Big data meets materials science: Training the future generation

By Elizabeth Dickey and Greer Arthur

Capitalizing on the promise of "big data" will require materials scientists who are trained in data informatics. Several universities are answering the call.

"Big data" is making big changes to all fields of science and engineering and revolutionizing the way researchers work and interact. The data revolution in materials and ceramics research has been driven principally by two major developments: multi-billion-dollar investments in scientific characterization instrumentation at federal laboratories[1] and universities[2] during the past two decades; and advances in high-throughput computational materials discovery.[3,4] Further, real-time sensing coupled with robust data analytics has transformed product development and manufacturing. This area has become a target for investment by several large manufacturing companies[5] and has since been referred to as the Industrial Internet of Things (IIOT).

Big data is characterized by the "three Vs"—volume, velocity, and variety—and materials research is seeing huge growth along each of these facets. As an example, the Advanced Photon Source at Argonne National Laboratory can generate more than one terabyte of data per day from some beamlines, which is expected to increase to hundreds of terabytes or even petabytes per day in 10 years.

With the diversity of sources from which materials scientists can now harvest big data, leading to increases in data variety, challenges come during the analysis phase, when something meaningful must be deduced from multiple large datasets. In all of these cases, adding statistical data sciences to the other three paradigms of materials science—empirical, theoretical, and computational—likely will prove significant in successfully handling and utilizing big data.[6]

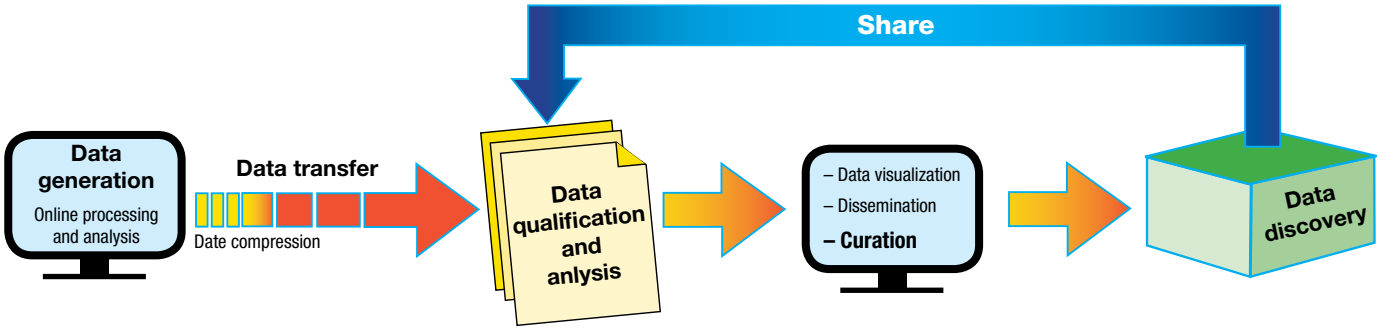


Figure 1. An infrastructure for data informatics can improve every stage of materials research and development, from initial collection of data and transmission, through its quantification and analysis, to its final dissemination and curation.

To keep pace, we must develop an infrastructure for data informatics that we can implement at every stage of materials research and development—from initial collection of data and transmission, through its quantification and analysis, to its final dissemination and curation (Figure 1).[7,8] Further, we must use data informatics as a sifting tool to reveal information hidden within big data to facilitate new discoveries in materials science.

Recently, researchers articulated a demand for "big-deep-smart data" within the Materials Genome Initiative (MGI),[9] which motivates integrating materials physics with advanced statistical and computational approaches to data analysis and machine learning. The ultimate goal is to create new knowledge from the flood of materials data. Further, unprecedented opportunities occur for us to integrate information from multiple characterization and modeling datasets. This perhaps is best exemplified in the field of materials structure determination. We can approach deducing the crystallographic structure of a new material from a variety of experimental and computational techniques (e.g., X-ray diffraction, neutron scattering, electron diffraction, and imaging on the experimental front and density functional theory and molecular dynamics on the computational front). Each has a unique sensitivity to various aspects and scales of the material's structure. Despite the fact that information from each technique is complementary, it is rare that data are fully assimilated into a coherent model of atomic structure, and analyses of individual datasets sometimes result in numerous—even conflicting—descriptions of the same material.[10]

Training the next generation of scientists and engineers

We will ultimately guide future materials science knowledge while accelerating material discovery and design by unifying the fourth paradigm of statistical data sciences with the empirical, theoretical, and computational paradigms of materials science. Importantly, this shift in scientific methodology demands a response in the way we train students to ensure the new investment in big data-driven science is prosperous and sustainable.

