Editor in Chief: **Forrest Shull**
Fraunhofer Center for Experimental Software Engineering, fshull@computer.org

Getting an Intuition for Big Data

Forrest Shull

LET ME PUT my personal experience right up front: as a researcher, I'm a data analyst by trade and have spent a large portion of my career combing through data sets of various sizes, domains, and quality. I even enjoy statistics humor (for example, http://xkcd.com/552). It's a rewarding job, but I've seen lots of ways that analysts can get things wrong, including having faith in untrustworthy data, bad assumptions made about what the data are really describing, and incorrect mathematics. (The now-infamous Reinhart-Rogoff spreadsheet error, which has had real-world implications in the extent to which it affected the evidential support for policies of economic austerity [www.nytimes.com/2013/04/30/opinion/debt-and-growth-a-response-to-reinhart-and-rogoff.html], is exactly the kind of thing that keeps me up at night.)

In this issue, we're tackling the topic of software analytics, and it's truly an exciting time to be following this field and watching the many capabilities being developed. But given the size of the datasets involved, how do you distinguish an "aha!" moment, where the size and richness of the data yield a surprising new insight, from a "Reinhart-Rogoff moment," where the size and richness of the data make it easy to miss an error somewhere along the line that spuriously affects the conclusions?

> It would be a mistake to slavishly follow the data miners to the point where you've lost the connection to reality.

Recently, I had the opportunity to speak with some of the experts doing very exciting work with big data to ask them, how do you build trustable big data systems?

Iterative Model Building

Paul Zikopoulos, director of technical professionals for IBM Software Group's Information Management division, also leads the World Wide Competitive Database and Big Data Technical Sales Acceleration teams. Several of his 16 published books are on the subject of big data. I started by expressing to him my worries that, given the size of a typical "big data" dataset, analysts can no longer intuit for themselves about what's really in their data. He didn't dismiss this concern, but he turned to a helpful metaphor and asked me to think about big data analytics as being like using GPS while driving a car. Both have helpful capabilities and can support a person in doing things that he or she couldn't do as well by themselves. But just like a driver can get into trouble by blindly following GPS and ignoring the reality outside the car window, it would be a mistake to slavishly follow the data miners to the point where you've lost the connection to reality.

Paul emphasized an idea that I've heard from other sources as well: the best way to think about using big data, if you want to make sure that

IEEE Software Mission Statement

To be the best source of reliable, useful, peer-reviewed information for leading software practitioners—the developers and managers who want to keep up with rapid technology change.

SATURN AWARDS

Continuing a tradition that was started in 2010, *IEEE Software* this year collaborated on recognizing top content at the Software Engineering Institute's SATURN 2013 conference, which was held 29 April to 3 May. The SATURN conference aims to help architecture practitioners from around the world connect with one another, share best practices, and find promising solutions. Within that already very forward-looking and practical community, *IEEE Software* helps call attention to the kind of rigorous yet pragmatic content that we like to feature by sponsoring two awards.

Selected based on votes from attendees, the Architecture in Practice Award is given to the presentation that best describes experiences, methods, and lessons learned from the implementation of architecture-centric practices. This year's winner was Simon Brown of Coding the Architecture, for his presentation titled "The Conflict between Agile and Architecture: Myth or Reality." The New Directions Award is given to the presentation that best describes innovative new approaches and thought leadership in the application of architecture-centric practices. This year's winner was Darryl Nelson of Raytheon for his presentation titled, "Next-Gen Web Architecture for the Cloud Era." General Chair Bill Pollak noted that "these awards … contribute to the maturation of the practice of software architecture by recognizing sound and innovative practices."

Several additional presentations at the conference were also recognized based on positive feedback from conference attendees, as well as their relevance to both architecture practice and the *IEEE Software* audience; they've been invited to submit articles for consideration in future issues.

outcomes are real and appropriate, is to think of the mode of interaction as being one of hypothesis testing. The data miners will produce some results, but it's up to the domain experts to come up with possible explanations for what those results really mean and then find ways to test those hypotheses. Such testing might involve more data mining or exploring ideas via data visualizations. On this latter point, a range of visualizations will work and need not be sophisticated to be useful—tag clouds can be surprisingly effective even on big datasets for understanding themes and patterns. The advantage of big data and the automated data mining that goes with it is that such hypothesis testing can be done at scale—thousands of runs of the models can be done overnight with different parameters. Moreover, doing this hypothesis testing correctly has to be viewed as a collaborative enterprise: the feeling of "I think I've found something" has to come from the business user, who can recognize actionable and insightful findings as they come along. But exploring and nailing down those findings requires working with the IT department to run the tests on various hypotheses.

