affect a broad group of students by using co-curricular structures.

Moore-Sloan Data Science Environments

The Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation sponsor three Moore-Sloan Data Science Environments at New York University, the University of California, Berkeley, and the University of Washington. These environments aim to increase data science research across multiple domains by

supporting cross-institutional efforts, specifically in areas of career development, education and training, tools and software, reproducibility and open science, physical and intellectual space, and data science studies. The Education Working Group for the environments is "developing ways to use innovative teaching methods and formats to offer both formal and informal training in data science skills at undergraduate, graduate, and professional levels. Viewing education in its widest sense, the aim is to make data science technology and expertise far more accessible both within universities and beyond" (MSDSE, 2017). Collaborations between academic institutions and private foundations have the potential to catalyze research efforts and foster unique partnerships.

University of California, Irvine

The University of California, Irvine integrates statistics and computer science into a data science bachelor's degree. The Department of Statistics at the university is housed within the Donald Bren School of Information and Computer Sciences.[e] The variety of courses offered within the statistics department include classes on statistical concepts that are supplemented with computerized applications, such as "Introduction to Probability and Statistics for Computer Science," "Statistical Computing and Exploratory Data Analysis," and "Statistical Computing Methods." Undergraduate students who choose to pursue the data science bachelor's degree have the ability to put equal focus on computer science and statistics when choosing their coursework, as opposed to having to select a specialty. The undergraduate program culminates with a capstone project in the final year that allows for students to apply the statistics and computer science skills learned within the classroom to a large-scale, interdisciplinary, real-world problem.

University of Massachusetts Amherst

University of Massachusetts, Amherst (UMass Amherst) offers an interdisciplinary approach to data science education with programs coordinated across computer science, statistics, public health, management, and the social sciences. UMass Amherst recently created new data science tracks within its computer science and informatics undergraduate majors and is in the process of developing new "X + Computing" majors, with linguistics, humanities, and social science departments. Its Center for Data Science[f] coordinates with multiple colleges and schools across campus to create data science programs tuned to their own students. Jointly with the Department of Computer Science, the Department of Mathematics and Statistics has created a Data Science Certificate program, while also adding new courses on methods leveraging large data. The faculty from these two departments are also creating an introductory undergraduate course in data science foundations, which will serve as a general education course for UMass and other institutions. The Isenberg School of Management has created a new certificate in business analytics, a version of data science for business majors, while sharing some courses with statistics and computer science. The Computational Social Science Institute has dramatically increased hiring of computationally and statistically minded social science faculty and organizes a series of short courses in data science topics.

[a] The website for Carnegie Mellon's Department of Statistics is http://www.stat.cmu.edu/academics, accessed August 21, 2017.
[b] The website for the University of Illinois, Urbana-Champaign's Department of Computer Science is https://cs.illinois.edu/academics/undergraduate/degree-program-options/cs-x-degree-programs, accessed August 21, 2017.
[c] The website for the University of California, Berkeley's Division of Data Sciences is http://data.berkeley.edu, accessed August 21, 2017.
[d] The website for the University of Michigan's Big Data Summer Institute is https://sph.umich.edu/bdsi/, accessed August 21, 2017.
[e] The website for the University of California, Irvine's Department of Statistics is http://www.stat.uci.edu/, accessed August 21, 2017.
[f] The website for the University of Massachusetts Amherst's Center for Data Science is https://ds.cs.umass.edu/, accessed August 21, 2017.

SUGGESTIONS FOR INSTITUTIONS

Evaluation, Assessment, and Accreditation

Assessment of student skill and conceptual development within formal courses benefits from being informed by the overall objective of the program and its associated curriculum. It is helpful for departments and institutions to consider criteria and methods for evaluating the entire program, particularly for interdisciplinary programs such as data science. Program evaluation approaches would be best implemented throughout the process of program/curriculum development from the time when the objectives of the program are first considered. Such an evaluation process can inform the formative and summative assessment methods used within the program and may suggest particular metrics for success. Indeed, the methods of data science could readily be applied to ascertain the data to be collected, the analysis methods to be used, and the metrics through which any analysis can specify program success or suggest efforts to adapt or modify the program to meet the success metrics. Having an evaluation process in place could assist institutions in preparation for any formal accreditation that is already being utilized or that might arise as the field of data science grows. It could also enable program designers to characterize the outcomes of students trained in data science relative to comparable students trained in other fields.

Fortunately, there is substantial literature available on ways to build evaluation and assessment into program design. In a very influential series of papers, Handelsman et al. (2004) argue that new types of science need to adopt active learning techniques. By this, they mean changing teaching from a lecture-based format to one that has both *inquiry-based and modular-learning* components and that *treats students as scientists* who "develop hypotheses, design and conduct experiments, collect and interpret data, and write about their results" (Handelsman et al., 2004). Handelsman et al. (2007) note that the line between active learning and assessment is difficult to define, but teaching that promotes students' active learning (e.g., asking students to perform an action or task) can then help assess understanding. Hoey (2008) suggests that the key measures include the following:

- Knowledge of concepts in the discipline;
- Ability to conduct independent research;
- Ability to use appropriate technologies;
- Ability to work with others, especially in teams; and
- Ability to teach others.

Of course, new educational techniques have been scientifically evaluated in other contexts. The Department of Education's What Works Clearinghouse[4] provides a compendium of the results of different interventions at different grade levels. Universities are partnering with government agencies to link their data to administrative records to trace the earnings and employment outcomes of students well beyond their graduation date. For example, a pilot partnership among a number of universities (the University of California system, the University of Michigan, the University of Wisconsin, and the University of Texas), the Institute for Research on Innovation and Science,[5] and the U.S. Census Bureau is linking individual-level transcript data to Longitudinal Employer-Household Dynamics[6] program data. The plan is to scale the pilot nationally, if successful. The advantage of linking transcript data to administrative records is not only that it provides longitudinal information on the educational experience of students, but also that it permits the construction of comparison groups—outcomes of groups of

[4] The website for the What Works Clearinghouse is ies.ed.gov/ncee/wwc/, accessed August 21, 2017.

[5] The website for the Institute for Research on Innovation and Science is iris.isr.umich.edu, accessed August 21, 2017.

[6] The website for the Longitudinal Employer-Household Dynamics is lehd.ces.census.gov, accessed August 21, 2017.

students who took data science classes can be compared with those of groups who did not take data science classes.

There is much to be learned from the experience in other fields in moving from a curriculum based on providing content to one that is both interdisciplinary and concept-driven. In the biological sciences, Gutlerner and Van Vactor (2013) argue forcefully for the development of modular classes—what they call "nanocourses." Their approach brings together students from multiple backgrounds, engages faculty from a variety of disciplines, and creates "small discussion group activities that allow students to practice framing experiments into larger scientific contexts and disciplines" (Gutlerner and Van Vactor, 2013).

Faculty Involvement

Faculty who are already busy teaching, engaging in faculty service, and conducting research, and who may not have deep expertise in data science topics, can find it hard to find time for curriculum development. The structure and content of faculty training and incentives are crucial, as is time and funding to support curriculum development. Networks built among multiple faculty from multiple disciplines and professional development offered on a regular basis can enhance interdisciplinary innovation. If these do not occur, faculty may be less equipped to teach their students, and institutions might not benefit from the cross-department educational collaborations that characterize successful data science programs. When data science programs or curricula are being established, it is important for academic institutions to consider the balance of talent throughout institutes and within departments.

> **Finding 3.2:** Structured faculty training, meaningful incentives, and available time and funding to support curriculum development are all crucial to preparing faculty for data science education.

Structures of Academic Institutions

An institution's infrastructure and organizational structure can shape, and perhaps limit, the possibilities for future data science programs. Economic structures, especially tuition-driven models, often form silos within departments and between disciplines. For example, if there are no mechanisms in place to adjust tuition payments, faculty pay rates, faculty course-load distributions, and general education requirements to accommodate cross-department and cross-disciplinary course offerings, data science course options could be more limited in scope and reach a smaller audience of students. Institutional considerations about whether to admit students to a particular program or a general program or to deliver only in-person courses may also affect a data science program's ability to reach a more diverse group of students. Institutions offering flexible options for students who would like to enter the workforce with a wider range of skills may be more successful than those that offer only very restrictive degree programs. Academic institutions could benefit from the creation of an office devoted to campus data initiatives, such as data science education programs. Such an office could include support for teaching and a data facility that would provide access to data, including a secure environment for confidential data as well as information about data standards and reusable code (such as Jupyter Notebooks). Another possible consideration would be a "quantitative sciences education center" that is similar in structure and objectives to writing, computing, and statistics centers at most institutions but focused on development of broad student appreciation for quantitative sciences, including data science.

