Roundtable
HIGHLIGHTS
June 2018

Roundtable on Data Science Postsecondary Education

Meeting #6 - March 23, 2018

The sixth Roundtable on Data Science Postsecondary Education was held on March 23, 2018, at the Hotel Shattuck Plaza in Berkeley, California. Stakeholders from data science education programs, government agencies, professional societies, foundations, and industry convened to discuss how data science can be used to help understand and improve reproducibility of scientific research and to highlight several courses and training offerings in reproducible data science. This Roundtable Highlights summarizes the presentations and discussions that took place during the meeting. The opinions presented are those of the individual participants and do not necessarily reflect the views of the National Academies or the sponsors. Watch meeting videos or download presentations at **nas.edu/DSERT**.

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

[1] Replicability "refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected," whereas reproducibility "refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" (NSF, 2015, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, May).

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

Victoria Stodden, University of Illinois, Urbana-Champaign

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

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

Applying this expectation for practicing transparent science to the notion of *teaching* data science, Stodden commented that effective data science curricula would include training in computational methods and tools as well as in theory and computational techniques. She suggested thinking about both tool and curricula development in terms of the data life cycle (i.e., acquire, clean, use/reuse, publish, preserve/destroy). Kolaczyk asked how faculty could modify their curricula based on the data life cycle. Stodden responded that using the data life cycle as a guide highlights where knowledge gaps exist and where new courses can be added in programs to address such gaps. Deb Agarwal, Lawrence Berkeley National Laboratory, elaborated that students should be trained to understand data as approximations of facts by considering how data sets are generated, examining uncertainty and underlying errors, and evaluating how errors could affect algorithms. In response to a question from Timothy Gardner of Riffyn, Stodden said that the audience for her data science curriculum includes any student who wants to work in any aspect of the data life cycle—from departments of statistics, computer science, information, and library science, for example. She added that classes with the word "data" in the title are so popular that it would be useful to begin to refine the curricula appropriately for students who plan to enter industry or to continue in academia, respectively.

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

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

be generalizable to other contexts. Stodden agreed that generalizability is the end goal; she added that computational reproducibility is a subset of this issue, and transparency is a key part of the process.

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

Fernando Perez, University of California, Berkeley

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

The course uses the Jupyter Notebook, which allows the combination of text, code, and mathematical language in a single document accessible via a web browser. The students' work environments include a personal installation on each of their devices, using Anaconda for dependency management, and an installation hosted by the Department of Statistics; this mimics real-world settings in which data science practitioners may have to use remote servers or the cloud. The course and its materials can be accessed via GitHub, which provides a natural workflow for content management. Working with this software allows students to develop habits for good "computational hygiene," according to Perez. Students learn how to automate tasks with the Make tool, using tutorials developed by Software Carpentry, as well as how to do continuous integration with validation using Travis. Students attempt to replicate real-world research in their first hands-on projects; then, they develop a practical "playbook" for reproducible research and use Binder to share a live, executable version of their completed work. For their final course project, Perez explained, students find their own data and conduct and document their analyses using the playbook they created earlier in the course.

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

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

ence courses. Deborah Nolan, University of California, Berkeley, replied that students benefit more when courses are co-taught and the content is integrated because they can learn more about how to use computational skills in the context of data analysis.

REPRODUCIBLE MACHINE LEARNING—THE TEAM DATA SCIENCE PROCESS

Buck Woody, Microsoft Research and AI

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

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

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

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

OPEN DISCUSSION

Incentive and Reward Structures

Nicholas Horton, Amherst College, wondered how incentive structures in academia could be modified to encourage faculty to teach data science courses and to develop data science tools. Gardner described the fundamental difference between incentive structures in academia (e.g., publishing results and earn-

ing grants) and in industry (e.g., creating products that work for a customer). The incentive structure in industry better drives collaboration and innovation, Gardner explained. Agarwal said that the reward system in academia has not yet emphasized team-based investigation over individually driven investigation. She suggested working to prioritize team-based investigation in the culture and in the practice of science by giving appropriate credit to everyone who participates in any part of the research and analysis process. She also noted that the *people* involved in reproducible scientific research are just as important as the *mechanisms* of reproducibility because myriad decisions get made over the course of an analysis. She encouraged recognizing and rewarding people for the contributions they make, specifically in the middle of their careers. Green and McKeown added that curriculum development is not incentivized or rewarded as much as it could be at many public research universities. Green suggested encouraging faculty to develop contracts with their deans that formalize reward structures for course development as well as educating students early about the importance of team-oriented and goal-oriented approaches so as to begin to change the culture.