One way to describe the process is as "rapid model development." Organizations build models of what's in the data, at rest, in the usual way. The model gets evaluated in the ways we're used to, starting with measures of precision and recall, and those metrics are used to fine-tine the model from there. The usual best practices (such as masking personal data) must be applied, but in ways capable of dealing with the volume of data streaming in. Big data changes the aperture on the model—we're dealing with many more attributes and data points than ever before—but not the underlying mechanics. As always, my conversation included many more nuggets than would fit into this column; interested readers will enjoy hearing more of Paul's thoughts on big data at www.computer.org/software-multimedia.

Building the Human Intuition in Big Data

Intrigued by the idea of visualizations helping humans better understand what's lurking in all that big data, I talked with experts at the University of Maryland's Human Computer Interaction Lab (HCIL). The HCIL is the oldest center in the US focusing on research in HCI, and is still going strong. I spoke with Catherine Plaisant, Associate Director of Research, and Megan Monroe, PhD student. Their "EventFlow" project represents an important effort in making "big data-size" datasets more tractable for human reasoning.

The project's goal is to summarize very large datasets of medical data, consisting of records from millions of patients, on a single display so that users can get an overview without scrolling or paging. In this view, EventFlow presents an aggregate of the data that shows the most common patterns. It also lets users query and interact with the dataset to look in more detail at specific subsets. (For more info, including demos, see the video at http://medianetwork.oracle.com/video/player/2079912021001 or the project homepage at www.cs.umd.edu/hcil/eventflow.)

This project grew out of prior work that focused on summarizing a single person's medical data. That earlier work focused on using a representation called "Lifelines" to provide an easy-to-understand summary of one person's

IBM IMPACT UNCONFERENCE

IEEE Software also had a presence last month at IBM's Impact conference, where we sponsored the "Unconference." In contrast to a traditional conference, in an unconference, potential speakers and topics come from volunteers and conference attendees (in this case, numbering over 8,000), who vote for the most interesting selections. *IEEE Software* is pleased to help support this innovative format, which so well reflects our emphasis on addressing the hot topics that software engineers need to master to keep up with their field.

To set the right tone, the unconference was kicked off with a thought-provoking discussion between Grady Booch, author of our "On Computing" department, and Tim O'Reilly, the founder and CEO of O'Reilly Media, two leading experts who are as familiar as anyone with the topics on the mind of today's software engineers. Enjoy excerpts from their fun and insightful back and forth at www.computer.org/software-multimedia.

medical history over years of care. The EventFlow work now builds another level of complexity where analysts can look across many such patients. Such work supports new approaches to medical research, in which an increasing number of investigations can now be done retrospectively—that is, by analyzing what has happened in previous cases where the illness or event was seen—and don't always require new clinical studies to investigate a specific hypothesis. The intended users for now are clinical researchers, although I think it's easy to imagine a physician in the not-too-distant future asking EventFlow to help identify records that match a current patient, to get a sense of how treatment options have worked in the past.

Catherine and Megan mentioned that the EventFlow dataset's size was a novelty for visualization work and represented new challenges related to scale. But at the same time, they stressed that the challenges weren't where you might expect: performance and processing power weren't areas of concern; rather, the hard problems were related to how to make the display usable given the amount of data that had to be presented intelligibly.

In describing the importance of supporting visualizations for such large datasets, Megan likes to reference the saying that "it's more about the journey than the destination." She painted a picture that was very similar to Paul's theme of "hypothesis testing." Certainly, it's possible to give researchers an answer to a given question just using data mining techniques. But often, the very definition of what constitutes a meaningful event pattern changes as researchers do the exploration through visualization tools.