> **Finding 3.3:** Data science programs often adapt to the existing infrastructure and organizational structure of an academic institution, but infrastructure innovations by the institution (e.g., in data provision, data and code access, and data documentation) can help data science programs be more collaborative and multidisciplinary.

Importance of Flexibility

The rapid flux in data science concepts, tools, and applications make flexibility and adaptability key to developing, achieving, and maintaining successful data science programs. Although students are generally becoming more comfortable with some computing and data science contexts (e.g., more accustomed to connectedness in every facet of society, more accepting of new technologies like artificial intelligence, and more adept with software and devices), it is important to identify and address potential gaps in knowledge to ensure data science programs are accessible to all future students, regardless of past experiences. Approaches to attract students who are interested in data science but have less quantitative backgrounds are important. Employers' needs will also likely evolve as new skill sets are required. Educational institutions are also likely to experience change in their options for delivery, program types, and course content. Faculty flexibility in developing and modifying appropriate and timely course content to keep pace with the changes in the surrounding world may be invaluable. Instead of creating one-size-fits-all approaches to teaching data science concepts, institutions will have to remain open-minded and flexible to best meet the needs of students and the workplace.

Finding 3.4: To keep up with the quickly evolving field of data science and recruit students with more diverse backgrounds, educational approaches in data science need to be flexible in terms of what concepts, skills, tools, and methods are taught; how students are recruited; and how departments and programs collaborate to provide a full data science experience to students.

4

Broad Participation in Data Science

Data science programs have the potential to attract broad participation, including diverse members from different disciplines (including the humanities, social sciences, and the arts) and from populations that are underrepresented in other similar science, technology, engineering, and mathematics (STEM) fields (see Box 4.1). Part of this potential comes from the various compelling application areas of data science, including digital humanities, computational social science, public policy, and many others. There are also numerous skill sets that are currently captured under a *data scientist* label that span multiple training and education levels.

There are many current and recent efforts aimed at increasing diversity, inclusion, and broadening participation in fields related to data science. These approaches can serve to inform emerging data science programs to encourage broad participation by design. The following highlights just a few of these efforts:

- *NSF INCLUDES* (Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science)[1] is a National Science Foundation (NSF) initiative designed to enhance U.S. leadership in STEM discoveries and innovations while supporting efforts to develop talent from all sectors of society to build the STEM workforce. The initiative aims to improve the preparation, increase the participation, and ensure the contributions of individuals from groups that have traditionally been underrepresented and underserved in the STEM enterprise, including women, members of racial and ethnic groups, persons with disabilities, and persons with low socioeconomic status. Significant advancement of these groups would result in a new generation of promising STEM talent and leadership to secure the nation's future in science and technology.
- *InGenIOus* (Investing in the Next Generation through Innovative and Outstanding Strategies)[2] is a collaboration among mathematics and statistics professional societies and NSF that culminated in a July 2013 workshop devoted to identifying and envisioning programs and strategies for increasing the flow of mathematical sciences students into the workforce pipeline.
- *CS for All*[3] is a program that aims to provide *all* U.S. students the opportunity to participate in computer science and computational thinking education in their schools at the K–12 levels. Funded by NSF, this program focuses on researcher–practitioner partnerships that foster the research and development needed to bring computer science and computational thinking to all schools. Specifically, the program aims to provide high school teachers with the

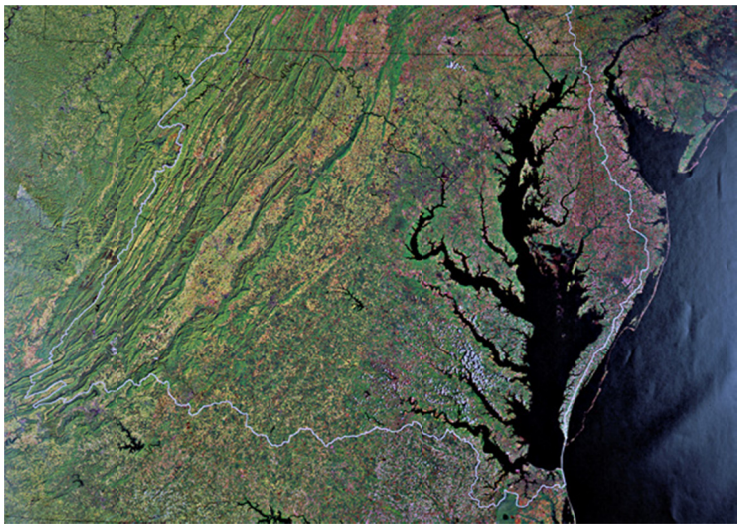
[1] The website for NSF INCLUDES is https://www.nsf.gov/pubs/2016/nsf16544/nsf16544.htm, accessed August 21, 2017.

[2] The website for InGenIOus is http://www.maa.org/programs/faculty-and-departments/ingenious, accessed August 21, 2017.

[3] The website for CS for All is https://www.nsf.gov/pubs/2017/nsf17525/nsf17525.htm, accessed August 21, 2017.

BOX 4.1
A Watershed Instead of a Pipeline

A "pipeline" metaphor has been a standard means to consider the flow of students through a STEM curriculum, with "leakage" used to indicate that some students step out of this path and potentially move to others. It has been argued that this metaphor should be replaced by a "watershed" in which there are multiple flow pathways by which students may enter a degree program dependent upon their own backgrounds. For inherently interdisciplinary degree programs with multiple potential routes for student success, such a metaphor structures a more open, collaborative approach toward building programs that attract diverse students than a fixed pipeline metaphor.



SOURCE: Library of Virginia (2014) via David Asai, Howard Hughes Medical Institute.

preparation, professional development, and ongoing support that they need to teach rigorous computer science courses, while providing K–8 teachers with the instructional materials and preparation they need to integrate computer science and computational thinking into their teaching.

- *StatFest*[4] is a conference hosted by the American Statistical Association's Committee on Minorities in Statistics. The goal of the program is to provide an opportunity for undergraduate students historically underrepresented in statistics to explore potential career options in the field and learn from industry and academic leaders. Each speaker session focuses on a different category of career trajectory, such as government, industry and consulting, academia, and graduate programs (ASA, 2017).
- *Math Alliance*[5] is an organization focused on assisting mathematics undergraduate students from historically underrepresented backgrounds in pursuing a doctoral degree in the mathematical sciences. Based out of Purdue University, the program strives to improve diversity and inclusion into mathematics doctoral programs (including pure and applied mathematics, mathematical and applied statistics, and biostatistics) while encouraging research collaborations and community within the broader mathematical community (National Alliance for Doctoral Studies in the Mathematical Sciences, 2013).

[4] The website for StatFest 2017 is http://community.amstat.org/cmis/events/statfest, accessed August 21, 2017.
[5] The website for Math Alliance is https://mathalliance.org/welcome/, accessed August 21, 2017.

There are also a variety of programs and projects that aim to foster interaction between individuals with varying levels of science background. For instance, a Music Data Science Hackathon hosted by Data Science London and EMI Music brought together data scientists and music specialists in a competition to develop a program that would predict the next hit in music. Other programs are housed at the institution level, such as the Mobile News App Design Class at the University of Texas, Austin, which brings computer science and journalism students together with the task of designing a novel mobile application for news. These, and many other efforts in this area, can inform emerging data science education programs and attract students from non-science disciplines to potentially pursue a path in data science. This chapter discusses some recruitment and retention strategies, institutional partnerships and K–12 outreach, and the role of evaluation and assessment.

