Data Science: Opportunities to Transform Chemical Sciences and Engineering Proceedings of a Workshop - in Brief

US National Academy of Sciences, 2018

Linda Casola and Ellen Mantus, Rapporteurs; Chemical Science Roundtable; Board on Chemical Sciences and Technology; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

New technologies and approaches are generating large, diverse data sets, and data science offers the tools that are needed to interrogate, analyze, and manage these data sets. Biology, material sciences, and other fields have embraced data science tools and used them to gain insights into, for example, gene–environment interactions, molecular mechanisms of disease, and implications of material characteristics on performance. Chemical sciences and engineering have also used data science tools to, for example, monitor and control chemical processes, predict activity depending on chemical structures or properties, and inform business and research decisions. However, data science applications in the chemical sciences and engineering communities have been relatively limited, and many opportunities for advancing the fields have gone unexplored. Accordingly, the Chemical Sciences Roundtable of the National Academies of Sciences, Engineering, and Medicine, in collaboration with the National Academies Board on Mathematical Sciences and Analytics, held a workshop to explore opportunities to use data science to transform chemical sciences and engineering on February 27–28, 2018, in Washington, DC. Stakeholders from academia, government, and industry convened to discuss the challenges and opportunities to integrate data science into chemical sciences and engineering practice and data science training for the future chemical sciences and engineering workforce. This Proceedings of a Workshop-in Brief summarizes the presentations and discussions that took place during the workshop. The workshop videos and presentations are available online.[1]

What Is Data Science?

Andrew Ferguson, an assistant professor at the University of Illinois at Urbana-Champaign, opened the workshop by providing a broad overview of data science and its applications. He described data science as knowledge discovery from often large and complex data sets and noted that it is interdisciplinary by nature, encompassing statistics, computer science, applied mathematics, and domain-specific tools. He explained that many data sets being used could qualify as big data, meaning they are difficult to store, analyze,

[1] *See http://nas-sites.org/csr/data-science-opportunities-to-transform-chemical-sciences-engineering.*

or share with traditional tools and approaches. He depicted data science as the "fourth pillar of science" (after theory, experimentation, and simulation) and noted that its success depends on the integration of data science paradigms and tools with domain-specific knowledge and expertise.

Data science can help to solve chemical science problems, for example, in predicting material properties or behaviors at new scales, Ferguson said. He described the typical workflow for data-driven science as an iterative process: (1) identify a database; (2) eliminate redundancies, reduce large uncertainties, and describe or annotate the data; and (3) use data science methods to develop and validate a data-driven model that can examine correlations, evaluate inferences, and gain new understanding of the system.

Ferguson provided an overview of various easy-to-use software packages and libraries that can introduce domain scientists to data science.[2] He highlighted a few mature applications in the design of such products as crystalline alloys and organic photovoltaics (e.g., AFLOW,[3] The Materials Project,[4] Harvard Clean Energy Project[5]) and proteins (e.g., Rosetta Commons,[6] Desmond[7]). Emerging data science applications in the chemical sciences community include those in nanoparticle packing and assembly (Damasceno et al. 2012) and data-driven materials design (Miskin et al. 2016). Industrial applications also exist-a new public-private partnership titled Accelerating Therapeutics for Opportunities in Medicine[8] aims to accelerate drug discovery by using high-throughput computational screening of drug candidates. That project emphasizes the importance of negative results as they can be vital for training data science models.

Ferguson noted that the hard materials community has already established mature databases and servers, while the soft materials community is experiencing more difficulty, owing largely to its more diverse systems and tools.[9] Additional challenges for the soft materials community include the diversity of length and time scales; the diversity of sequences and configurations, which complicates sampling, screening, and design; the out-of-equilibrium or active- material systems; and biological interactions. He emphasized the need for chemical sciences curricula to include training in data science, statistics, machine learning, and computer programming so that graduates will be better prepared to overcome the challenges in the future. Changes are already happening in policy and training workshops; for example, the National Science Foundation (NSF) hosted a workshop[10] in 2017 on advancing and accelerating materials innovation that included a large component on data science.

Ferguson concluded his presentation by highlighting some strengths of the chemical sciences community as they relate to the emergence of data science: (1) statistical thermodynamics and molecular simulation integrate well with data science and machine learning paradigms; (2) the chemical science community is set to inspire new computational tools and algorithms; (3) entropy can be exploited as a driving force to stabilize desired molecular configurations; and (4) new fields

---

2 *See http://nas-sites.org/csr/files/2018/04/2.-Ferguson.pdf#page=13.*

3 *See http://www.aflowlib.org.*

4 *See https://materialsproject.org.*

5 *See http://cleanenergy.molecularspace.org.*

6 *See https://www.rosettacommons.org.*

7 *See https://www.schrodinger.com/desmond.*

8 *See https://atomscience.org.*

9 *Hard materials are those that are resistant to deformation from applied forces, such as diamonds, ceramics, and many metals. Soft materials are those that can be easily deformed, such as liquids, gels, foams, and biological materials.*

10 *See https://mrsec.uchicago.edu/mat_summit.*

within the chemical sciences, such as bioengineering, systems biology, and-omics, already integrate data science into their disciplines.