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

Reproducible Research

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

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

TRAINING AS A PATHWAY TO IMPROVE REPRODUCIBILITY

Tracy Teal, The Carpentries

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

the majority of them are eager to learn new data analytics skills.

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

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

These four goals can be achieved, according to Teal, by revising existing courses, hosting short courses and workshops, developing massive open online courses, or offering just-in-time training. Teal suggested that educational resources be built collaboratively, reused, and continually updated. Based on her experience, these materials would be most useful if made discoverable and open, and they are most effective when aligned with the needs and goals of the individual learners. She emphasized the importance of changing the culture around who works with data by creating a community of practice in which people help one another to learn.

Teal explained that the Carpentries is a "non-profit organization that develops curriculum, trains instructors, and teaches workshops on the skills and perspectives to work effectively and reproducibly with software and data." The Carpentries offers 2-day active learning workshops led by trained instructors. In these workshops, students are given formative feedback, have opportunities to collaborate with one another and with instructors, and develop skills applicable to data workflow and software development best practices. Teal recounted that the Carpentries hopes that students recognize the possibilities for data-driven discovery, develop confidence in using computational and data science skills, and will continue learning upon completion of a workshop. The Carpentries has hosted more than 1,300 workshops on 7 continents with 1,300 volunteer instructors for 35,000 learners. Teal noted that the Carpentries conducts both short- and long-term pre- and post-workshop surveys to gauge participant interest and success. Responses to these surveys indicate that, overall, students have improved their attitudes toward reproducible research and use the skills they have acquired on a regular basis.

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

PERSPECTIVES ON ENHANCING RIGOR AND REPRODUCIBILITY IN BIOMEDICAL RESEARCH THROUGH TRAINING

Alison Gammie, National Institute of General Medical Sciences

Gammie explained that because issues of scientific rigor and transparency (especially in the field of biomedical research) are being discussed more frequently in the popular press, representatives of Congress are now paying more attention to the notion of reproducibility. Surveys conducted by Nature revealed a number of causes that contribute to irreproducible results, the top three of which are selective reporting, pressure to publish, and low statistical power or poor analysis. The biomedical research incentive structure, in particular, represents an underlying systemic factor that can affect reproducibility. Academic researchers are under constant pressure to secure funding, innovate, publish, and gain tenure. These issues are complicated by the fact that only 10 percent of National Institute of Health (NIH)-funded principal investigators receive more than 40 percent of NIH funding, according to Gammie. Through its program called Maximizing Investigators' Research

Awards, the National Institute of General Medical Sciences (NIGMS) works to better distribute these funds among researchers and enhance scientific discovery.

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

Gammie explained that increased training is one pathway to enhance reproducibility. NIGMS developed a clearinghouse for new training resources that contribute to rigor and transparency, as well as multiple funding announcements to develop training modules in enhanced reproducibility or local courses in experimental design and analysis. NIH also offers a predoctoral training grant program to ensure that rigor and transparency are threaded throughout the graduate curriculum and reinforced in the laboratory. The principal investigator and program faculty on these grants are required to have a record of doing rigorous and transparent science and to submit a specific plan for how the instruction will enhance reproducibility. Such programs will help trainees develop the technical, operational, and professional skills needed to enter the biomedical research workforce. Gammie emphasized the need for academic institutions to recognize training and mentoring activities in tenure and promotion packages and to decrease the pressures on principal investigators that negatively impact the research culture. Gammie concluded by reiterating that rigor and transparency, responsible and safe conduct of research, and diversity and inclusion are integral to excellence in training.