RECRUITMENT AND RETENTION STRATEGIES

The focus on recruitment and retention extends beyond the obvious choices for data science majors and minors; other academically diverse student populations would benefit from the addition of data science to the curriculum. For example, it may be necessary to do targeted outreach to recruit students who are interested in enrolling in data science courses but may not find course titles immediately relevant or appealing. It may also encourage such students to know that a lack of preparation in data science does not equate to a lack of ability; developing multiple pathways to incorporate data science concepts into varied curricula via specialized connector courses or other "on ramps" could address these students' concerns about knowledge gaps and allow them to gain the level of expertise appropriate for their interests and career goals. Recruiting students from diverse disciplines to data science courses could also improve retention in such courses due to the increased interest in and value added to the courses. Retention may also improve if the content of and the faculty for introductory data science courses are selected with the diverse backgrounds and interests of the student population in mind. Throughout this process, it is important to consider strategies to retain faculty members as well. Data science skills and expertise are highly sought-after in industry, and educators in these areas are strong candidates for industry employment (Kaminski and Geisler, 2012).

Challenges also persist in both recruiting and retaining underrepresented minorities and women in the sciences more broadly. It is useful to consider whether data science's diversity and inclusion issues are unique as compared to those of other disciplines as well as what can be learned from other STEM programs that address diversity and inclusion well. Hiring a more diverse faculty may also help to attract a more diverse student population.

Recruitment and retention continue to be challenging in the workplace. In a professional research environment, employers want to hire people with literacy in computing, data science, and a domain science, but it can be difficult to find individuals who fit this description (Agarwal, 2016). Instead, employers often hire data science-literate people with domain expertise and provide more in-depth training in specific technical or professional skills. To increase diversity, inclusion, and data science literacy, employers could increase cross-disciplinary collaboration opportunities, create supportive team environments, counteract bias, and build mentor cohorts. Dedicated recruiting, inclusive recognition, and active support would also help.

> **Finding 4.1:** Data science has the potential to draw in a diverse set of students and build in broad participation from the onset, rather than trying to broaden participation later. However, strategies are needed to recruit and retain these students.

INSTITUTIONAL PARTNERSHIPS

Community colleges are well qualified to be highly effective providers of data science education while also serving as important partners for 4-year institutions that are considering the emerging role of data science education. Community college programs can serve to (1) be an entry point to inspire and attract diverse student populations to data science; (2) permit existing members of the workforce to retrain or obtain specific new skill sets to complement their education and experience; (3) create mechanisms by which students can certify specific or general skill sets with certificates or associate's degrees; (4) build foundational, translational, ethical, and professional skills to support matriculation into 4-year college data science programs; and (5) provide opportunities for advanced high school students to begin data science training early. The majority of these purposes support undergraduate education objectives, while also targeting the specific needs of industry. Institutional, industry, and government partnerships are all important to the development of data science education that meets these objectives for community colleges.

Funding agencies can help support formal partnerships with interested community colleges and 4-year institutions by providing funding mechanisms that allow for the development of new curricula as well as professional learning opportunities for faculty.

> **Finding 4.2:** Partnerships between 2- and 4-year institutions provide a valuable opportunity to develop innovative curricula, reach more diverse student populations, and expand the reach of data science education.

K–12 OBJECTIVES

Elementary, middle, and high schools play an important role in developing data science education and preparing students to thrive in a modern workforce. With changes in federal legislation that call for students to be prepared to succeed in college and careers, states are looking to national content standards to provide a vision for K–12 education. These standards of practice include content areas that are relevant to data science education, such as science and engineering (Next Generation Science Standards[6]) and mathematics and statistics (Common Core State Standards[7]). Some of the practices called for in these standards include analyzing and interpreting data; using mathematics and computational thinking; and obtaining, evaluating, and communicating information. Embedded in these practices are such skills as being able to (1) identify significant features and patterns in data through tabulation, graphical interpretation, visualization, and statistical analysis; (2) make and test predictions through constructing simulations and recognizing, expressing, and applying quantitative relationships; and (3) communicate orally or in writing using tables, diagrams, graphs, and equations (NRC, 2012, pp. 49-53). Through the adoption of these national standards, data scientists may be positioned to play a role in curriculum development by working with curriculum designers to ensure alignment between the practices highlighted above and the requisite skills that are needed upon entry into data science programs.

PUBLIC OUTREACH

In addition to efforts that could be achieved in formal educational spaces, there are outreach efforts to students in more informal spaces, including year-long afterschool programs, summer camps,

[6] The website for the Next Generation Science Standards is https://www.nextgenscience.org/, accessed August 21, 2017.

[7] The website for the Common Core State Standards is http://www.corestandards.org/, accessed August 21, 2017.

high school internship programs, competitions, and websites designed to foster motivation and interest in data science and other STEM fields. Focused data science summer programs can be effective in attracting high school students to postsecondary STEM fields. For example, the Michigan Institute for Data Science (MIDAS) at the University of Michigan has been running a data science summer camp for the past 2 years that uses art and sports activities to gently introduce teenagers to mathematics, computing, signal processing, and statistics. Such use of familiar activities to introduce data science makes it fun and sustains student interest in STEM. Such camps can also be used to reach a more diverse group of students.

These programs may also be offered by organizations, such as museums, that not only collaborate with the K–12 education system, but also seek to engage the broader community. For example, the Exploratorium[8] has a repository of online materials and activities as well as community-based programs designed to foster the development of the skills highlighted above. Other online resources, such as Data.gov, provide parents and teachers access to materials to help advance children's understanding of concepts associated with data science (Data.gov, 2013).

EVALUATION AND ASSESSMENT

Good data science practices can inform the evaluation of programs targeted toward broad participation. It is useful for program evaluation to follow established best practices, including following an appropriate model for inclusion of metrics for participation in the overall goals of the program, clearly articulating these to all participants from the beginning of the program, establishing procedures for assessing these metrics on a regular basis, and specifying adaptation and modification procedures based on these formative and summative assessments.

In establishing approaches for measuring success, the tools of experimental design and analysis can be incorporated when appropriate (using, for example, comparison of treatment and control, randomized trials, nationally normed instruments, exploitation of natural experiments, appropriate descriptive analyses of observational data accounting for confounding factors, etc.). It may be necessary to consider an overall data plan at the start of the program as part of the evaluation plan, which would account for Institutional Review Board requirements if the data might be used for research rather than just within institutional planning.

Data sources useful for measuring diverse participation might include transcripts that reveal who takes data science courses and who completes a data science degree. Such data can be used to make both programmatic and cross-institutional comparisons. Institutional constraints may encourage a return-on-investment perspective as part of the evaluation, incorporating the impacts and costs of various targeted educational interventions to broaden participation.

> **Finding 4.3:** Data science programs would benefit from ongoing curricular evaluation, especially with respect to how well curricular objectives are being met and the degree of curricular integration. Taking a cue from its own domain, these data could be used to inform data science instruction and curriculum.

[8] The website for the Exploratorium is https://www.exploratorium.edu/education, accessed August 21, 2017.

5

Reflections

Given the wide-ranging applications, potential impacts, and important implications for society, the committee began its reflections on the future of data science with aspects of ethical conduct as part of a broader set of skills and capacities.

HIPPOCRATIC OATH

Emerging data science technologies and methodologies (1) blur differences between "public" and "private" data, (2) offer more widespread access to data and related tools, (3) influence and affect society at large, and (4) create greater opportunities for deeper insights through the use and integration of multiple data sources. As a result, data ethics take on an ever more prominent role in both data science curricula and data science practice.

The Hippocratic Oath, which details the ideal conduct of physicians in terms of their treatment of patients and interactions with colleagues, has historically been affirmed by physicians to acknowledge their understanding of key ethical principles for their profession (Box 5.1). Similarly, the Canadian "Calling of an Engineer" ceremony for engineering graduates helps establish shared moral and social responsibilities (NSPE, 2009). The pervasive impact of data science suggests that a similar oath would be beneficial for data scientists, whose work has a direct impact on individuals throughout society and on the advancement of the body of scientific knowledge. Data science students learn to solve complex problems in the world and use data to make decisions, while understanding limitations of data sets and methods.