A participant posited that the data problems in the chemical industry are uniquely difficult. Because domain knowledge is critical, companies are interested in hiring strong chemists and chemical engineers whose data science skills can be expanded, rather than hiring pure data scientists. Another participant highlighted kinetics, catalysis, and molecular design as areas in which the chemical community could push forward with data science applications, considering the complexity of how certain molecules interact with materials. Ferguson agreed that data science tools will be useful in countless areas, especially those with large, complex data sets. In response to another participant's concern about appropriate training, Ferguson said that data science is context dependent and that easy-to-use tools need to be developed to capture errors without requiring the user to have extensive expertise in data science. He emphasized that domain expertise is crucial in the successful application of data science tools to avoid a situation in which a limited model is built and applied too broadly. A participant emphasized the importance of capturing the experimental parameters associated with the data in the data science workflow described earlier. Ferguson agreed that capturing the metadata and ensuring that all data are searchable and unified is essential, although each community will likely have a unique protocol for such processes.

What can Data Science do for me? Three answers from real projects

Patrick Riley, a principal engineer at Google, continued the discussion by providing three examples of the successful application of data science tools and methods in the chemical sciences domain. First, Google collaborated with the California Institute of Technology to conduct high-throughput searches for new metal oxides. The team used an inkjet printer with metal salts dissolved in gel and "printed" the gel onto glass plates that were then baked. Crystalline oxides formed and were examined by using hyperspectral microscopy. As the data were interpreted, anomalies were detected and ultimately corrected. Riley explained that this work succeeded because of the iterative approach taken and the maintenance of the metadata associated with the experiments. He emphasized that high-throughput experiments often decrease costs to run replicates and provide more opportunity to be innovative. He added that people often think that the goal of high-throughput experiments is to gather massive data sets, but one clear benefit is greater confidence that can be obtained in the experiments conducted.

Riley's second example was a collaboration of Google with TriAlpha Energy[11] to help optimize experimental design in the context of nuclear fusion. Riley explained that there are myriad choices that must be made to operate the machines, which leads to an optimization problem. Although the underlying physical processes might be understood by scientists, getting the system to work is extremely complicated and involves extensive trial and error. So, Google built an algorithm that could assist TriAlpha Energy in exploring its complex parameter space and ultimately in helping to identify new operating regimes for the machines. A key aspect of this work was that a human was kept in the loop.

The third example Riley presented used the QM9 data set[12] that contains estimates of various physical properties of small organic molecules to develop a machine learning model that could ultimately be used for virtual screening. The approach relied on a type of neural network that involves messaging and updating

[11] *TriAlpha Energy is now called TAE Technologies. See https://tae.com.*

[12] *See http://quantum-machine.org.*

functions at each step rather than just "learning" at the last step. Riley noted that this is an example of a more advanced machine learning model that surpasses other methods.

Riley concluded his presentation with four take-away messages: (1) attention to design, verification, and validation can prevent problems and allow the development of a more useful model; (2) working with big data might require a change to the current way of thinking; (3) machine learning might not be the best way to solve all data problems; and (4) the production of reliable data remains challenging, even with the emergence of new methods.

In response to a question from an audience participant, Riley noted that when building a machine learning model, one needs to think carefully about how broadly the model can be applied given the training set used to develop it. In response to another participant's question about potential areas for collaboration between industry and academia, Riley noted that Google's goal is to inspire more people to adopt and apply computational techniques and tools, thus increasing the rate of scientific discovery. Google is interested in areas for collaboration that can use complex data and demonstrate the compelling power of data science methods.

García-Muñoz concluded his presentation by reiterating that students need to have the opportunity to develop better skills in data analysis, basic programming, and algorithmic thinking before they enter the workforce. Equally important, students still need to learn fundamental chemical engineering principles. He emphasized the gap between academia and industry that urgently needs to be closed, perhaps with greater opportunities for collaboration. He also noted that data science can be thought of as a business case; agencies have the difficult task of determining whether it is profitable in the long term to implement data science tools.

Untapped data resources from a catalysis science perspective

Thomas Bligaard, a senior staff scientist at the SLAC National Acceleratory Laboratory and the deputy director of theory at the SUNCAT Center for Interface Science and Catalysis, explained that the mission of catalysis science is to integrate theory with material synthesis, characterization, and testing. The goals are to improve understanding of interface phenomena and move toward science-based design of solid catalysts that can help provide sustainable processes

for fuels and chemicals. He noted that although data science is often used to manage big data, it can also be used to optimize having only sparse data, which is often the case in catalysis science.

Simulation costs have fallen drastically over the past 30 years. Bligaard noted that quantum computing has the potential to accelerate that trend further. The tremendous advances in computational power mean that the value of data generated via simulations will decrease as a function of time; thus, it is important to think about what data should be put into the databases and why. He emphasized the importance of storing descriptions of the scientific process and the relevant metadata to reproduce and build on results. In response to a question from the audience, he encouraged the chemical sciences community to embrace the tools that are becoming available to promote the reproducibility of science.

