Beyond Big Data?

By Judith Bayard Cushing

BIG DATA IS INDEED AN EXCITING, ALBEIT CHALLENGING, TIME FOR SCIENTIFIC DISCOVERY. WHY EXCITING? EXPONENTIAL GROWTH IN DATA ACQUISITION AND GENERATION—AND SOME PROGRESS IN DATA DOCUMENTING, ARCHIVING, VALIDATION, AND RETRIEVAL—HAVE MADE MORE (IN QUANTITY and diversity) and better (less "dirty") data available to scientists. Prime examples of readily available scientific data include nationally and internationally funded repositories and data portals, where scientists including computational scientists increasingly publish and document their data: DNA-RNA-Gene-Genomics-Protein-Proteomics-Sequence databases, biomedical databases (such as more than 40 entries in the US National Institute of Health's Data Sharing Repositories), Dryad, Long-Term Ecological Research (LTER, for long-term ecology data), the *National Ecological Observatory Network* (NEON, for remotely sensed environmental data), the *Consortium of Universities for the Advancement of Hydrologic Science* (CUAHSI, a consortium for US hydrology data), Earth Observing System (EOS), Sloan Digital Sky Survey, DataONE, *Global Lake Ecological Observatory Network* (GLEON), and many others.

Partially in response to this data deluge, *Computing in Science & Engineering* has once again broadened its scope beyond computation. Initially focused on computational problems (in particular, for computational physicists who have led the way), *CiSE* addresses data as well as computation, programming, and software engineering as applied to a broad range of the physical, natural, and engineering sciences. Previous *CiSE* issues have articulated both the challenges of Big Data[1] and responses to those challenges from the computer science research community.[2] Although research, development, deployment, and education in computational speed, programming techniques, data browsing, and access will be needed for the foreseeable future, it's likely that soon scientists will be able to find and access more data than they can deal with using current tools. With the most pressing computational and data problems solved, scientists generate, document, organize, and manage data in public and private repositories. As for why the era of Big Data is challenging, scientists must now figure out how to *exploit* those data.

Data analytics, including information visualization and data mining, is industry's answer to business' Big Data problem, and IBM, Microsoft, Google, and others have geared up to provide data analytics solutions for business. For the sciences, visualization had a long history even before Jim Thomas coined the term *visual analytics*, and CiSE's *Visualization Corner* offers suggestions in every issue about new visualization techniques. What of data mining, the other Big Data exploitation strategy? For many scientists, the term connotes poor scientific method—conjuring images of fishing expeditions or even "cooking the data." The popular press hasn't helped a somewhat tarnished reputation. Articles such as "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"[3] didn't help, and author Chris Anderson's entreaty "There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?" prompted a dialog in respected scientific journals about whether hypotheses or data "come first."[4,5] Scientists might well ask what they could learn from Google, but the fact remains that current data-mining techniques were developed for business applications. Will they work for science?

Scientists don't doubt they need help with the data deluge; the data-mining challenge, however, is not only technical, but cultural—integrating a culture of hypothesis-driven science with data mining.[8,9] Scientific theory is far from dead, and scientists do cling to their "old ways," namely the scientific method. While business isn't without theory, the scientific method's long history of hypothesis-driven inquiry is deeply ingrained through education, tradition, and success. In medicine, for example, the clinical trial remains the gold standard, and many practitioners find retrospective studies—even when based on thousands of data points—"unscientific." Other scientists, well versed in statistics, ask how data-mining methods differ from the sophisticated Bayesian, Monte Carlo, or CART analyses that they already use effectively. As Deb Peters, the

 Copublished by the IEEE CS and the AIP 1521-9615/13/$31.00 © 2013 IEEE

principle investigator of the Jornada Long Term Ecological Research Site, relates:

> *The interactive use of Big Data within the scientific method could reap tremendous … rewards, but Big Data is not readily accepted by most scientists as an integral part of their research because of the way they think about and study the natural world. In fact, only a small fraction of current data is actually used by scientists,[6] and most data that are used (ca. 50 percent) are from relatively small, locally collected and stored datasets.[7,8]*

Some forward-thinking scientists are asking how future systems might integrate scientific theory and prior knowledge that isn't easily codified with data mining. For example, Bernardo Gonçalves and Fabio Porto point out that the "age of data-driven science opens the possibility of developing database technology for integrating both data and theories in the same framework."[10]

As several researchers[8–10] and articles in this issue suggest, we hope to find beyond Big Data not the end of theory, but exemplar research and new technology that uses Big Data to develop, pose, and test scientific hypotheses. I look forward to learning from this and future *CiSE* issues about how data-mining techniques are extended for and applied by leaders in computer science and in the scientific domains. Please let us at *CiSE* know what you're thinking about these questions!

References

1. F.J. Alexander, H. Adolfy, and A. Szalay, special issue on Big Data, *Computing in Science & Eng*, vol. 13, no. 6, 2011.
2. J.B. Cushing and J. French, special issue on science data management, *Computing in Science & Eng*, vol. 15, no. 3, 2013.
3. C. Anderson, "The End of Theory," *Wired*, 23 June 2008.
4. R. Weinberg, "Point: Hypotheses First," *Nature*, vol. 464, no. 7289, 2010; doi:10.1038/464678a.
5. T.R. Golub, "Counterpoint: Data First," *Nature*, vol. 464, no. 679, 2010; doi:10.1038/464679a.
6. O.J. Reichman, M.B. Jones, and M.P. Schildhauer, "Challenges and Opportunities of Open Data in Ecology," *Science*, vol. 331, no. 6018, 2011, pp. 703–705.
7. *Science,* "Staff, Challenges, and Opportunities," *Science*, vol. 331, no. 6018, 2011, pp. 692–693.
8. D. Peters, research scientist, US Dept. of Agriculture (USDA) Agricultural Research Service (ARS), Jornada Experimental Range and Lead Principal Investigator, Jornada Basin Long-Term Ecological Research (LTER) project, personal communication, July 2013.
9. O. Trelles et al., "Big Data, But Are We Ready?" *Nature Reviews, Genetics*, vol. 12, no. 224, 2011; doi:10.1038/nrg2857-c1.
10. B. Gonçalves and F. Porto, "Research Lattices: Towards a Scientific Hypothesis Data Model," *Proc. 25th Int'l Conf. Scientific and Statistical Data Management*, ACM, 2013, article no. 41.

Welcome Aboard!

CiSE welcomes the following new editorial board member. Matthew Turk is a US National Science Foundation (NSF) postdoctoral fellow at Columbia University. His research interests include the formation of the first stars in the universe, analysis and visualization of large-scale volumetric datasets, community building in scientific software development, and the development of cyberinfrastructure for computational science. Turk has a PhD in physics from Stanford University. Contact him at matthewturk@gmail.com.

Judith Bayard Cushing, a member of the CiSE editorial board, teaches computer science at The Evergreen State College in Olympia, Washington. Her research focuses on science data management across many disciplines, and she's currently the principle investigator for a US NSF-supported project called Visualization of Terrestrial and Aquatic Systems (http://blogs.evergreen.edu/vistas). Cushing has a PhD in computer science and engineering from the Oregon Graduate Institute. Contact her at judyc@evergreen.edu.