An oath of this sort may be helpful in formalizing the role of data ethics and to inspire future data scientists to practice with honor, "do[ing] no harm" to the subjects involved in or affected by their work. This oath also formalizes the professional role of the data scientist, offering guidance on appropriate conduct to those entering the field and encouraging collaboration across diverse communities.

What might a Hippocratic Oath for data science include? To explore this question, the committee developed the text in Box 5.2 as a preliminary form of a possible pledge for future data scientists. The proposed Data Science Oath highlights aspects of data ethics and the value of incorporating societal impact as part of data science education.

SUMMARY OF PRELIMINARY COMMITTEE FINDINGS AND OPEN QUESTIONS

At the midpoint of its study, the committee finds that it is important that data science education incorporate real data, broad impact applications, commonly deployed methods, and ethical considerations, as well as provide support for work in teams. Other critical content areas include data description and curation, mathematical foundations, computational thinking, statistical thinking, data modeling, computing, reproducibility, and data ethics. Students would also benefit from developing deep analytic and communication skills so as to better work with large, complex data sets and engage with diverse audiences about real-world problems that data science can help solve. All of these promote the

BOX 5.1
Hippocratic Oath[a]

I swear to fulfill, to the best of my ability and judgment, this covenant:
I will respect the hard-won scientific gains of those physicians in whose steps I walk, and gladly share such knowledge as is mine with those who are to follow.
I will apply, for the benefit of the sick, all measures which are required, avoiding those twin traps of overtreatment and therapeutic nihilism.
I will remember that there is art to medicine as well as science, and that warmth, sympathy, and understanding may outweigh the surgeon's knife or the chemist's drug.
I will not be ashamed to say "I know not," nor will I fail to call in my colleagues when the skills of another are needed for a patient's recovery.
I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know. Most especially must I tread with care in matters of life and death. If it is given me to save a life, all thanks. But it may also be within my power to take a life; this awesome responsibility must be faced with great humbleness and awareness of my own frailty. Above all, I must not play at God.
I will remember that I do not treat a fever chart, a cancerous growth, but a sick human being, whose illness may affect the person's family and economic stability. My responsibility includes these related problems, if I am to care adequately for the sick.
I will prevent disease whenever I can, for prevention is preferable to cure.
I will remember that I remain a member of society, with special obligations to all my fellow human beings, those sound of mind and body as well as the infirm.
If I do not violate this oath, may I enjoy life and art, respected while I live and remembered with affection thereafter. May I always act so as to preserve the finest traditions of my calling and may I long experience the joy of healing those who seek my help.

[a] Lasagna (1964).

BOX 5.2
Data Science Oath

I swear to fulfill, to the best of my ability and judgment, this covenant:
I will respect the hard-won scientific gains of those data scientists in whose steps I walk and gladly share such knowledge as is mine with those who follow.
I will apply, for the benefit of society, all measures which are required, avoiding those twin traps of data-fishing and analytic nihilism.
I will remember that there is art to data science as well as science, and that consistency, candor and compassion should outweigh the algorithm's precision or the interventionists influence.
I will not be ashamed to say, "I know not," nor will I fail to call in my colleagues when the skills of another are needed for solving a problem.
I will respect the privacy of my data subjects, for their problems are not disclosed to me that the world may know, so I will tread with care in matters of privacy and security. If it is given to me to save life with my analyses, all thanks. But it may also be within my power to do harm and this responsibility must be faced with humbleness and awareness of my own limitations.
I will remember that my data are not just numbers without meaning or context, but represent real people and situations and that my work may lead to unintended societal consequences, such as inequality, poverty, and disparities due to algorithmic bias. My responsibility must consider potential consequences of my extraction of meaning from data and ensure my analyses help make better decisions.
I will do personalization where appropriate, but I will always look for a path to fair treatment and non-discrimination.
I will remember that I remain a member of society, with special obligations to all my fellow human beings, those who need help and those who don't.
If I do not violate this oath, may I enjoy vitality and virtuosity, respected for my contributions and remembered for my leadership thereafter. May I always act to preserve the finest traditions of my calling and may I long experience the joy of helping those who seek my help.

development of data acumen. Highly trained and flexible faculty, innovative cross-disciplinary pedagogical approaches, and diverse participation would enhance learning experiences. Such programs' successes can then be evaluated and assessed using the very tools of experimental design and analysis common in the field of data science.

The findings from the preceding chapters are restated below along with key questions on which the committee would like to gather public input.

Finding 2.1: A critical component of data science education is to guide students to develop data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts related to developing data acumen include the following:

- Mathematical foundations,
- Computational thinking,
- Statistical thinking,
- Data management,
- Data description and curation,

- Data modeling,
- Ethical problem solving,
- Communication and reproducibility, and
- Domain-specific considerations.

The necessary levels of exposure to each area will vary based on the overall objectives and duration of the data science program as well as the goals for the students.

Questions

- *Which key components should be included in data science curriculum, both now and in the future?*
- *How could these components be prioritized or best conveyed for differing types of data science programs?*
- *How can opportunities to enhance data acumen (i.e., the ability to make good judgments and decisions with data) be integrated into data science educational programs?*
- *How can data acumen be measured or evaluated?*

Finding 2.2: It is important for data science education to incorporate real data, broad impact applications, and commonly deployed methods.

Questions

- *How can partnerships between industry and educational programs be encouraged?*
- *Could a focus on real problems serve as a means to attracting more diverse students?*
- *How can students gain access to real-world data sets?*

Finding 2.3: Incorporating ethics into an undergraduate data science program provides students with valuable skills that can be applied to complex, human-centered questions across disciplines.

Questions

- *How can ethical considerations be best incorporated throughout the data science curriculum?*

- *How can students be taught to apply ethical decision making throughout the problem-solving process?*

Finding 2.4: Strong oral and written communication skills and the ability to work well in multidisciplinary teams are critical to students' success in data science.

Questions

- *How can communication and teamwork be fostered in data science programs?*
- *What type of multidisciplinary teams serve as effective models for the real world? Will these groupings be different in the future?*

Finding 3.1: Data science curricula are enhanced by bringing together faculty from different disciplines, utilizing diverse pedagogical approaches, and building upon existing educational programs.

Questions

- *What are known good practices for fostering collaboration between departments and existing programs?*
- *What new directions and opportunities exist for new curricular initiatives?*

- *What pedagogical approaches are particularly relevant to data science, both now and in the future?*

Finding 3.2: Structured faculty training, meaningful incentives, and available time and funding to support curriculum development are all crucial to preparing faculty for data science education.

Questions

- *What types of training would be beneficial to faculty?*
- *How could incentives be restructured to encourage more faculty development in data science?*

Finding 3.3: Data science programs often adapt to the existing infrastructure and organizational structure of an academic institution, but infrastructure innovations by the institution (e.g., in data provision, data and code access, and data documentation) can help data science programs be more collaborative and multidisciplinary.

Questions

- *What are current infrastructure obstacles and how can they be rethought going forward?*
- *How could organizational structures be modified and/or incentives added to encourage data science collaboration and innovation?*

Finding 3.4: To keep up with the quickly evolving field of data science and recruit students with more diverse backgrounds, educational approaches in data science need to be flexible in terms of what concepts, skills, tools, and methods are taught; how students are recruited; and how departments and programs collaborate to provide a full data science experience to students.

Questions

- *How can data science programs build in flexibility and adaptability so they can be most responsive to changes in the field?*
- *How can flexibility encourage more diverse students?*

Finding 4.1: Data science has the potential to draw in a diverse set of students and build in broad participation from the onset, rather than trying to broaden participation later. However, strategies are needed to recruit and retain these students.

Questions

- *How can broad participation, diversity, and inclusion be ingrained in data science programs?*
- *What strategies to recruit and retain diverse students can data science programs deploy, and what examples can inform these efforts?*

Finding 4.2: Partnerships between 2- and 4-year institutions provide a valuable opportunity to develop innovative curricula, reach more diverse student populations, and expand the reach of data science education.

Questions

- *How can partnerships between 2- and 4-year institutions be facilitated?*
- *How do the skills and concepts taught at a 2-year institution vary based on students' goals?*
- *What aspects of data science education are appropriate and feasible to develop at 2-year institutions?*

Finding 4.3: Data science programs would benefit from ongoing curricular evaluation, especially with respect to how well curricular objectives are being met and the degree of curricular

integration. Taking a cue from its own domain, these data could be used to inform data science instruction and curriculum.