Bligaard described failed experiments as a source of untapped data resources; for example, using a machine learning model with all experimental data-successes and failures-can lead to the discovery of new materials (Raccuglia et al. 2016). He noted that machine learning will not replace human intelligence; rather, it will allow all data to be factored into decision-making and result in better directions for next steps. Bligaard described another example of untapped data resources in the context of sample characterization. A compressed sensing method was

used to decrease the electron flux through a sample to reduce heating or perturbation so that sample stability could be maintained (Kovarik et al. 2016). Integrating synthesis, characterization, and testing data also reveals an untapped data resource, according to Bligaard. Using cluster techniques with machine learning tools can reveal correlations among composition, structure, and various optimal properties (Suram et al. 2015). His final example of an untapped data resource related to using data science to calibrate scientific equipment to increase performance (McIntire et al. 2016).

He explained that data science tools reduce the computational work needed to explore challenging design problems and to identify optimal materials. Bligaard concluded that he hoped that databases containing experimental data on synthesis, characterization, and testing will be more integrated and useful in the near future so that catalyst discovery can be accelerated.

Panel Discussion

How Can Data Science Be Used to Gain Insights or New Knowledge from Data Resources?

Riley began the discussion by noting that data science techniques and tools have been used for simulations in areas where there is a good understanding of the underlying physics but that these techniques and tools need to be applied more broadly. He stated that data science should help in the development of models that can provide more mechanistic insights but added that the techniques will have to advance before those benefits can be realized. Ferguson noted that increased insight will come from increased data sharing and that the chemical sciences community needs to facilitate the cultural shift that encourages scientists to publish their codes, provide more details about their methods, and make their databases more accessible. Bligaard added that machine learning models can help to reduce complexity by identifying which signals are real and which can be ignored and thus improve physical and chemical models that describe the desired system.

García-Muñoz emphasized that it is important to define the problem before using data science methods to extract information, especially given the high cost associated with generating various data. He suggested using hybrid models in which prior knowledge can be embedded and encouraged people to recognize that they know more about systems than they realize and to integrate that information into data science approaches. Riley countered that the value of many data science methods is partly to allow the data to speak for themselves as much as possible and reveal to researchers when they are wrong. García-Muñoz agreed that the data can reveal errors but reiterated the importance of researchers relying on their fundamental knowledge to know when to trust the data. Ferguson noted that although it is important to ask the right questions and to rely on domain expertise, it is equally valuable to do exploratory analyses to discover unanticipated patterns, which can spark new inquiries. Riley added that producing large data sets offers a way to ask questions that might otherwise seem statistically unsound. García-Muñoz, however, cautioned against assuming that big data are necessarily equivalent to informative data.

Moderator Bruce Garrett, director of the Chemical Sciences, Geosciences, and Biosciences Division in the Office of Basic Energy Sciences at the US Department of Energy (DOE), pressed the panelists and workshop participants to identify opportunities to apply data science in areas where it is not currently being used. A participant noted that an algorithm that reveals underlying correlations in thousands of primary data streams that are not evident from engineering models could provide critical information for chemical manufacturing. Riley noted that such analytical techniques already exist and explained that similar problems are already being tackled in the technology industry to detect anomalies. García-Muñoz agreed that sophisticated analytical techniques exist, but the barrier to implementation might be the lack of a business case.

A participant noted that encoding physical constraints and knowledge into machine learning methods presents some challenges and asked the panelists to discuss those difficulties. Bligaard said that it is surprisingly simple to combine machine learning models with normal search algorithms and emphasized the need to resist the "business as usual" culture and rethink the way simple methods are used. Riley noted that encoding invariances into machine learning models is an active area of research. Ferguson added that the physical sciences can inform the data science so that new algorithms that better embed physical constraints can be developed. García-Muñoz reiterated that there is a gap between academic research and industry practice; for example, techniques to address complex problems exist, but they have not been adopted by industry.

What Are the Barriers Preventing Broader Adoption of Data Science in the Chemical Sciences and Engineering?

Riley stated that a substantial barrier to the adoption of data science is a lack of exposure to data science methods and tools in the various domains. He added that although academic institutions are starting to develop data science curricula, there is still much progress to be made. In response to a question about the difficulty of recruiting people with the right combination of domain expertise and data science training to work on projects, Riley proposed hiring a data scientist and then teaching that person materials science or hiring a materials scientist and teaching that person data science, with preference for the former. Alternatively, García-Muñoz proposed hiring both a data scientist and a domain expert to work as a team. Another participant suggested that programs similar to the now defunct NSF Integrative Graduate Education and Research Traineeship[13] are needed so that doctoral students have an opportunity to work on domain-specific problems using data science skills. Ferguson suggested that more women and minorities also need to be engaged in this work.

García-Muñoz stated that another barrier relates to structuring and storing data so that they can be analyzed effectively. Riley noted that it is important for organizations to invest in ways to monitor the production of data continually to help to overcome that problem. García-Muñoz described another barrier as the lack of standards for capturing data on chemical reactions, which prevents effective data mining.

An audience participant asked about the roles that data science, machine learning, and cognitive computing play in chemical synthesis and other chemical processes. García-Muñoz noted that reaction kinetics are relatively well understood and struggled to see the applicability of machine learning methods in this area. Another participant, however, suggested that reaction data might be well suited to study linear free energy relationships.

