Big data are shaping the future of materials science

Ashley A. White

Big data mean different things to different people. In commerce, retailers extract trends from millions of consumer purchases to target advertising and increase profits. In health, Google and the US Centers for Disease Control and Prevention analyze vast amounts of search data to identify and curb potential flu outbreaks. In biology, genomic scientists use the human genome map to develop disease treatments. And in materials science, advances in data analysis have placed the field on the verge of a revolution in how researchers conduct their work, analyze properties and trends in their data, and even discover new materials.

The size of data sets that once would have boggled the mind is now almost commonplace. The entire book collection of the US Library of Congress could be stored in 15 terabytes, paling in comparison to the 1.2 zettabytes (1.2 billion terabytes) of data created by humankind in 2010. But big data are about much more than size.

Data scientists may disagree on the exact definition of big data, but they discuss data-related issues in terms of V's. The size, or volume, of the data is one component, but equally important are variety (the degree of complexity or heterogeneity in the data) and velocity (the speed of data access and processing). IBM recently coined a fourth V, veracity (inherent trustworthiness of the data), while others include viability or value among their V's.

In any field, data sets are considered "big" when they are large, complex, and difficult to process and analyze. Materials science data tend to be particularly heterogeneous in terms of their type and source compared with data encountered in other fields. "We can easily generate large sets of data, whether it's from experiments like those performed using the Advanced Photon Source here at Argonne or from large simulations," said Olle Heinonen of the Materials Science Division at Argonne National Laboratory in Illinois. "What matters more are your capabilities for processing that data and ultimately getting something useful out of it."

One of the first steps in processing large data sets is data reduction. Experiments at the Large Hadron Collider, for example, retain only a small fraction of 1% of the data they produce. Storing and analyzing any more than the hundreds of megabytes per second deemed most valuable become impractical with current technologies. It is up to sophisticated software to determine which data are most relevant.

The Spallation Neutron Source at Oak Ridge National Laboratory (ORNL) in Tennessee, a user facility that carries out hundreds of materials science experiments each year, is capable of creating hundreds of gigabytes of data in a single experiment. This rate is beyond what a materials scientist can effectively analyze with typical technologies. In addition to reducing these data to something manageable, fast and easy data access (the "velocity" component) is particularly critical to experimentalists. ORNL has made strides toward addressing these issues through ADARA—the Accelerating Data Acquisition, Reduction, and Analysis Collaboration project.

Thomas Proffen, director of the Neutron Data Analysis and Visualization Division at ORNL, likens the advances ADARA will enable to transporting water. Old technology was like filling a bucket and carrying it from one place to another. "It's the same way we used to store data on tapes or disks and take it from point A to point B," Proffen said. "Now we'll have a pipe that allows data to flow quickly and continuously." In this streaming mode, data reduction and analysis can happen *in situ*, during the course of the experiment. Having real-time access to experimental data from the neutron beam means scientists can make immediate decisions as to how to steer their experiments, resulting in less wasted time and better data.

Users at most get one or two days a year at the facility, so time is precious. "The benefit is more efficient science," said Galen Shipman, director of the Compute and Data Environment for Science at ORNL. "If your calibration takes tens of minutes to an hour per sample, you can easily use up your entire allocation for experimental setup. With the real-time feedback ADARA provides, a user can immediately see if they'll get better results if they rotate their crystal by a few degrees." Previously, a researcher might only realize the setup was suboptimal after taking the data home on a disk for analysis. ADARA's capabilities are so attractive that both ISIS, a neutron and muon source in the United Kingdom, and the European Spallation Source, planned for construction in Sweden, have sought advice from ORNL on adopting ADARA or implementing similar technologies.

Anatole von Lilienfeld, a computational scientist at Argonne National Laboratory and a chemistry professor at the University of Basel, works on the application of machine learning to atomistic simulation. According to von Lilienfeld, the most sophisticated way of analyzing large data sets is by using artificial intelligence to detect trends in the data, then to quantify those trends and use them as models that implicitly make use of all the data in the set. The resulting models can then be used to design better materials.

This type of approach is used by researchers like Gerbrand Ceder, leader of the Materials Project at the Massachusetts Institute of Technology; Stefano Curtarolo, director of the Center for Materials Genomics at Duke University; and

Ashley A. White, ashley.ann.white@gmail.com

Alán Aspuru-Guzik, leader of the Clean Energy Project at Harvard University. Aspuru-Guzik hopes to open the door to the rational and systematic design of future high-performance materials, like organic solar cells. Using IBM's World Community Grid, his group will compile and analyze their results in a reference database that will be available for public use. Curtarolo's current focus is topological insulators, which conduct electrical current on their surfaces while their interiors behave as insulators, a property useful in quantum computing. Rather than rely on experimental trial and error to find crystals that exhibit this behavior, Curtarolo's data-driven methods create a mathematical formulation that serves as a recipe for discovering topological insulators with the desired properties. Likewise, Ceder's work aims to accelerate the materials discovery and design process by helping materials researchers predict new materials with desired properties by computing the fundamental properties of all known inorganic compounds.

This predictive simulation work is not meant to replace experiment, but rather to focus it. The true scientific advances are expected to come from the intersection of computation, data, and experiment. Once candidate materials are identified, experimentalists can confirm the predicted properties by synthesizing the material in a laboratory. Additional iterations of simulation and experiment then fine-tune the formulation to fit a particular application. This close collaboration between computational experts and experimentalists, combined with big data, is the key idea behind the US government's Materials Genome Initiative, which celebrated its second anniversary this past June. The initiative's ultimate goal is to use this iterative process to significantly reduce the time and cost to bring new materials from the laboratory to the marketplace.

The US government is also investing in other areas relevant to data. In March 2012, President Obama announced $200 million in new investments to support a Big Data Research and Development Initiative. Its goals are to advance state-of-the-art technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data; to harness these technologies to accelerate the pace of discovery in science and engineering; and to expand and train the workforce needed to develop and use big data technologies.

Proffen believes workforce issues are non-trivial. "One of the biggest challenges is finding people who live at the intersection of experiment, data, and computation." Heinonen agrees. While he says collaborations between mathematicians and scientists are common at Argonne, he still acknowledges "a real gap in training" in interdisciplinary areas. Shipman predicts the rise of a new discipline. "Just as we've branded computational science as a discipline, next we'll have data science as a discipline." Several institutions, including Cornell, the California Institute of Technology, George Mason University, and Rensselaer Polytechnic Institute have started data science programs. Materials science departments looking to reform their curricula in light of the growing role of data are facing tough choices between integrating informatics into traditional courses or creating separate courses to teach informatics.

Another area of concern is data curation and standards, as data are often generated or collected without a strategy for storage, sharing, or reuse. In order to enable long-term use of data collected today, they must be stored in such a way that they are broadly accessible and interoperable across repositories. Rigorous metadata standards and practices will also ensure that the research community can fully understand and reproduce the data.

One organization working to address these issues is the Research Data Alliance, a new international community-based alliance that aims to better coordinate data infrastructure and activities, improving standards, policies, and technologies for data sharing. The group has received initial funding from the US and Australian governments, as well as the European Commission. US participation is led by Rensselaer computer science professor Francine Berman. Its international launch and first plenary were held in Gothenburg, Sweden, in March 2013, and the group plans follow-up meetings for September 2013 in Washington, DC, and March 2014 in Dublin, Ireland.

With recent advances in experiment, computation, and data analytics, big data have the potential to result in significant materials advances as they did for genomics. Proffen believes "we're on the cusp of having lots of real examples of the impact of big data. When we have those examples, people will believe it because they'll see it."