Questions

- *What evaluation and assessment objectives are currently being used in data science programs, and how will these differ in the future?*
- *What best practices in evaluation and assessment can inform data science programs?*
- *What data are available to evaluate the effectiveness of different data science approaches?*
- *What standard evaluation approaches should be adopted?*

INPUT NEEDED

The committee seeks input from the growing data science community and the public on the following topics:

- Additional content for its study, including but not limited to case studies from institutions providing data science education, innovative ways to bring researchers together, best practices for program evaluation, and ideas for future topical webinars;
- The proposed Data Science Oath outlined at the beginning of this chapter; and
- The questions posed in the previous section.

Please visit the following webpage to provide input: http://www.nas.edu/EnvisioningDS.

References

Agarwal, D. 2016. "Data Science Diversity from the Perspective of a National Laboratory," presentation at the National Academies' Data Science Education Workshop, Washington, D.C., December 12.

ASA (American Statistical Association). 2014. "Curriculum Guidelines for Undergraduate Programs in Statistical Sciences." http://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf.

ASA. 2017. "StatFest 2017." http://community.amstat.org/cmis/events/statfest. Accessed June 28, 2017.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2017. *Fairness in Criminal Justice Risk Assessments: The State of the Art.* University of Pennsylvania Departments of Statistics, Criminology, and Computer and Information Science. Philadelphia, Pa. https://arxiv.org/pdf/1703.09207.pdf.

BHEW (Business-Higher Education Forum). 2017. *Investing in America's data science and analytics talent: The case for action*. http://www.pwc.com/us/dsa-skills. Accessed August 21, 2017.

Boyd, D. 2010. "Making Sense of Privacy and Publicity." SXSW, Austin, Tex., March 13.

Butler, D. 2013. When Google got flu wrong. *Nature* 494: 155-156.

Center for Theory of Change. 2016. "What is Theory of Change?" http://www.theoryofchange.org/what-is-theory-of-change/.

Cetinkaya-Rundel, M. 2017. "Teaching Data Science and Statistical Computation to Undergraduates." United States Conference on Teaching Statistics, State College, Pa., May 20. https://www.causeweb.org/cause/uscots/uscots17/keynote/3.

Chen, H., R.H.L. Chiang, and V.C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36(4): 1165-1188.

Codella, N.C.F., Q.B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J.R. Smith. 2017. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development* 61(4/5).

Columbia University. 2013. "University Writing: Readings in Data Sciences." https://www.college.columbia.edu/core/node/3290/.

Columbus, L. 2017. "IBM Predicts Demand For Data Scientists Will Soar 28% By 2020." *Forbes*, May 13.

Culler, D. 2016. "Data Science at the Heart of a 21st Century University," presentation to the National Academies' Committee on Envisioning the Data Science Discipline: the Undergraduate Perspective, Washington, D.C., December 12.

Danyllo, W.A., V.B. Alisson, N.D. Alexandre, LM.J. Moacir, B.P. Jansepetrus, and R.F. Oliveira. 2013. "Identifying Relevant Users and Groups in the Context of Credit Analysis Based on Data from Twitter." Paper presented at the 2013 IEEE Third International Conference on Cloud and Green Computing, September/October, Karlsruhe, Germany.

Data.gov. 2013. "Data in the Classroom." https://www.data.gov/education/classroom/.

De Veaux, R., M. Agarwal, M. Averett, B. Baumer, A. Bray, T. Bressoud, L. Bryant, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Applications* 4: 15-30.

Groves, R.M. 2011. Three eras of survey research. *Public Opinion Quarterly* 75(5): 861-871.

Gutlerner, J.L., and D. Van Vactor. 2013. Catalyzing curriculum evolution in graduate science education. *Cell* 153(4): 731-736.

Handelsman, J., S. Miller, and C. Pfund. 2007. *Scientific Teaching*. New York: W.H. Freeman.
Handelsman, J., R. Ebert-May, P. Beichner, A. Bruns, R. Chang, J. DeHaan, S. Gentile, J. Lauffer, J. Stewart, S.M. Tilghman, and W.B. Wood. 2004. Scientific teaching. *Science* 304: 521-522.
Hicks, S., and R. Irizarry. 2017. A guide to teaching data science. *American Statistician* http://dx.doi.org/10.1080/00031305.2017.1356747.
Hoey, J.J. 2008. Tools and assessment methods specific to graduate education. Pp. 149-167 in *Designing Better Engineering Education Through Assessment* (J.E. Spurlin, S.A. Rajala, and J.P. Lavelle, eds.). Sterling, Va: Stylus.
Hvistendahl, M. 2016. "Can 'predictive policing' prevent crime before it happens?" *Science,* October 5. http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens.
IOM (Institute of Medicine). 2009. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* Washington, D.C.: The National Academies Press.
Jefferson, T. 1803. "Jefferson's Instructions for Meriwether Lewis." In The Thomas Jefferson Papers, Library of Congress. https://www.loc.gov/exhibits/lewisandclark/transcript57.html.
Johnstown Area Heritage Association. 2013. "Discovery Center: Larger than Life, Part 1." http://www.jaha.org/edu/discovery_center/work/folk_hero.html.
Jordan, M. 2013. On statistics, computation and scalability. *Bernoulli* 19(4): 1378-1390.
Kaminski, D., and C. Geisler. 2012. Survival analysis of faculty retention in science and engineering by gender. *Science* 335(6070): 864-866.
Kitchin, R. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* 79: 1-14.
Kuh, G.D. 2008. *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter.* http://secure.aacu.org/imis/aacur.
Lasagna, L.C. 1964. *Hippocratic Oath*, Modern Version. The Johns Hopkins Sheridan Libraries and University Museums. http://guides.library.jhu.edu/c.php?g=202502&p=1335759.
Lazar, N.A., J. Reeves, and C. Franklin. 2012. A capstone course for undergraduate statistics majors. *The American Statistician* 65(3): 183-189.
Legler, J. 2017. "ID 280: Comparative Public Health: the US and the World." St. Olaf College. http://stolaf.studioabroad.com/_customtags/ct_FileRetrieve.cfm?File_ID=05027172754F73020D020306070B1C04080C0014757800006E06030E7A0577737302710307750471 72. Accessed June 28, 2017.
Library of Virginia. 2014. "Watershed." http://www.lva.virginia.gov/exhibits/mapping/map_images/satellite.jpg.
Meisler, D. 2017. "Google, U-M to build digital tools for Flint water crisis." *MDST Projects*, January 25.
Miller, S., and D. Hughes. 2017. *The Quant Crunch: How the Demands for Data Science Skills is Disrupting the Job Market.* https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IML14576USEN&. Accessed June 21, 2017.
MSDSE (The Moore-Sloan Data Science Environments). 2017. "Themes." http://msdse.org/themes/. Accessed August 21, 2017.
National Alliance for Doctoral Studies in the Mathematical Sciences. 2013. "Math Alliance Goals." https://mathalliance.org/goals/.
NRC (National Research Council). 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, D.C.: The National Academies Press.
NRC. 2014. *Training Students to Extract Value from Big Data*: *Summary of a Workshop*. Washington, D.C.: The National Academies Press.
NSF (National Science Foundation). 2017. "Research Experience for Undergraduates (REU)." https://www.nsf.gov/crssprgm/reu/. Accessed June 21, 2017.
NSPE (National Society of Professional Engineers). 2009. "Called to Order." https://www.nspe.org/resources/pe-magazine/called-order. Accessed August 21, 2017.
O'Neil, C. 2016 *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishers.

Pratt, M.K. 2016. "Big Data's Big Role in Humanitarian Aid." Computer World, February 8. http://www.computerworld.com/article/3027117/big-data/big-datas-big-role-in-humanitarian-aid.html.

Purdue University. 2013. "Statistics Living-Learning Community." http://llc.stat.purdue.edu/.