Another participant asked what tools are needed to integrate predictive modeling into synthesis, characterization, and testing processes. Bligaard said that it will take years before everything is integrated into a closed loop system. He noted that integrating theory and synthesis is more difficult than integrating theory and characterization or theory and testing. Progress will be reached sooner if everyone in the chemical sciences community focuses on these integration challenges. Riley added that the process will be a gradual evolution rather than a rapid progression.

Creating Data Science Programs in Academia

Thomas Ewing, a professor at Virginia Tech, opened his presentation with a discussion of data science as a mechanism for storytelling in different domains. As an example, he explained that in the early 20th century people began to analyze the war efforts from a statistical perspective and used data visualizations to illuminate various findings. He emphasized the value

[13] *See http://www.igert.org.*

of linking data science with the humanities-the humanities bring the "flexibility of scalability, an appreciation of change, and a deep understanding of complexity to the interpretation of human behavior, ideas, and interactions as represented in numbers, statistics, and data."

Ewing then discussed considerations for creating data science programs in academia and suggested that the first question to answer is what are the skills, capacities, and abilities that you want the students to learn? He shared the Association of American Colleges and Universities' "Top 10"[14] employer wishes for new hires and highlighted two: "the ability to think clearly about complex problems" and "the ability to understand numbers and statistics." He showed similar lists from other organizations and stated that it is essential for academic institutions to provide students with the appropriate skills to build a data science talent pipeline.

Ewing continued that curriculum designers also need to consider how new courses can integrate with general education requirements; at what level the courses should be offered; what connections exist with the natural and social sciences, arts, and humanities; how to incorporate the core concepts of mathematics, statistics, and computer science; and how best to equip students to communicate data. He also noted that it is important that data science courses teach students to think about what data mean as opposed to just learning how to manipulate data. He acknowledged that curriculum development is challenging, especially given the different definitions of data science and the ongoing discussions about the essential components of data science education. His concluding advice for those engaged in designing new data science programs was to discuss desirable student skills, coordinate existing courses, identify motivated faculty members, outline the sequence of courses that meets degree requirements, find partners within and beyond the institution for capstone projects, and balance creativity with practicality.

A participant from industry noted the value of ensuring that all people understand how to use data in their particular disciplines and job functions. Ewing agreed and added that the ability to make good judgments and decisions about data is a key component of emerging undergraduate data science curricula. He advocated for more experiential learning, internships, and project-based learning in a variety of academic disciplines.

A participant asked how Ewing's course incorporates mathematics, which often prevents students from enrolling in data science courses. Ewing noted that his course is not meant to be a substitute for a mathematics course; instead, he hopes students will consider how mathematics matters in context. A participant commented that it is possible

to teach better data-driven decision-making without having to teach equations. Another participant added that it is important for data science tools to be accessible to a variety of people-complex equations could be a divisive factor in data science courses meant to target a broad audience.

In response to a participant's request for more advice on adopting a data science program, Ewing suggested consulting the National Academies' report *Data Science for Undergraduates: Opportunities and Options* (NASEM 2018). He reiterated the value of employer input in curriculum development. Competing demands on campuses can prove to be problematic, so Ewing emphasized the need for faculty to present a clear argument to administrators about how such programs could be valuable for the universities.

[14] *See https://www.aacu.org/leap/students/employers-top-ten.*

Breakout Group Discussions

Participants spent the afternoon of the first day of the workshop in small groups discussing data analytic methods, data quality assessment, and data management and translation in the chemical sciences and chemical engineering communities. Highlights of the discussions were provided by the chairs of the breakout sessions the following day.

Data Analytics

Bryce Meredig, the co-founder and chief science officer of Citrine Informatics, provided an overview of his group's discussion of recent applications of data analytics. For example, there have been major advances in measurement capabilities in the environmental sciences that have led to the generation of vast amounts of data, and data analytic methods have been valuable for analyzing, integrating, and interpreting the information. Other applications include the Chemistry Dashboard[15] of the US Environmental Protection Agency, which links chemical structures, physical-chemistry properties, and toxicity data and the work of the Oak Ridge National Laboratory to centralize and standardize large, fast-growing volumes of image data. Meredig also mentioned a few specific methods used by the chemical sciences community, including retrosynthetic analysis tools and generative models that use machine learning to design new molecules.

Meredig described several challenges to using various data analytic methods. He noted that machine learning algorithms can learn human biases and that machine learning methods can often overwhelm a person's ability to process the results. To be successful, some methods present the results as a short list of options, but designing that short list can be difficult. Meredig echoed the need to train chemists and chemical engineers to take advantage of data analytic tools and methods and stressed the need to standardize data and possibly define ontologies. He pointed out that federated (interconnected) databases and other similar infrastructure efforts are beginning to emerge, and these could help resolve accessibility issues.

Although the data science community often touts machine learning or artificial intelligence, Meredig stated that important advances are occurring in the development of tools that simply help people visually interpret data and tools that help to combine, interconnect, or transfer data from multiple sites. He also described exciting research to incorporate more physics or physical constraints into models to reduce the black box nature of algorithms. Meredig emphasized the need to develop data analytic methods to predict side reactions, for example, in catalysis and to address issues of data quality and reproducibility. He stressed that data analytic methods are only as good as the data themselves, and he reiterated the value of negative results in model building. He encouraged increased investment in data preparation and infrastructure, as systems have not kept pace with some methods, and concluded his overview by discussing the need for incentives or mechanisms to promote greater data sharing.

