Keeping Pace with Extreme Data

by Judith Bayard Cushing
The Evergreen State College

After perusing the articles and guest editors' introduction for this issue on extreme data, I googled "extreme," and hundreds of links bubbled up—from computing, coupons, data, and fatigue, to science, scientists and science fun, to "extreme x" (don't ask). I then concluded that "extreme" has become an unreliable buzzword, used somewhat indiscriminately for its shock value. That said, the fact remains that the problems of dealing with very large sets of data do plague us scientists, despite emerging new technologies. And still, our models, sensing devices, telescopes, and even the Internet, threaten to bury us ever deeper in data. So, maybe "extreme data" is a relevant term for us to use. New technology is certainly necessary, but is it sufficient?

As the guest editors Manish Parashar and George Thiruvathukal point out in their introduction to this special issue (see p. 8), "growing data volumes, expanding distribution and dynamism of the data, [and] … growing costs … associated with transporting and processing this data, require new paradigms and practices in data management and analytics." What's the conscientious scientist to do? How can she keep up with both her own field *and* the dynamic science of data management, applying new (and as yet unproven) technologies to her own work? More fundamentally, to what extent do "paradigms and practices" include more than technology? How does technological innovation—not directly related to the scientific question at hand—come about in the sciences?

CiSE's Efforts

The Computing in Science and Engineering (CiSE) community tries to help. Contributors have recently shared in special issues their findings in scientific databases, Big Data, science data management, cloud computing, and machine learning. In each issue, *CiSE* editors or guest editors ask probing questions or make daring predictions about technological change in computational science: for example, in the guest editor's introduction to *CiSE*'s November/December 2011 Big Data issue, Francis Alexander, Adolfy Hoisie, and Alexander Szalay predict that "emerging petabytes could change every aspect of scientific disciplines" (p. 12). Francis Sullivan posits that while we might doubt that in 2019 we would do all our computing via $100 laptops and never suffer problems related to service or security, we can be sure that scientific computing environments will have evolved in unpredictable ways.

Just last year, Thiruvathukal asked how we can maintain productivity in the face of cognitive overload, and a few years ago Norman Chonacky and Dante Choi asked us point blank if we believe or trust our computers' output. Douglass Post in his introduction to the November/December 2007 issue of *CiSE* on Software Engineering for Computational Science and Engineering (pp. 10–11) perhaps gets to the heart of the data matter, asking why computational scientists and engineers don't generally adopt new software engineering practices that would greatly improve their productivity and the quality and maintainability of their software products, if not their scientific results. If we computational scientists indeed eschew new software engineering practices as we create computational artifacts, what makes us think we'll employ new technologies in data management and analytics?

In this issue, Jordan Raddick and his colleagues describe an analysis of 10 years of Web and SQL traffic from access logs of the Sloan Digital Sky Server astrophysics project, and demonstrate widespread use of Sky Server by research and educational institutions and even the lay public. Those of us interested in software engineering or

1521-9615/14/$31.00 © 2014 IEEE Copublished by the IEEE CS and the AIP

data exploitation for science and engineering might ask—what can such successes tell us about technology transfer? One important aspect of this work is that measurable, relatively long-term usage data are available for analysis; their assessment isn't based on hearsay or ex post facto surveys. But something more than just data availability seems to be occurring.

Looking Elsewhere

In my own data analytics project, we've turned to social scientists—as apparently have several national labs and large companies (including Intel and Microsoft)—to understand underlying drivers of technological change and innovation among scientists. And I have cajoled, coaxed, and even coerced my software engineering, database, and graphics students, who ask "how this relates to computer science," into reading about new findings in certain (social) sciences. Nicholas Christakis and James Fowler,[1] for example, suggest that size limitations (usually no more than ~150) of (even online) communities are genetic in nature, that language hasn't evolved for enhancing communication but to help humans deal efficiently with relatively large social networks, that online communities with their increased efficiency and selectivity keep our networks compact and efficient but deter innovation, and that colleagues of colleagues might influence us more than colleagues themselves.

How might findings from network science inform the software engineering and databasing of science and engineering artifacts? Do new centers of study such as the Center for Extreme Data Management Analysis and Visualization at the University of Utah (http://cedmav.sci.utah.edu), which focuses on "theoretical and algorithmic research, systems development, and tool deployment for dealing with extreme data," do more than develop technology? Do such centers also serve to break down barriers and act as bridges or boundary objects among different closely knit scientific communities? One possible reason for Sky Server's success is that a community of potential users existed prior to Sky Server's release. Perhaps we should ask how the use of that technology spread within the community. Who in that community were the early adopters, and were they centrally located in that network? Did Sky Server expand the existing community, or spawn new communities? In sum, I propose we ask what respect roles might scientific network theory and software or scientific advances play in the adoption of extreme data solutions in the sciences and engineering.

This issue doesn't offer all the answers to our "extreme data" problems. But it continues a strong series of special issues that have made *Computing in Science & Engineering* a source of serious thinking about big, wild, extreme, gorgeous, proliferating, profligate scientific data. Please send us your comments and thoughts on this, and for future topics. ■

Reference

1. N.A. Christakis and J.H. Fowler, *Connected,* Back Bay Books, 2009.

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 National Science Foundation-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.