During the late 20th century, Integrated Computational Materials Engineering (ICME) emerged as an answer to the growing importance of modeling and simulation in materials science and engineering. Today, we integrate computational materials science into most undergraduate and graduate curricula.[11] Likewise, the educational landscape for what has been termed "Data Enabled Science and Engineering" (DESE) is just beginning to emerge as an interdisciplinary collaboration among physical scientists, statisticians, applied mathematicians, and computer scientists. To support this development, the National Science Foundation (NSF) is investing in transformative graduate research models that transcend the traditional boundaries between disciplines, while simultaneously equipping graduate students with the skills needed to become competent, professional leaders in a broad range of career paths. In 2014, NSF launched its Research Traineeship (NRT) program, which awards projects that use bold and innovative approaches to graduate education while focusing on cross-disciplinary teaching and learning within a nationally important topic. Currently, DESE is one such high-priority research area.

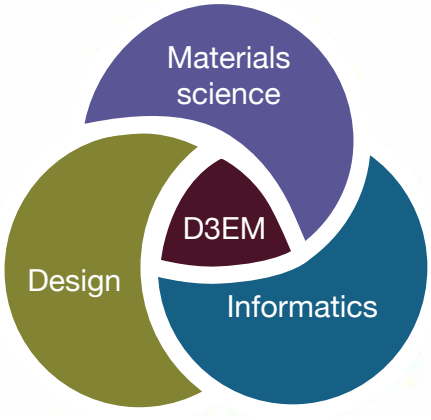
Creation and utilization of resources with key stakeholders within private sectors, nongovernmental organizations, national laboratories, governmental agencies, field stations, teaching and learning centers, informal science centers, and academic partners is a key component of NRT programs. Broadening participation of under-served populations of students in STEM disciplines is a second component of the NRT programs. To date, NSF has awarded $76.7 million to 40 NRT projects, 14 of which have an emphasis on DESE.

Several NRT-DESE programs will have a direct impact on ceramics research. At Texas A&M University (College Station, Texas), faculty in the Colleges of Science and Engineering have teamed with faculty from the Center for Teaching Excellence to develop an interdisciplinary Data-Enabled Discovery and Development of Energy Materials (D^3EM) graduate program. By combining expertise from materials science, informatics, engineering systems design, and STEM graduate education, the program aims to bridge the gap between materials science and data science by merging new accelerated mate-

Report identifies major hurdles and provides recommendations for materials data infrastructure

The Minerals, Metals & Materials Society (Pittsburgh, Pa.) recently organized a study on behalf of the National Science Foundation that aims to inform, guide, and motivate development of a materials data infrastructure to overcome challenges in harnessing data so that it can effectively serve the materials community. The rich 2017 report from that study, *Building a Materials Data Infrstructure: Opening New Pathways to Discovery and Innovation in Science and Engineering*, is available at www.tms.org/mdistudy.

Big data meets materials science: Training the future generation



Credit: D³EM

Figure 2. Texas A&M University's interdisciplinary Data-Enabled Discovery and Development of Energy Materials (D³EM) graduate program merges new accelerated materials development with engineering systems design, utilizing skills from materials science, design, and informatics.

rials development with the discipline of engineering systems design, utilizing skills from materials science, design, and informatics (Figure 2). The program pursues the hypothesis that by learning from theories, concepts, and methods of various disciplines, it can create a cultural tool to facilitate this interdisciplinary transition.

Established in 2015, D³EM consists of a technical interdisciplinary component and a strong professional development curriculum. The technical component is implemented at Texas A&M University as an Interdisciplinary Graduate Certificate, which consists of cross-disciplinary and interdisciplinary components. After a first year of grounding in their respective disciplines, D³EM students enroll in three courses: materials science, materials informatics, and advanced product design. The following semester, students enroll in a materials design studio, a project-based, capstone design-inspired course developed by D³EM's principal investigator, Raymundo Arróyave. In the design studio, students use ideas from informatics, design, decision theory, optimization, and search to solve realistic materials design, development, or discovery problems that ideally are connected to the student's research. The first D³EM student cohort has gone through the certificate curriculum and successfully completed materials design studio projects.

Debra Fowler, D³EM coprincipal investigator, designed the professional development component of the D³EM certificate. It consists of

- A learning community, where students explore various ideas related to ethics, interdisciplinarity, collaboration, conflict resolution, etc.;
- A writing community centered around the Texas A&M Department of Education and Human Development's Promoting Outstanding Writing for Excellence in Research (POWER) writing program, lead by a POWER-certified consultant; and
- A coffee session facilitated by D³EM-affiliated faculty, where various aspects of academic- and industry-based research are discussed in an informal setting.