Data Quality Assessment

Dave Higdon, a professor in the Social Decision Analytics Laboratory at the Biocomplexity Institute at Virginia Tech, provided an overview of his group's discussion. He stated that data can be physical measurements, expert judgments, or predictions from computational models and described data quality as a property of the data that depends on the application. Thus, there is not one generic assessment that can be conducted to validate the data for every type of decision that the data could inform. The data are part of a larger process in which a decision is typically going to be made, and the acceptable level of accuracy or uncertainty associated with the data will depend on the magnitude or significance of that decision.

[15] *See https://www.epa.gov/chemical-research/chemistry-dashboard.*

Group member Karl Mueller, the chief science and technology officer for the Physical and Computational Sciences Directorate at the Pacific Northwest National Laboratory, explained that data quality assessment plays different roles for those in academia, industry, and government but ultimately focuses on whether the data can be trusted for use in decision-making. He noted the importance of understanding the risk associated with decision-making and when one needs to repeat experiments. Group member Miguel García-Garibay, a professor at the University of California, Los Angeles, described the Cambridge Crystallographic Data Centre[16] as an exemplary example of a high-quality data collection that is used by many in the field. He reiterated that the need for high-quality data depends on the application and the expense. Higdon then shared a number of best practices for data quality assessment in the chemical sciences and engineering communities. They included instituting and describing instrument calibration procedures, using standard materials and samples, recording metadata and history associated with the measurements, recording data provenance and data ownership, describing the domain for which data will be useful, building relationships with new data sources, and replicating experiments.

Higdon continued that the chemical sciences community could benefit from thinking more about standards for metadata and ways to interconnect databases to improve data accessibility and could leverage advances in computer science and statistics. Group member Pablo Rolandi, a process development director at Amgen, emphasized the importance of metadata by describing a personal experience in which he was trying to model the performance of a chemical process reactor and found the metadata particularly valuable for better understanding the industrial process behavior. Higdon concluded his overview by sharing a list of challenges for data quality assessment, including issues surrounding automation and training; intellectual property; legacy culture, data, and equipment; data integration and storage; timelines for decision-making; and a lack of a common language among statisticians, data scientists, chemists, and chemical engineers.

A participant commended the group for highlighting the importance of metadata and data provenance and asked about how best to start recording and compiling the data. Group co-leader Carlos Gonzalez, the chief of the Chemical Sciences Division of the National Institute of Standards and Technology, noted that approaches will vary by discipline. He suggested considering ontologies that can evolve. A participant referenced the Allotrope Foundation[17] as a good example of an evolving ontology that the chemical sciences community could use as a model. A participant noted that "trusting data" is a confusing concept; he advocated coupling data science tools with experimental data sets. As a final comment, a participant emphasized that timelines for making decisions in research and development are much longer than those in manufacturing.

Data Management and Translation

John Gregoire, a principal investigator and research thrust coordinator at the Joint Center for Artificial Photosynthesis, provided an overview of his group's discussion. He noted that a goal of data management and translation is to make the data useful for other problems, and this presents new challenges. Because data scarcity is a persistent issue in chemistry research, the community is motivated to develop methods to integrate disparate data streams.

Gregoire explained that careful data management is essential in experimental chemistry and added that automation has the potential to facilitate the standardization of data storage and management. He stated that improved metadata management is the most urgent need for the chemical sciences

[16] *See https://www.ccdc.cam.ac.uk.*

[17] *See https://www.allotrope.org.*

community and continued that data integration can be challenging in academia, where instruments and methods are constantly evolving. As such, data management has to be flexible enough to accommodate new types of data.

He also referenced the Cambridge Crystallographic Data Centre as an example of a successful research database.

He noted that because this model cannot be expanded to accommodate the entire chemistry and engineering communities, the communities should foster local data management, connect relevant databases, and build algorithms to draw from different sources. He described NSF's and DOE's work in algorithm development and suggested that their programs be maintained and that the community be trained to use these algorithms. Gregoire also highlighted Citrine Informatics, which has been successful in developing tools that can pull data from multiple sources. He emphasized that algorithms and models used to manage the data are as valuable as the data themselves and need to be managed accordingly.

Regarding data sharing, Gregoire suggested that scientists in academia might be even more protective of their data than members of industry, particularly in the case of failed experiments. Strategies to incentivize data sharing could include conducting peer reviews of how researchers execute their data plans; developing a scheme to quantify the impact of a data point or a data set; and encouraging the publication of data sets in peer-reviewed journals. The group also discussed the potential benefits of publishing partial data sets and developing an algorithm that could match these data sets among laboratories as a novel validation approach. Gregoire mentioned the development of a Web-based data repository similar to GitHub and noted the need to adopt a common language or ontology that can evolve. Group co-leader Jens Hummelshøj, a materials design and discovery program manager at the Toyota Research Institute, reiterated the importance of the community's role in establishing scientific standards.