We might think that the volume of big data ensures that the answers are always buried in there, just needing to be found. However, Catherine and Megan have found that, just like hypothesis testing in the small, analysts often realize that they need yet more data as new hypotheses arise. They described a cycle they've often seen, in which data needs to narrow down relatively quickly; as a user discovers what's interesting and relevant, it's often only a subset of the available data fields that become important. But even

IEEE Software

STAFF

Lead Editor
Brian Brannon
bbrannon@computer.org

Content Editor
Camber Agrelius

Manager, Editorial Services
Jenny Stout

Senior Editor
Linda World

Publications Coordinator
software@computer.org

Production and Design Editor, Webmaster
Jennie Zhu-Mai

Contributor
Alex Torres

Cover Artist
Eero Johannes

Director, Products & Services
Evan Butterfield

Senior Manager, Editorial Services
Robin Baldwin

Senior Business Development Manager
Sandra Brown

Membership Development Manager
Cecelia Huffman

Senior Advertising Coordinator
Marian Anderson
manderson@computer.org

Editorial: All submissions are subject to editing for clarity, style, and space. Unless otherwise stated, bylined articles and departments, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *IEEE Software* does not necessarily constitute endorsement by IEEE or the IEEE Computer Society.

To Submit: Access the IEEE Computer Society's Web-based system, ScholarOne, at http://mc.manuscriptcentral.com/sw-cs. Be sure to select the right manuscript type when submitting. Articles must be original and not exceed 4,700 words including figures and tables, which count for 200 words each.

IEEE prohibits discrimination, harassment and bullying: For more information, visit www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

WELCOME NEW ADVISORY BOARD MEMBERS

I'm very pleased to welcome two new members to our advisory board.

Rafael Prikladnicki is a professor in the Computer Science School at PUCRS, Brazil. His areas of expertise are distributed software development and agile methods for software development. Prikladnicki is director of the Technology Management Agency (AGT) at PUCRS, where he's responsible for managing the interaction between PUCRS and its industry and government partners for the development of R&D projects. His 2007 book, *Distributed Software Development: Developing Software with Distributed Teams*, was the first Portuguese book on this topic, and he also leads one of the main research groups in this area in Brazil. In 2011, Prikladnicki received the PhD innovation award, promoted by FAPERGS (the state of Rio Grande do Sul funding agency) for his research on global software engineering, conducted in collaboration with companies at Tecnopuc (PUCRS' Scientific and Technology Park).

Walker Royce is the Chief Software Economist in IBM's Software Group and a principal consultant and practice leader specializing in measured improvement of systems and software development capability. Royce is the author of three books: *Eureka! Discover and Enjoy the Hidden Power of the English Language* (Morgan James, 2011), *The Economics of Software Development* (Addison-Wesley, 2009), and *Software Project Management, A Unified Framework* (Addison-Wesley, 1998). From 1994 through 2009, Royce was the vice president and general manager of IBM's Rational Services organization and built a worldwide team of 500 technical specialists in software delivery best practices and $100 million in consulting services. Before joining Rational/IBM, he spent 16 years in software project development, software technology development, and software management roles at TRW Electronics & Defense. Royce was a recipient of TRW's Chairman's Award for Innovation for his contributions in distributed architecture middleware and iterative software processes in 1990 and was a TRW Technical Fellow.

in big data, it often becomes important to expand the dataset to draw in more types of information as new questions arise. Data analysis often has cycles of dataset contraction and expansion as users get a better idea about what's interesting, and that cycle doesn't change just because we're dealing with big data.

For all of the above reasons, my interviewees stressed that visualization isn't in an adversarial role to data mining, although it's sometimes presented as if computer-assisted and automated pattern detection are two incompatible and competing philosophies. The reality is quite the opposite: visualization and data mining should proceed in tandem, each helping the other to deliver a holistic look at the data's meaning.

And as to how those visualizations should be designed in the era of big data, Ben Shneiderman, the HCIL's founder, has a mantra that visualization tools should provide an overview first, then allow zooming and filtering, and provide deeper details on demand. Megan and Catherine have a lot of experience that shows this still provides a useful approach even when dealing with big data; to hear more about these experiences, listen to our conversation at www.computer.org/software-multimedia.

Thus the directions that research is taking for visualizing big data are a mix: both applying existing principles such as "details on demand" at higher levels of scale (for example, providing more levels of drill-down between the overview and the lowest level of granularity) and also coming up with new and specialized visualizations that allow larger quantities of data to be represented intelligibly. And as such tools become more sophisticated and more mainstream, I can hope we're building toward a scenario where humans can be just as comfortable with the nuts and bolts of a big dataset as we've been for other analyses.

FORREST SHULL is a division director at the Fraunhofer Center for Experimental Software Engineering in Maryland, a nonprofit research and tech transfer organization, where he leads the Measurement and Knowledge Management Division. He's also an adjunct professor at the University of Maryland College Park and editor in chief of *IEEE Software*. Contact him at fshull@computer.org.