SUNY (State University of New York System). 2012. "Guide for Undergraduate Teaching Assistantships." http://system.suny.edu/media/suny/content-assets/documents/faculty-senate/UndergraduateTAGuideFinalversion.pdf.

University of California, Berkeley. 2017. "UC Berkeley CS 169 Software Engineering." http://cs169.saas-class.org/faq. Accessed June 21, 2017.

University of Chicago. 2017. "Learning Opportunities, Center for Data Science and Public Policy." http://dsapp.uchicago.edu/research-areas/learning-opportunities/. Accessed June 22, 2017.

VCU (Virginia Commonwealth University). 2017. "Undergraduate Fellowships for Community Engaged and Translational Research." http://www.research.vcu.edu/ugresources/ce_cctr_fellowship.htm. Accessed June 28, 2017.

Virginia Polytechnic Institute and State University. 2007. "Capstone Courses." http://www.cs.vt.edu/undergraduate/capstones.

Zacharakis, A.L., and D.A. Shepherd. 2001. The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing* 16(4): 311-332.

Appendixes

A

Biographies of the Committee

LAURA HAAS, *Co-Chair*, joined the University of Massachusetts Amherst in August 2017 as dean of the College of Information and Computer Sciences, after a long career at IBM, where she was accorded the title IBM Fellow in recognition of her impact. At the time of her retirement from IBM, she was director of IBM Research's Accelerated Discovery Lab (2011–2017), after serving as director of computer science at IBM's Almaden Research Center from 2005 to 2011. She had worldwide responsibility for IBM Research's exploratory science program from 2009 through 2013. From 2001–2005, she led the Information Integration Solutions architecture and development teams in IBM's Software Group. Previously, Dr. Haas was a research staff member and manager at Almaden. She is best known for her work on the Starburst query processor, from which DB2 LUW was developed, on Garlic, a system which allowed integration of heterogeneous data sources, and on Clio, the first semi-automatic tool for heterogeneous schema mapping. She has received several IBM awards for Outstanding Innovation and Technical Achievement, an IBM Corporate Award for information integration technology, the Anita Borg Institute Technical Leadership Award, and the Association for Computing Machinery (ACM) Special Interest Group on Management of Data Edgar F. Codd Innovation Award. Dr. Haas was Vice President of the Very Large Data Bases Endowment Board of Trustees from 2004–2009 and served on the board of the Computing Research Association from 2007–2016 (vice chair 2009–2015). She currently serves on the National Academies' Computer Science and Telecommunications Board (2013–2019). She is an ACM Fellow, a member of the National Academy of Engineering, and a Fellow of the American Academy of Arts and Sciences.

ALFRED O. HERO III, *Co-Chair*, is the John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and R. Jamison and Betty Williams Professor of Engineering at the University of Michigan. He received the B.S. (summa cum laude) from Boston University (1980) and 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 Ecole Superieure d'Electricite, Gif-sur-Yvette, France. He is an Institute of Electrical and Electronics Engineers (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 Millennium 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.

ANI ADHIKARI is a senior lecturer in statistics at the University of California, Berkeley, and she has received the Distinguished Teaching Award at Berkeley and the Dean's Award for Distinguished Teaching at Stanford University. While her research interests are centered on applications of statistics in

the natural sciences, her primary focus has always been on teaching and mentoring students. She teaches courses at all levels and has a particular affinity for teaching statistics to students who have little mathematical preparation. She received her undergraduate degree from the Indian Statistical Institute and her Ph.D. in statistics from the University of California, Berkeley.

DAVID CULLER received his B.A. from the University of California, Berkeley in 1980, and an M.S. and Ph.D. from the Massachusetts Institute of Technology in 1985 and 1989, respectively. He joined the electrical engineering and computer science (EECS) faculty in 1989, is the founding director of Intel Research, University of California, Berkeley, and was associate chair (2010–2012) and chair (2012–June 2014) of the EECS Department. He won the Okawa Prize in 2013. He is a member of the National Academy of Engineering, an ACM Fellow, and an IEEE Fellow. He has been named one of *Scientific American*'s Top 50 Researchers and the creator of one of *MIT Technology Review*'s "10 Technologies that Will Change the World." He was awarded the National Science Foundation Presidential Young Investigator and the Presidential Faculty Fellowship. His research addresses networks of small, embedded wireless devices; planetary-scale internet services; parallel computer architecture; parallel programming languages; and high-performance communication. It includes TinyOS, Berkeley Motes, PlanetLab, Networks of Workstations, Internet services, Active Messages, Split-C, and the Threaded Abstract Machine.

DAVID DONOHO is an Anne T. and Robert M. Bass Professor of Humanities and Sciences and professor of statistics at Stanford University. Dr. Donoho is a mathematician who has made fundamental contributions to theoretical and computational statistics, as well as to signal processing and harmonic analysis. His algorithms have contributed significantly to the understanding of the maximum entropy principle, of the structure of robust procedures, and of sparse data description. His theoretical research interests have focused on the mathematics of statistical inference and on theoretical questions arising in applying harmonic analysis to various applied problems. His applied research interests have ranged from data visualization to various problems in scientific signal processing, image processing, and inverse problems. He is a member of the National Academy of Sciences. Donoho received an A.B. from Princeton University and a Ph.D. from Harvard University.

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. His education included a B.A. from Williams College and a Ph.D. in history from the University of Michigan. He teaches courses in Russian, European, Middle Eastern, and world history; gender/women's history; and historical methods. 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). His articles on Stalinist education have been published in *Gender & History*, *American Educational Research Journal*, *Women's History Review*, *History of Education Quarterly*, *Russian Review*, and *The Journal of Women's History*. He has received funding from the National Endowment for the Humanities, the Spencer Foundation, and the National Council for Eurasian and East European Research.

LOUIS J. GROSS is a James R. Cox and Alvin and Sally Beaman Distinguished Professor of Ecology and Evolutionary Biology and Mathematics and director of the Institute for Environmental Modeling at the University of Tennessee, Knoxville (UTK). He is also director of the National Institute for Mathematical and Biological Synthesis, a National Science Foundation-funded center to foster research and education at the interface between math and biology. He completed a B.S. in mathematics at Drexel University and a Ph.D. in applied mathematics at Cornell University, and has been a faculty member at UTK since 1979. His research focuses on applications of mathematics and computational methods in

many areas of ecology, including disease ecology, landscape ecology, spatial control for natural resource management, photosynthetic dynamics, and the development of quantitative curricula for life science undergraduates. He led the effort at UT to develop an across-trophic level modeling framework to assess the biotic impacts of alternative water planning for the Everglades of Florida. He has co-directed several courses and workshops in mathematical ecology at the International Centre for Theoretical Physics in Trieste, Italy, and he has served as program chair of the Ecological Society of America, president of the Society for Mathematical Biology, president of the UTK Faculty Senate, treasurer for the American Institute of Biological Sciences, and chair of the National Research Council Committee on Education in Biocomplexity Research. He is the 2006 Distinguished Scientist awardee of the American Institute of Biological Sciences and is a fellow of the American Association for the Advancement of Science and of the Society for Mathematical Biology. He has served on the National Research Council Board on Life Sciences and was liaison to the National Research Council Standing Committee on Emerging Science for Environmental Health Decisions.

NICHOLAS HORTON is a professor of statistics at Amherst College. As an applied biostatistician, Dr. Horton's work is based squarely within the mathematical sciences but spans other fields in order to ensure that biomedical research is conducted on a sound footing. He has published more than 160 papers in the statistics and biomedical literature and 4 books on statistical computing and data science. He has taught a variety of courses in statistics and related fields, including introductory statistics, data science, probability, theoretical statistics, regression, and design of experiments. He 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 think with data. Dr. Horton received the American Statistical Association (ASA) Waller Award for Distinguished Teaching, the Mathematical Association of America Hogg Award for Excellence in Teaching, the Mu Sigma Rho Statistics Education Award, and the ASA Founders Award. He was a co-principal investigator of the National Science Foundation-funded Project MOSAIC, serves as the chair of the Committee of Presidents of Statistical Societies, is a fellow of the ASA, and was a research fellow at the Bureau of Labor Statistics. Dr. Horton earned his A.B. from Harvard College and his Sc.D. in biostatistics from the Harvard T.H. Chan School of Public Health.