A participant highlighted the importance of data monitoring to resolve data inconsistencies. Gonzalez stated that computational tools to improve and streamline the data monitoring process would be helpful. A participant suggested designing databases so the data can be used by different communities but acknowledged the challenge of communicating information to different audiences and noted that computational tools to translate the information into formats familiar to respective communities might help. Gonzalez added that data visualization is another method to help bridge the communication gap between domain scientists and data scientists.

Panel Presentations on Training A Future Generation of Chemists and Chemical Engineers

The workshop concluded with a panel discussion on training the future generation. Panelists were first asked to provide examples of educational gaps for chemists and chemical engineers and to discuss ways in which academic programming and other initiatives could address these gaps.

Leo Chiang, Associate Technology Director, The Dow Chemical Company

Chiang posed the following question: What training is necessary to create a workforce that can sustain our scientific and technologic success? He posited a training model in which 70 percent is on-the-job experience, 20 percent is mentoring and coaching, and 10 percent is classroom courses and assignments. He suggested that undergraduate students would benefit from more training in statistics, programming, and data-driven decision-making; master's students would benefit from developing a more balanced breadth and depth in technical topics; and doctoral students would benefit from a greater incorporation of data science into chemical engineering and chemistry research. Chiang emphasized that the chemical industry has to compete with other industries for top talent, especially for those who have data science experience.

Lloyd Colegrove, Director of Data Sciences & Fundamental Problem Solving, The Dow Chemical Company

Colegrove referenced *Fault Detection and Diagnosis in Industrial Systems* (Chiang et al. 2001) as an important text on integrating data science into process monitoring for members of industry. He described *Competing on Analytics: The New Science of Winning* (Davenport and Harris 2007) and *Analytics at Work: Smarter Decisions, Better Results* (Davenport et al. 2010) as texts that highlight the advantages of data-driven decision-making in industry. Although the texts emphasize the interest in integrating data science into industrial practice, Colegrove stated that it is within only the past 5 years that academia has become more receptive to training requests from industry. He advocated for more relevant statistics instruction through practical problem solving at the undergraduate level and noted Dow partnerships with Northwestern University and The Pennsylvania State University to help better train its workforce. He added that if science is to continue to progress in the United States, it will be necessary for *all* people to develop better data acuity. He remarked that NSF, DOE, and the US Department of Defense could play important roles in achieving that goal.

Rebecca Nugent, Director of Undergraduate Studies and Professor, Carnegie Mellon University

Nugent explained that the decentralized infrastructure at Carnegie Mellon University enables collaboration. Carnegie Mellon's Department of Statistics and Data Science hosts or co-hosts five undergraduate majors, three master's degree programs, and four or five doctoral programs, and the university sponsors various data competitions. All the academic programs integrate interdisciplinary research in the coursework, require few textbooks, and introduce a multitude of methods. She suggested that curricula and training be scalable, modern, diverse, and interdisciplinary, and she cautioned against overemphasizing coding at the expense of conceptual understanding in introductory courses. She added that any program that incorporates data science elements should stress such concepts as reproducibility, replicability, and responsibility. Nugent stated that data literacy is essential for all people, not just science and engineering students. She highlighted the importance of diversity and accessibility in data science and suggested conducting outreach in middle and high schools to begin the path toward widespread data literacy.

Al Hero, John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and R. Jamison and Betty Williams Professor of Engineering, University of Michigan

Hero suggested that the data life cycle—planning, collecting, analyzing, visualizing, confirming, and communicating— be taught as a principle concept in data science courses. He mentioned the National Academies report *Data Science for Undergraduates: Opportunities and Options* (NASEM 2018) and distinguished two possible outcomes of data science education: data literacy and data acumen. He defined data literacy as overcoming data and math phobias and developing basic data visualization skills and an understanding of the data life cycle so that one can identify obvious problems. He defined *data acumen* as being able to ask the right questions to inform an experiment; finding and collecting the right data; mastering data curation, modeling, computation, and inference; and shepherding data through the full life cycle. He added that data science continues to evolve rapidly and noted the introduction of a new journal titled *SIAM Journal on Mathematics of Data Science*.[18]

[18] *See https://www.siam.org.*

Hero said that the University of Michigan offers a BS in data science jointly administered by the Electrical Engineering and Computer Science Department and the Statistics Department. Students complete the program with a rigorous foundation in computer science, statistics, and mathematics and an understanding of data science methods and real-world experience from a capstone course. The university offers a graduate certificate in data science: a nine-credit program that trains users in modeling, technology, and practice. The university now also offers an MS in data science that best serves students who plan to enter the workforce instead of a doctoral program. Hero said that specific strategies to instill data science acumen in chemistry and chemical engineering students are still in development.

Melissa Cragin, Executive Director, Midwest Big Data Hub

Cragin noted that NSF established four regional big data innovation hubs to address scientific and societal challenges, spur economic development, and foster a national big data ecosystem among academia, industry, nonprofit and civic groups, and government. The mission of the Midwest Big Data Hub[19] is to cultivate cross-sector communities and research networks to transfer and share cutting-edge work, increase access and use of data and technology, and build capacity in data science. The Midwest Big Data Hub focuses on issues related to agriculture; food, energy, and water; smart and resilient communities; health and biomedicine; unmanned systems; and materials and manufacturing.

