**Your Questions About Big Data Answered**

Experts help define what the growing field is all about

BY MONICA ROZENFELD

Our September issue on [big data](http://theinstitute.ieee.org/static/special-report-big-data) raised a lot of thought-provoking questions by our readers about what *big* actually means, if the field is all just hype, and which industries will be most affected. To help answer these questions, *The Institute* invited [three experts](http://theinstitute.ieee.org/ieee-roundup/opinions/ieee-roundup/ask-the-experts-big-data) to respond. They include IEEE Fellow [Grady Booch](http://researcher.watson.ibm.com/researcher/view.php?person=us-gbooch), chief scientist for software engineering at the [IBM Almaden Research Laboratory](http://www.research.ibm.com/labs/almaden/index.shtml), in San Jose, Calif.; IEEE Fellow [Manish Parashar](http://parashar.rutgers.edu/), founding director of the [Rutgers Discovery Informatics Institute](http://rdi2.rutgers.edu/) (RDI2), in Piscataway, N.J.; and [Dennis Shasha](http://cs.nyu.edu/shasha/), professor of computer science at New York University, in New York City. Here is what they had to say.

**What is the difference between “big data” and “large amounts of data,” and where is the line drawn?**

**SHASHA:** Semantically, there is no real difference. However, in practice, big data has come to mean data that is either collected from physical sensors or generated from the activity of many people (such as financial ticker data or [social media](http://theinstitute.ieee.org/ieee-roundup/opinions/ieee-roundup/how-to-mine-social-media)). As such, it tends to be of lower quality than curated data in say an employee database. But because it’s aggregated, big data can often help answer new questions. For example, taxi pickup and drop-off data can give information about a city’s traffic flow. Biological sequence data can tell about the effects of some genetic mutation. Thus, the term at least partly reflects the new opportunities that such data provides.

**PARASHAR:** While increasing volumes, velocity, and variety are important dimensions of big data, I believe a more appropriate interpretation of “big” is the tremendous impact that the pervasive access and the ability to transform this data into knowledge and insights can have on business, engineering, medicine, science, and society at large. The data does not necessarily have to be large, but rather the quality of the data and our ability to transform that data into meaningful and timely insights are often more important.

**Is “big data” a recent marketing term, or can it legitimately provide us with new information we couldn’t get just a few years ago, say in the field of health care?**

**BOOCH:** Honestly? Yes, it really is a marketing term. I worked with military and industrial systems in the 1980s and 1990s that used large volumes of data. The difference is, however, that data is no longer used for these niche purposes alone but instead have become part of the mainstream, available to just about any industry.

**PARASHAR:** While “big data” has become a marketing term and has the expected hype associated with it, unprecedented instrumentation and the growth of digital data sources does have the potential for fundamentally transforming our ability to understand and manage our lives and our environment.

 For example, data collection using personal wearable devices, smart homes, cars and transportation infrastructure, and so on, are already having an impact on our lifestyles and daily lives. Looking ahead, the transformative vision of personalized medicine aims to harness these advances in technology, along with genetics and biomedical research, to understand the roots of disease, develop targeted therapies, and ultimately provide predictive, preventive, and precise [care to patients](http://theinstitute.ieee.org/technology-focus/technology-topic/better-health-care-through-data).

**For those of us in the industry, we’ve been working with “big data” for years now. Some of these programs available today give us insights we already knew, or could be figured out some other way. Is there a good approach to integrating existing business knowledge, expertise, and common sense into the process when churning out facts, alongside analyzing data?**

**Shasha:** Your inclination is correct. The ideal use of big-data analysis in fields such as biology and telecommunications is to meld experience-derived insight with data analysis. The biologist who knows how important auxin is to plants uses big data to understand exactly which pathways are affected. The engineer who knows that millimeter waves may be attenuated by rain can make better use of channel models when laying out base stations in Honduras.

**Parashar:** I do believe that the “new” data sources should be used as complements to existing business knowledge, expertise, and common sense in decision-making as part of any business process. The specific mechanisms for achieving this will depend on the nature of the business and its processes.

**I am concerned about the abuse of personal data, especially by some of the larger entities such as government agencies that are using our data for their purposes. How is our data being protected from misuse?**

**BOOCH:** Not very well, unfortunately, although many are trying very hard to solve this. The problem, by the way, is not just the release of information into the wild without our permission. It is the problem of what IBM Fellow Jeff Jonas speaks of as non-obvious connections. Individually, we may leave a trail of seemingly innocuous bits, but when those bits are put together in context…well, then the trouble really begins. It’s impossible to pin responsibility on any one source when the real problem lies in the conjunction of data. There really is no good technical solution for this yet.

**PARASHAR:** There are multiple dimensions to protection that are essential for safeguarding data from misuse, including technical—such as the algorithms and technologies for managing data security, integrity, and privacy. There are also

legal and regulatory laws and societal awareness. All of these components have to work together along with good common sense to address this issue. This is often a moving target. It is an important issue and needs to be continually addressed.

**For engineers without a big-data background, how can we break into the field? What classes and conferences might be useful to attend?**

**PARASHAR:** Many universities offer courses in data sciences that are targeted to working engineers and other professionals, and many online courses are also available. One example is the [Rutgers Professional Science Master’s Program](http://mbs.rutgers.edu), which offers a concentration in analytics and data sciences.

SHASHA: Big data expertise is based on the technologies of data cleaning, database management, and data mining/machine learning/statistics. These are all vast fields, data cleaning relies on a combination of approximate matching techniques and relational constraints. Machine learning and statistics are also vast but there are excellent packages available for both. The best way to learn is just to try a project. If you feel there are severe gaps in your knowledge, then consider taking courses.

**Which tech areas are going to be most affected by big data?**

**PARASHAR:** The exponential and pervasive growth of digital data sources and need to transport, manage, process, analyze, and pervasively access this data presents new technical challenges and opportunities at all levels, including for new technologies such as [software-defined networks](http://sdn.ieee.org/).

**SHASHA:** The main tech areas to be affected will be the ones that handle the most data. So network capacities, storage capacities will increase as will the need to sample data, process it on the fly, and draw conclusions. This is a ripe area for all kinds of engineering invention.