These activities are complemented by a comprehensive optional internship program in which students acquire experience in industry and apply interdisciplinary technical skills that they acquired during the certificate program.

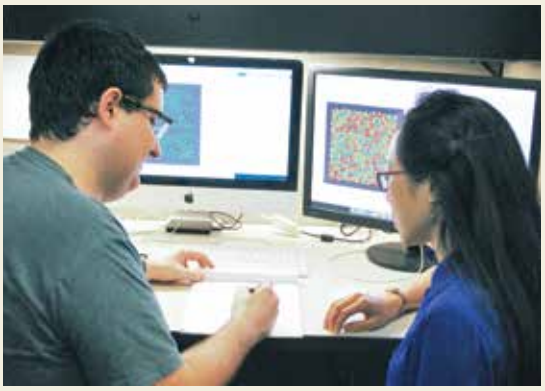
D³EM recently established a partnership with the U.S. Air Force Research Laboratory to double the size of the program through creation of an AFRL-Minority Leadership Program extension that will train underrepresented groups in this revolutionary method of materials development. The Texas A&M Office of Graduate and Professional Studies and the College of Engineering Academic and Student Affairs Office make the D³EM expansion possible with additional support. D3EM currently is seeking other avenues to expand the reach of the program. For more information, visit https://d3em.tamu.edu.

Credit: NCSU SEAS

Figure 3. North Carolina State University's Data-Enabled Science and Engineering of Atomic Structure (SEAS) graduate training program assembles measurement scientists, computational materials scientists, applied mathematicians, and statisticians to address atomic structure determination.

More recently, North Carolina State University (Raleigh, N.C.) launched another NRT-DESE program in the fall

Quantifying order and disorder in ferroelectric materials



The Center for Dielectrics and Piezoelectrics (CDP) is an NSF Industry/University Cooperative Research Center (I/UCRC) that aims to provide international leadership and train next-generation scientists in the fundamental science and engineering that underpin dielectric and piezoelectric materials. CDP supports industries based on capacitor and piezoelectric materials and devices through the development of new materials, processing strategies, electrical testing, and nanoscale characterization and modeling methodologies.

Several research projects within CDP focus on topics of relevance to DESE, most notably those associated with local structure determination via aberration-corrected transmission electron microscopy, electron diffractometry, and X-ray diffractometry. SEAS-NRT fellows Matthew Cabral (Department of Material Science and Engineering) and Jocelyn Chi (Department of Statistics) use spatial statistical analysis of aberration-corrected STEM images to quantify local variations in chemistry and atomic positions in complex dielectrics, such as relaxor ferroelectrics. Their goal is to understand the origins of very high electromechanical responses of this class of ceramics. The link to CDP provides a natural mechanism to translate statistical concepts and methods to industry research and development. For more information, visit the CDP website at www.cdp.ncsu.edu.

of 2016. Here, the Data-Enabled Science and Engineering of Atomic Structure (SEAS) NRT program addresses the demand for a new generation of interdisciplinary, data-driven scientists who can apply advanced statistical and mathematical methods to atomic structure data generated from cutting-edge scattering and imaging experiments as well as high-throughput atomistic computation. This interdisciplinary graduate training program assembles measurement scientists, computational materials scientists, applied mathematicians, and statisticians, who together address research and educational challenges associated with atomic structure determination (Figure 3).

The SEAS NRT program at N.C. State developed a comprehensive graduate curriculum that aims to

- Produce scientists and engineers who can respond to opportunities and challenges resulting from state-of-the-art experimental, computational, and statistical tools that produce and manage big data;
- Bridge knowledge gaps across disciplines;
- Foster collaborative interactions;
- Develop fluency across disciplines and in public communication of scientific ideas;
- Promote diversity;
- Develop leaders with strong professional identities; and
- Build a network of professional colleagues and mentors.

As well as addressing research challenges posed by big data, N.C. State's SEAS NRT program recognized underrepresentation of minorities in STEM disciplines

Bayesian inference meets materials science

With the complexities of big data comes the necessity to develop more powerful statistical methods. In diffractometry, the leading method for studying material phases, classical statistics has dominated the analysis. New methods are arriving, although they are not yet broadly adopted. For example, if we were to ask a group of experimental ceramists if they have heard of "Bayesian," we likely would receive only a handful of positive responses. Although advanced statistical methods, such as Bayesian statistics, have been readily incorporated in computational materials science, most experimental research utilizes classical statistics.