The Midwest Big Data Hub facilitates workshops, participates in national and international industry activities, and provides awards through federal funding programs, such as the Spoke awards. The awards provide funding to a particular sector to build a community around data science applications and develop data tools and techniques to accelerate application (e.g., the Midwest Big Data Spoke for Integrative Materials Design[20]). The Midwest Big Data Hub also participates in regional and national meetings to encourage data science education and workforce development, with a specific interest in broadening participation, and offers data science training opportunities. In addition to these many initiatives, the Midwest Big Data Hub aims to expand data science access in small (non-R1) colleges and universities.

Richard Braatz, Edwin R. Gilliland Professor of Chemical Engineering, Massachusetts Institute of Technology

Braatz noted that data science education in chemical engineering curricula ranges from a few lectures, to a multi- week unit within a course, to a statistics and probability course taught by faculty from the statistics or mathematics department, to an engineering statistics course taught by an engineer, to an engineering statistics course taught specifically by a member of the chemical engineering faculty. At the Massachusetts Institute of Technology, all graduate students take two courses that cover advanced statistical methods. He suggested that chemical engineering educators visit the Computer Aids for Chemical Engineering[21] website for a helpful collection of statistics instructional materials. He emphasized the value of having chemical engineering students work on real engineering problems, and he referenced The University of Texas at Austin as a leader in this space. He also suggested that educators focus on developing students' skills in chemical engineering–relevant areas, such as common distribution and hypothesis testing, design of experiments, parameter estimation with confidence regions or intervals, statistical process control charts, and software tools and applications to real data.

[19] *See http://midwestbigdatahub.org.*

[20] *See https://www.nsf.gov/awardsearch/showAward?AWD_ID=1636910&HistoricalAwards=false.*

[21] *See https://cache.org.*

Panel Discussion on Training A Future generation of Chemists and Chemical Engineers

Data Science Curriculum

Chiang asked whether data science education needed to be domain specific or whether it could take a more generalized approach. Colegrove stated that data science training does not have to be domain specific and that data science coursework could be integrated into current curricula. Braatz stressed that all engineers need to learn how to analyze data, particularly in their chosen discipline. Hero stated that students can gain exposure to data science the domain curriculum but will still need to develop skills to think about problems outside of their own disciplines. Nugent said that a mixture of chemical-engineering dependent material and exposure to different kinds of problems and tools is the best way for students to learn to think more critically and creatively about data. She suggested that departments review their curricula to determine whether new data science content can be substituted for outdated material.

Meredig said that Citrine Informatics has offered machine learning training to employees in various companies and has started to enter curricula conversations with academic institutions. He noted that summer schools and boot camps are options, but neither addresses the dearth of statistics and linear algebra in curricula. According to Meredig, a better option might be to insert data science modules into current curricula. Nugent noted that although statistics plays a role in most disciplines, it is unrealistic to expect everyone to become experts in statistics. She suggested that departments instead work together to develop interdisciplinary project-based courses in which people from different disciplines learn to collaborate and communicate with each other. Hero said that Iowa State University hosts a pre- freshman accelerator course in data science and statistics, and the University of Michigan offers a summer school experience for high school sophomores and juniors to learn about data science concepts through music, sports, and art activities. Hero and Nugent emphasized that many data science education opportunities exist beyond those in the traditional curriculum, and Braatz stated that a multipronged effort will be needed because different approaches will be necessary at the various universities and companies.

A participant said that it is possible to find space in any curriculum for data science instruction, for example, by eliminating unnecessary repetition of a straightforward concept or skill that spans multiple courses. This participant noted the value of consulting resources from other academic institutions that have successfully integrated data science into their curricula. Braatz and Nugent highlighted the need to change the culture by providing incentives for faculty to innovate in the classroom. Colegrove reiterated the urgent need to train undergraduates in a way that prepares them for the current and future demands of industry. Another participant asked how industry might help academia to meet these needs. Hero responded that industry could offer more partnerships, investments, student internships, and speaker sessions, in addition to offering advice on curricular changes. A participant suggested the need to focus on supporting continuing education so that employees can keep pace with emerging technologies.

Data Science Certification and Accreditation

A participant asked about certification for data scientists. Hero said that there is not currently a national data science organization to certify competency but that professional certificates in data science give some indication of training. Nugent said that “certification” in data science comes in many forms; for example, Johns Hopkins University offers a massive open online course in data science that provides a certificate on completion of a series of 10

courses and receipt of a small fee. She explained that such programs are most often used by professionals seeking new skills.

Another participant asked how the Accreditation Board for Engineering and Technology (ABET) and the American Chemical Society (ACS) view data science in chemical engineering curricula. Hero hoped that ABET will become more active in determining the right balance of domain knowledge and data science to best prepare graduates for a variety of careers in the future. Braatz said that ABET's perspective on engineering curricula has changed during the past 20 years and now has a greater emphasis on identifying goals, collecting data to determine whether these goals have been met, and using data to modify curricula accordingly. A participant said that data science or statistical training is only encouraged, not required, by the ACS to obtain a bachelor's degree in chemistry. The ACS is in the process of revising the guidelines for curricular approval to be more outcomes based, so input from the chemical sciences community would be helpful. Colegrove added that the presence of degrees in applied statistics indicates progress.