JULIA LANE is a professor at the Center for Urban Science and Progress (CUSP), and at New York University's (NYU's) Wagner Graduate School of Public Service. She also serves as a Provostial Fellow for Innovation Analytics and Senior Fellow at NYU's GovLab. As part of the CUSP team, Dr. Lane works with the research team to build the CUSP Data User Facility. Dr. Lane is an economist who is the co-founder of the Longitudinal Employer-Household Dynamic (LEHD) partnership with the Census Bureau. LEHD data has been used to analyze commuting patterns for transportation planning, and the study of workforce turnover, pensions, and low-wage work. Dr. Lane has authored over 65 refereed articles and edited or authored 7 books. She has been working with a number of national governments to document the results of their science investments. Her work has been featured in several publications including *Science* and *Nature*. Work Dr. Lane started at the National Science Foundation (as senior program director of the Science of Science and Innovation Policy Program) to quantify the results of federal stimulus spending is the basis of the new Institute for Research on Innovation and Science at the University of Michigan. The data will be used to describe the structure of the research workforce, the nature and evolution of research collaborations, and the diffusion of sponsored research results. Dr. Lane has had leadership positions in a number of policy and data science initiatives at her other previous appointments, which include senior managing economist at the American Institutes for Research; senior vice president and director, Economics Department at NORC/University of Chicago; various consultancy roles at The World Bank; and assistant, associate, and full professor at American University. Dr. Lane received her Ph.D. in economics and Master's in statistics from the University of Missouri.

ANDREW MCCALLUM is a professor and director of the Center for Data Science, as well as the Information Extraction and Synthesis Laboratory, in the College of Information and Computer Science at

the University of Massachusetts Amherst. He has published over 250 papers in many areas of artificial intelligence, including natural language processing, machine learning, and reinforcement learning; his work has received over 45,000 citations. He obtained his Ph.D. from the University of Rochester in 1995 with Dana Ballard and a postdoctoral fellowship from Carnegie Mellon University with Tom Mitchell and Sebastian Thrun. In the early 2000s he was vice president of research and development at WhizBang Labs, a 170-person start-up company that used machine learning for information extraction from the Web. He is an Association for the Advancement of Artificial Intelligence Fellow, the recipient of the UMass Chancellor's Award for Research and Creative Activity, the UMass NSM Distinguished Research Award, the UMass Lilly Teaching Fellowship, and research awards from Google, IBM, Microsoft, and Yahoo. He was the general chair for the International Conference on Machine Learning 2012 and is the current president of the International Machine Learning Society, as well as member of the editorial board of the *Journal of Machine Learning Research*. For the past 10 years, McCallum has been active in research on statistical machine learning applied to text, especially information extraction, entity resolution, social network analysis, structured prediction, semi-supervised learning, and deep neural networks for knowledge representation.

RICHARD McCULLOUGH has a B.S. in chemistry from the University of Texas, Dallas and earned his M.A. and Ph.D. in chemistry at Johns Hopkins University. He did his postdoctoral fellowship at Columbia University. Since 2012, Dr. McCullough has been the vice provost for research, working with the president and provost to encourage, cultivate, and coordinate high-impact academic research across all of Harvard's schools and affiliated institutions. The Office of the Vice Provost for Research (VPR) has broad responsibility and oversight for the development, review, and implementation of strategies, planning, and policies related to the organization and execution of academic research across the entire university. Dr. McCullough leads a new office of Foundation and Corporate Development. He also assists in oversight of many of the interdisciplinary institutes, centers, and initiatives across Harvard. Under Vice Provost McCullough's leadership, the Office of the VPR is particularly focused on removing barriers to collaboration, whether in university policies or financial or administrative systems. Additionally, the vice provost for research works with the president and provost to foster and encourage entrepreneurship among undergraduates, graduate students, and faculty members. He also helps to lead the development of the new innovation campus. Richard McCullough is also a professor of materials science and engineering at Harvard and is a member of numerous professional societies and boards. Prior to being named vice provost for research at Harvard, Dr. McCullough was the vice president for research at Carnegie Mellon University in Pittsburgh, where he previously served as the dean of the Mellon College of Science and professor and head of the Department of Chemistry. Dr. McCullough has founded two companies: Plextronics, Inc. and Liquid X Printed Metals.

REBECCA NUGENT is a teaching professor in the Department of Statistics at Carnegie Mellon University (CMU) and has been teaching at CMU since she completed her Ph.D. in statistics from University of Washington in 2006. Prior to that, she received her B.A with majors in mathematics, statistics, and Spanish at Rice University and her M.S. in statistics at Stanford University. She recently was awarded top teaching honors with the ASA Waller Education Award; The William H. and Frances S. Ryan Award for Meritorious Teaching; and Statistician of the Year by the ASA Pittsburgh Chapter. Nugent's research interests lie in clustering, record linkage, educational data mining/psychometrics, public health, tech/innovation/entrepreneurship, and semantic organization.

LEE RAINIE is the director of internet, science, and technology research at Pew Research Center. Under his leadership, the Center has issued more than 500 reports based on its surveys that examine people's online activities and the internet's role in their lives. He also directs the Center's new initiative on the intersection of science and society. The American Sociological Association gave Dr. Rainie its award for "excellence in the reporting on social issues" in 2014 and described his work as the "most authoritative source of reliable data on the use and impact of the internet and mobile connectivity." Dr. Rainie is a co-

author of *Networked: The new social operating system* and five books about the future of the internet that are drawn from the Center's research. He gives several dozen speeches a year to government officials, media leaders, scholars and students, technology executives, librarians, and nonprofit groups about the changing media ecosystem. He is also regularly interviewed by major news organizations about technology trends. Prior to launching Pew Research Center's technology research, Dr. Rainie was managing editor of *U.S. News & World Report*. He is a graduate of Harvard University and has a Master's in political science from Long Island University.

ROB RUTENBAR received his Ph.D. from the University of Michigan in 1984 and then joined the faculty at Carnegie Mellon University (CMU). He spent 25 years in electrical and computer engineering at CMU, ultimately holding the Stephen J. Jatras (E'47) Chair. He was the founding director of the Center for Circuit and System Solutions (called "C2S2"), a large consortium of U.S. schools (e.g., CMU, Massachusetts Institute of Technology, Stanford, Berkeley, Caltech, Cornell, Columbia, Georgia Tech, University of California, Los Angeles) supported by the Defense Advanced Research Projects Agency and the U.S. semiconductor industry, focused on design problems at the end of Moore's Law scaling. In 2010 he moved to the University of Illinois, Urbana-Champaign, where he is Abel Bliss Professor and head of Computer Science. At Illinois, he pioneered the novel "CS + X" program, which combines a core computer science curriculum with a disciplinary "X" curriculum, leading to a Bachelor's degree in "X"; student pipelines for CS + anthropology, astronomy, chemistry, linguistics, are now under way, with several more CS + X degrees under design. His research has focused in three broad areas: tools for integrated circuit design, statistics of nanoscale chip designs, and custom architectures for machine learning and perception. In 1998 he founded Neolinear, Inc., to commercialize the first practical synthesis tools for non-digital ICs, and served as Neolinear's chief scientist until its acquisition by Cadence in 2004. In 2006 he founded Voci Technologies Inc. to commercialize enterprise-scale voice analytics. He has won numerous awards, including the IEEE CASS Industrial Pioneer Award and the Semiconductor Research Corporation Aristotle Award. His work has been featured in venues ranging from *Slashdot* to the *Economist* magazine.