Bayesian statistics models uncertainty in a way fundamentally different from classical statistics, such as linear regression, or what we commonly refer to as the "frequentist" viewpoint. The frequentist treats an event's probability as its relative frequency in a large number of trials. This is useful when we sample data from large populations (e.g., drug trials). However, not all materials problems are conducive to this perspective—if we have a single sample from a process, we may want to know something specific about that sample, as opposed to considering future samples. Hence, a parameter value's confidence interval describing that specific sample does not have much meaning. In contrast, Bayesian statistics treats hypotheses and solutions as finite probabilities (i.e., the strengths of models and hypotheses). Given available data, we calculate the probability of certain solutions, such as the probability that a sample has a monoclinic crystal structure, within which we can quantify specific locations of atoms.

Building from this statistical framework, researchers from the Materials Science and Engineering, Statistics, and Mathematics Departments at N.C. State University, in collaboration with researchers at Oak Ridge National Laboratory and the National Institute of Standards and Technology, have recently applied Bayesian statistical approaches to rigorous quantification of diffraction data. Using new programs in education and outreach similar to those described in this article, we can readily adopt such new methods for big data problems in materials research.

and formed a strategic partnership with North Carolina Central University (NCCU), a public and historically African-American university located in Durham, N.C. Prominent NCCU faculty in the physics and mathematics departments have research interests that align with the technical focus of the SEAS NRT program. Combined with prior or ongoing collaborations with N.C. State faculty, NCCU students can participate as trainees within the same environment. For more information about SEAS, visit https://research.mse.ncsu.edu/seas.

DESE graduate training efforts at institutions such as N.C. State and Texas A&M will increase the number of interdisciplinary scientists who are fluent in foundational principles of physical, statistical, and computer science disciplines and can become future leaders and innovators in data-intensive interdisciplinary research. Further, we anticipate these programs to establish models and best practices for this type of interdisciplinary graduate education, which other institutions can adopt.

The revolution of big data is upon us, and academic, national laboratory, and professional society communities are compelled to respond to the enormous opportunities and challenges that accompany the stream of data. NSF DESE traineeships are one critical component of this paradigm transition in graduate education.

About the authors

Elizabeth Dickey is professor and director of graduate programs in the Department of Materials and Engineering and director of the Center for Dielectrics and Piezoelectrics at North Carolina State University (Raleigh, N.C.). Greer Arthur is a postdoctoral research scholar specializing in molecular biology and immunology at the North Carolina State University.

References

[1]User facilities of the Office of Basic Energy Sciences: A national resource for scientific research, Argonne National Laboratory (2009) http://science.energy.gov/~/media/bes/suf/pdf/BES_Facilities.pdf.

[2]"Midsize facilities: Infrastructure for materials research," National Academy of Sciences, The National Academies Press (2005) http://www.nap.edu/catalog/11336.html.

[3]S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nat. Mater.*, **12**, 191–201 (2013) http://doi:10.1038/nmat3568.

[4]A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Mater.*, **1**, 011002 (2013) http://dx.doi.org/10.1063/1.4812323.

[5]https://sloanreview.mit.edu/case-study/ge-big-bet-on-data-and-analytics/.

[6]A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' in science in materials science," *APL Mater.*, **4**, 053208 (2016) doi: 10.1063/1.4946894

[7]J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, and B. Meredig, "Materials science with large-scale data and informatics: Unlocking new opportunities," *MRS Bull.*, **41**, 399–409 (2016) https:/www.cambridge.org/core/terms. https://doi.org/10.1557/mrs.2016.93.

[8]A. Dima, S. Bhaskarla, C. Becker, M. Brady, C. Campbell, P. Dessauw, R. Hanisch, U. Kattner, K. Kroenlein, M. Newrock, A. Peskin, R. Plante, S.-Y. Li, Pierre-Franc¸O. Rigodiat, G. Sousa Amaral, Z. Trautt, X. Schmitt, J. Warren, and S. Youssef, "Informatics infrastructure for the Materials Genome Initiative," *JOM*, **68** [8] 2053–64 (2016) doi: 10.1007/s11837-016-2000-4.

[9]S.V. Kalinin, B.G. Sumpter, and R.K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nat. Mater.*, **14**, 973–80 (2015) http://dx.doi.org/10.1038/NMAT4395.

[10]S.J.L. Billinge and I. Levin, "The problem with determining atomic structure at the nanoscale," *Science*, **316** [5824] 561–65 (2007) http://dx.doi.org/10.1126/science.1135080.

[11]K. Thornton and M. Asta, "Current status and outlook of computational materials science education in the U.S.," *Modell. Simul. Mater. Sci. Eng.*, **13** [2] R53–R69 (2005) http://dx.doi.org/10.1088/0965-0393/13/2/R01. ■