References

- Chiang, L.H., E.L. Russell, and R. Braatz. 2001. Fault Detection and Diagnosis in Industrial Systems. London, England: Springer- Verlag.
- Damasceno, P.F., M. Engel, and S.C. Glotzer. 2012. Predictive self-assembly of polyhedra into complex structures. Science 337(6093):453-457.
- Davenport, T.H., and J.G. Harris. 2007. Competing on Analytics: The New Science of Winning. Boston, MA: Harvard Business School Press.
- Davenport, T.H., J.G. Harris, and R. Morison. 2010. Analytics at Work: Smarter Decisions, Better Results. Boston, MA: Harvard Business Review Press.
- Kovarik, L., A. Stevens, A. Liyu, and N.D. Browning. 2016. Implementing an accurate and rapid sampling approach for low- dose atomic resolution STEM imaging. Applied Physics Letters 109:164102.
- McIntire, M., D. Ratner, and S. Ermon. 2016. Sparse Gaussian processes for Bayesian Optimization. Pp. 517-526 in Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence.
- Miskin, M.Z., G. Khaira, J.J. de Pablo, and H.M. Jaeger. 2016. Turning statistical physics models into materials design engines.
- Proceedings of the National Academy of Sciences 113(1):34-39.
- NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press.
- Raccuglia, P., K.C. Elbert, P.D. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, and A.J. Norquist. 2016.
- Machine-learning-assisted materials discovery using failed experiments. Nature 533(7601):73-76.
- Suram, S.K., J.A. Haber, J. Jin, and J.M. Gregoire. 2015. Generating information-rich high-throughput experimental materials genomes using functional clustering via multitree genetic programming and information theory. ACS Combinatorial Science 17(4):224-233.

Disclaimer

This Proceedings of a Workshop—in Brief was prepared by Linda Casola and Ellen Mantus as a factual summary of what occurred at the workshop. The committee's role was limited to planning the workshop. The statements made are those of the rapporteurs or individual workshop participants and do not necessarily represent the views of all workshop participants, the planning committee, the Chemical Sciences Roundtable, the Board on Mathematical Sciences and Analytics, or the National Academies. To ensure that this Proceedings of a Workshop—in Brief meets institutional standards for quality and objectivity, it was reviewed in draft form by Andrew Ferguson, University of Illinois at Urbana-Champaign; John Gregoire, California Institute of Technology; Nicholas Horton, Amherst College; and David Sholl, Georgia Institute of Technology. The review comments and draft manuscript remain confidential to protect the integrity of the process.

Planning committee members were Michelle Chang, University of California, Berkeley; Leo Chiang, The Dow Chemical Company; Bruce Garrett, DOE; Carlos Gonzalez, National Institute of Standards and Technology; John Gregoire, California Institute of Technology; and Angela Wilson, NSF.

About the Chemical Sciences Roundtable

The Chemical Sciences Roundtable provides a neutral forum to advance understanding of issues in the chemical sciences and technologies that affect government, industry, academic, national laboratory, and nonprofit sectors, and the interactions among them; and to furnish a vehicle for education, the exchange of information, and the discussion of issues and trends that affect the chemical sciences. The Roundtable accomplishes its objectives by holding annual meetings with its members and by organizing workshops on relevant and important topics for which published proceedings are made broadly available throughout the chemical sciences community.

Chemical Sciences Roundtable members are Jennifer Sinclair Curtis (*Co-Chair*), University of California, Davis; Mark E. Jones (*Co-Chair*), The Dow Chemical Company; Tina Bahadori, US Environmental Protection Agency; Brian Baynes, MODO Global Technologies; Michael R. Berman, Air Force Office of Scientific Research; Donna G. Blackmond, The Scripps Research Institute; Emilio Bunel, Argonne National Laboratory; Allison Campbell, Pacific Northwest National Laboratory; Richard R. Cavanagh, National Institute of Standards and Technology; Michelle Chang, University of California, Berkeley; Richard Dickinson, NSF; Miles Fabian, National Institutes of Health; Maria Flytzani- Stephanopoulos, Tufts University; Michael J. Fuller, Chevron Energy Technology Company; Miguel Garcia-Garibay, University of California, Los Angeles; Bruce Garrett, DOE; Franz Geiger, Northwestern University; Carlos Gonzalez, National Institute of Standards and Technology; Malika Jeffries-El, Boston University; Jack Kaye, National Aeronautics and Space Administration; Mary Kirchhoff, American Chemical Society; Robert E. Maleczka, Jr., Michigan State University; David Myers, GCP Applied Technologies; Ashutosh Rao, US Food and Drug Administration; Leah Rubin Shen, Legislative Assistant/Office of Senator Chris Coons (D-DE); Angela Wilson, NSF; and Jake Yeston, American Association for the Advancement of Science.

This activity was supported by DOE under Grant No. DE-FG02-07ER15872, the National Institutes of Health under Contract No. HHSN26300024, and NSF under Grant No. CHE-1546732. 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.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science: Opportunities to Transform Chemical Sciences and Engineering: Proceedings of a Workshop—in Brief*. Washington, DC: The National Academies Press. doi: http://doi. org/10.17226/25191.