KRISTIN M. TOLLE is the director of the Data Science Initiative in Microsoft Research Outreach, Redmond, Washington. Since joining Microsoft in 2000, Dr. Tolle has acquired numerous patents and worked for several product teams including the Natural Language Group, Visual Studio, and the Microsoft Office Excel Team. Since joining Microsoft Research's outreach program in 2006, she has run several major initiatives from biomedical computing and environmental science to more traditional computer and information science programs around natural user interactions and data curation. She also directed the development of the Microsoft Translator Hub and the Environmental Science Services Toolkit. Dr. Tolle is an editor, along with Tony Hey and Stewart Tansley, of one of the earliest books on data science, *The Fourth Paradigm: Data Intensive Scientific Discovery*. Her current focus is developing an outreach program to engage with academics on data science in general and more specifically around using data to create meaningful and useful user experiences across devices platforms. Prior to joining Microsoft, Tolle was an Oak Ridge Science and Engineering Research Fellow for the National Library of Medicine and a research associate at the University of Arizona Artificial Intelligence Lab managing the group on medical information retrieval and natural language processing. She earned her Ph.D. in management of information systems with a minor in computational linguistics. Dr. Tolle's present research interests include global public health as related to climate change, mobile computing to enable field scientists and inform the public, sensors used to gather ecological and environmental data, and integration and interoperability of large heterogeneous environmental data sources. She collaborates with several major research groups in Microsoft Research including eScience, computational science laboratory, computational ecology and environmental science, and the sensing and energy research group.

TALITHIA WILLIAMS takes sophisticated numerical concepts and makes them understandable and relatable to everyone. As illustrated in her popular TedTalk "Own Your Body's Data," she demystifies

the mathematical process in amusing and insightful ways, using statistics as a way of seeing the world in a new light and transforming our future through the bold new possibilities inherent in the science, technology, mathematics, and engineering (STEM) fields. As an associate professor of mathematics at Harvey Mudd College, she has made it her life's work to get people—students, parents, educators, and community members—more excited about the possibilities inherent in a STEM education. In her present capacity as a faculty member, she exemplifies the role of teacher and scholar through outstanding research, with a passion for integrating and motivating the educational process with real-world statistical applications. Her educational background includes a Bachelor's degree in mathematics from Spelman College, Masters' degrees in both mathematics from Howard University and statistics from Rice University, and a Ph.D. in statistics from Rice University. Her professional experiences include research appointments at the Jet Propulsion Laboratory, the National Security Agency, and NASA. Dr. Williams develops statistical models which emphasize the spatial and temporal structure of data and has partnered with the World Health Organization in developing a cataract model used to predict the cataract surgical rate for countries in Africa. Through her research and work in the community at large, she is helping change the collective mindset regarding STEM in general and math in particular—rebranding the field of mathematics as anything but dry, technical, or male-dominated but instead a logical, productive career path that is crucial to the future of the country.

ANDREW ZIEFFLER is a senior lecturer and researcher in the Quantitative Methods in Education program within the Department of Educational Psychology at the University of Minnesota. He teaches undergraduate- and graduate-level courses in statistics and trains and supervises graduate students in statistics education. Prior to receiving his Ph.D., Dr. Zieffler taught mathematics and A.P. Statistics at ROCORI High School in Cold Spring, Minnesota. Dr. Zieffler's scholarship primarily focuses on statistics education. He has authored/co-authored several papers and book chapters related to statistics education and has been a co-principal investigator on many National Science Foundation-funded statistics education research projects. Additionally, he has co-authored two textbooks that serve as an introduction to modern statistical and computational methods for students in the educational and behavioral sciences. Dr. Zieffler currently serves as co-editor of the journal *Technology Innovations in Statistics Education* and as a member of the Research Advisory Board for the Consortium for the Advancement of Undergraduate Statistics Education.

B

Meetings and Presentations

FIRST COMMITTEE MEETING
Washington, D.C.
December 12–13, 2016

Lessons from current data science programs and future directions
Rebecca Nugent, Carnegie Mellon University
Rob Rutenbar, University of Illinois, Urbana-Champaign
David Culler, University of California, Berkeley
William Yslas Velez, University of Arizona
Duncan Temple Lang, University of California, Davis

Envisioning the field of data science and future directions and implications to society
David Donoho, Stanford University
Lee Rainie, Pew Research Center

Expanding diversity in data science—among student populations and in topic areas embraced by data science
Bhramar Mukherjee, University of Michigan
Deb Agarwal, Lawrence Berkeley National Laboratory
Andrew Zieffler, University of Minnesota

Questions that should be asked to envision the future of data science for undergraduates
Tom Ewing, Virginia Tech
Louis Gross, University of Tennessee, Knoxville
Chris Mentzel, Gordon and Betty Moore Foundation
Patrick Perry, New York University
John Abowd, U.S. Census Bureau

SECOND COMMITTEE MEETING
Webinar
April 25, 2017

Overview of the study
Michelle Schwalbe, National Academies of Sciences, Engineering, and Medicine
Alfred Hero, University of Michigan
Laura Haas, IBM Almaden Research Center
Louis Gross, University of Tennessee, Knoxville

Facilitated discussion
Andy Burnett, Knowinnovation

WORKSHOP
Washington, D.C.
May 2-3, 2017

Opening comments
Study Co-Chairs: Laura Haas, IBM, and Alfred Hero, III, University of Michigan

Comments from the National Science Foundation
Chaitan Baru, National Science Foundation

Overview of the workshop
Andy Burnett, Knowinnovation

Workshop Themes

Skills and knowledge for future data scientists
Rob Rutenbar, University of Illinois, Urbana-Champaign

Broadening participation in data science education
Julia Lane, New York University

Future delivery of data science education
Nick Horton, Amherst College

Table Discussions About Key Questions

Question exploration groups
Small breakout groups to discuss all three questions

Feedback from question groups
Present ideas and discuss questions with full group

Integrate ideas into three thematic areas
Form three groups aligned with the thematic questions or possible new questions

Feedback from question groups
Share the integrated ideas with the full group

Plenary discussion of feedback
Study Co-Chairs: Laura Haas, IBM, and Alfred Hero, III, University of Michigan

New questions and ideas which emerged overnight
Full group discussion led by Andy Burnett, Knowinnovation

Identify the most promising ideas and possible findings for the committee's interim report

Small table groups

Backcast the most promising ideas

Small table groups discuss what steps would have to be taken in order to implement the most promising ideas

Participants

Ani Adhikari—University of California, Berkeley
Stephanie August—National Science Foundation
Chaitan Baru—National Science Foundation
Quincy Brown—American Association for the Advancement of Science
Andy Burnett—Knowinnovation
Eva Campo—National Science Foundation
Linda Casola—National Academies of Sciences, Engineering, and Medicine
Alok Choudhary—Northwestern University
Catherine Cramer—New York Hall of Science
David Culler—University of California, Berkeley
Renee Dopplick—Association for Computing Machinery
Jon Eisenberg—National Academies of Sciences, Engineering, and Medicine
E. Thomas Ewing—Virginia Tech
William Finzer—Concord Consortium
Greg Goins—North Carolina A&T State University
Louis Gross—University of Tennessee, Knoxville
Laura Haas—IBM
Alfred Hero—University of Michigan
Nicholas Horton—Amherst College
Charles Isbell—Georgia Tech
Ryan Jones—Middle Tennessee State University
Nandini Kannan—National Science Foundation
Danny Kaplan—Macalester College
Brian Kotz—Montgomery College
Jay Labov—National Academies of Sciences, Engineering, and Medicine
Julia Lane—New York University
Sharon Lane-Getaz—St. Olaf College
Jeff Leek—Johns Hopkins University
Andrew McCallum—University of Massachusetts Amherst
Richard McCullough—Harvard University
Mary Kehoe Moynihan—Cape Cod Community College
Bhramar Mukherjee—University of Michigan
Claudia Neuhauser—University of Minnesota
Deborah Nolan—University of California, Berkeley
Rebecca Nugent—Carnegie Mellon University
Dennis Pearl—Pennsylvania State University
Gabriel Perez-Giz—National Science Foundation
Lee Rainie—Pew Research Center
Patrick Riley—Google
Andee Rubin—TERC
Rob Rutenbar—University of Illinois, Urbana-Champaign
Michelle Schwalbe—National Academies of Sciences, Engineering, and Medicine

Amy Stephens—National Academies of Sciences, Engineering, and Medicine
Victoria Stodden—University of Illinois, Urbana-Champaign
Kristin Tolle—Microsoft
Ron Wasserstein—American Statistical Association
Ben Wender—National Academies of Sciences, Engineering, and Medicine
Elena Zheleva—National Science Foundation
Andrew Zieffler—University of Minnesota