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

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

Timothy Gardner, Riffyn

Gardner commented that it is important to bridge the gap between industry and academia. Riffyn's mission is to help scientists deliver reusable data and trustworthy results. He emphasized the value of focusing on the fundamental *causes* of irreproducibility rather than the *symptoms*, and he explained that researchers are failing to harness reproducibility lessons learned more than 50 years ago and apply them to scientific research. More than $420 billion is spent on research and development globally each year, and, if even 25 percent of the results are irreproducible, $105 billion will be lost each year. He continued that researchers hope to achieve a world of science in which published results can be built upon, but this goal has not yet been realized, primarily because researchers spend 80 percent of their time cleaning and organizing data instead of learning from them. He categorized data-related challenges in research and development in terms of data quality, access, integration, interpretation, and system flexibility. Gardner agreed with Agarwal that data are only approximations, not facts. Clean data begins with quality experiments, and it is important to teach principles, develop tools, and build a culture of quality in research and development throughout foundational undergraduate curricula.

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

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

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

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

SMALL GROUP DISCUSSIONS AND CONCLUDING CONVERSATIONS

Roundtable participants divided into two groups to discuss key questions that emerged earlier in the day. On behalf of his group, Green summarized discussions in response to the following questions: ***How could reproducibility be taught within a particular course or program? What are the implications of***

resource limitations, class size, teaching structure, and other incentives? Should reproducibility be taught on its own or integrated into other topics? Green described his group's discussion of how to balance programming with statistics education in data science courses. In reflecting on Perez's presentation, Green posited that perhaps only one-fifth of the curriculum would focus on programming, while the remainder would focus on issues such as testing and validation. He added, however, that such a curricular decision would vary by audience and that several approaches such as the following exist:

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

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

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

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

DISCLAIMER: This meeting recap was prepared by the National Academies of Sciences, Engineering, and Medicine as an informal record of issues that were discussed during the Roundtable on Data Science Postsecondary Education at its sixth meeting on March 23, 2018. Any views expressed in this publication are those of the participants and do not necessarily reflect the views of the sponsors or the National Academies.

ROUNDTABLE MEMBERS PRESENT: Eric Kolaczyk, Boston University, Co-Chair; Kathleen McKeown, Columbia University, Co-Chair; Deb Agarwal, Lawrence Berkeley National Laboratory; Alok Choudhary, Northwestern University; James Frew, University of California, Santa Barbara; Mark Green, University of California, Los Angeles; Alfred Hero, University of Michigan; Nicholas Horton (via webcast), Amherst College; Bill Howe, University of Washington; Charles Isbell, Georgia Institute of Technology; Mark Krzysko, U.S. Department of Defense; Brandeis Marshall, Spelman College; Chris Mentzel, Gordon and Betty Moore Foundation; Nina Mishra, Amazon; Deborah Nolan, University of California, Berkeley; Peter Norvig, Google; Antonio Ortega, University of Southern California; Victoria Stodden, University of Illinois, Urbana-Champaign; Duncan Temple Lang, University of California, Davis; Mark Tygert, Facebook Artificial Intelligence Research; Jeffrey Ullman, Stanford University; and Jessica Utts, University of California, Irvine.

GUESTS PRESENT: Alison Gammie, National Institute of General Medical Sciences; Timothy Gardner, Riffyn; Charlotte Mazel-Cabasse, University of California, Berkeley; Mary Beth McLendon, Accel.AI; Laura Montoya, Accel.AI; Fernando Perez, University of California, Berkeley; Josh Quan, University of California, Berkeley; Anthony Suen, University of California, Berkeley; Tracy Teal, The Carpentries; Tom Treynor, Treynor Consulting; Eric Van Dusen, University of California, Berkeley; Adam Wolisz, Berlin University of Technology; and Buck Woody, Microsoft Research and AI.

STAFF PRESENT: Linda Casola (via webcast), Janki Patel, Michelle Schwalbe (via webcast), and Ben Wender.

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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