

# Big and Useful: What's in the Data for Me? (Panel Description)

Rada Chirkova  
North Carolina State University  
rychirko@ncsu.edu

Jun Yang  
Duke University  
junyang@cs.duke.edu

## 1. BACKGROUND AND GOALS

In the context of extracting value to users from the available data, the database community has historically been focusing on efficient processing of primarily structured queries posed by expert users mostly on structured, pre-organized data. The rather recent “Big-Data” phenomenon has been shaping a world where extreme quantities of data are collected in an ad-hoc, almost accidental, way. Further, the user targets for Big Data are broader than those traditionally considered by the database research community.

The main goals of this panel are to identify the pain points when end users attempt to extract real value from ad hoc collections of Big Data, and to provide alternative viewpoints on what the database community should work on, if it is to play a bigger role in bringing the benefits of Big Data to the masses.

**Potential issues of controversy** Are there things more important to work on than what the community is currently doing? Specifically:

- We argue for more focus on the *variety* aspect of Big Data—as opposed to *volume* and *velocity*, which much of the current research has focused on;
- We argue for more attention to pre-analysis (e.g., data wrangling) and post-analysis (e.g., generating narratives) besides analysis itself; and
- We argue for more support for new user archetypes (e.g., users who own the data being collected and need to understand it, but are not experts in it), as opposed to just the traditional ones (e.g., programmers and analysts).

## 2. TENTATIVE QUESTIONS

1. Analysis is often the first thing that comes to mind when one mentions Big Data. What are the main bottlenecks there? Where would it make sense for the community to concentrate its efforts?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Articles from this volume were invited to present their results at the 39th International Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy.

*Proceedings of the VLDB Endowment, Vol. 6, No. 12*  
Copyright 2013 VLDB Endowment 2150-8097/13/10... \$ 10.00.

2. Now suppose the data is perfectly ready for the Big-Data Analysis. What are important directions for us as far as the analysis goes?
3. Who are the end consumers of the analysis results? How does it make sense to post-process the analysis results for typical consumer categories?
4. How about non-business owners/users of Big Data? What can be done so that they could understand the data?
5. How do we define value of the data for users?
6. We have seen Big-Data “killer apps” for business and science. What are the killer apps for ordinary people? Why would they actually care about Big Data?
7. Besides static representation of Big Data in a visual, textual, or other form, is there any need for interaction with it? What are the right primitives? Are they any different from what has been the case in the pre-“Big-Data” era?
8. The “citizen scientist” is an emerging new important type of user of Big Data. Are there any special requirements that it brings to the table? Is there any difference with the “scientist” or the “citizen” types?

## 3. PANELISTS

**Rada Chirkova (panel moderator)** Rada Chirkova is an Associate Professor in the Department of Computer Science at North Carolina State University. She received her PhD from Stanford University in 2002. She has been a recipient of the US National Science Foundation Career Award and of multiple IBM Faculty Awards, and is a senior member of the ACM. Her current research interests include reformulation of data and queries for a variety of purposes and application domains.

**Minos Garofalakis** Minos Garofalakis is a Professor and the Director of the SoftNet Lab at the Department of Electronic & Computer Engineering of the Technical University of Crete, Greece. He received his PhD in Computer Sciences from the University of Wisconsin in 1998, and held senior researcher positions at Bell Labs, Intel Research Berkeley, and Yahoo! Research, as well as an Adjunct Associate Professor appointment with UC Berkeley EECS (2006–2008). His current research focuses on centralized and distributed data streams, data synopses and approximate query processing,

uncertain databases, and Big-Data Analytics and data mining. He serves as a PI for a number of European research projects in these areas. Minos is an ACM Distinguished Scientist (2011), and a recipient of the Bell Labs President's Gold Award (2004) and the IEEE ICDE Best Paper Award (2009).

**Joseph M. Hellerstein** Joseph M. Hellerstein is a Chancellor's Professor of Computer Science at the University of California, Berkeley, whose work focuses on data-centric systems and the way they drive computing. He is an ACM Fellow, an Alfred P. Sloan Research Fellow and the recipient of two ACM-SIGMOD "Test of Time" awards for his research. In 2010, Fortune Magazine included him in their list of 50 smartest people in technology, and MIT's Technology Review magazine included his Bloom language for cloud computing on their TR10 list of the 10 technologies "most likely to change our world." He serves on the technical advisory boards of EMC, SurveyMonkey, Platfora and Caprictivity. Hellerstein is co-founder and CEO of Trifacta, which develops productivity software for data analysts.

**Yannis Ioannidis** Yannis Ioannidis (UC Berkeley PhD, 1986) is the President and General Director of the ATHENA Research and Innovation Center and a Professor at the Department of Informatics and Telecommunications of the University of Athens. His research work focuses on data and information management systems (especially query processing and optimization as well as complex dataflow processing), user modeling and attitude management systems, scientific data infrastructures, digital libraries and repositories, and computer-human interaction. He is coordinating several European and national projects on the above topics. Yannis Ioannidis is an ACM and IEEE Fellow and a member of Academia Europaea, he has received the 2003 VLDB "10-Year Best Paper" Award as well as several other research and teaching awards, and he has recently finished his term serving as the ACM SIGMOD Chair.

**Zachary Ives** Zachary Ives is an Associate Professor and the Markowitz Faculty Fellow at the University of Pennsylvania. His research interests include data integration and sharing, "Big Data," sensor networks, and data provenance and authoritativeness. He is a recipient of the NSF CAREER award, and an alumnus of the DARPA Computer Science Study Panel and Information Science and Technology advisory panel. He has also been awarded the Christian R. and Mary F. Lindback Foundation Award for Distinguished Teaching. He serves as the undergraduate curriculum chair for Penn's Singh Program in Networked and Social Systems Engineering. He is a co-author of the textbook *Principles of Data Integration*. His recent projects include a highly scalable cluster compute engine called REX, the ieeg.org portal for cloud-based sharing and analysis of network neuroscience data, and the Q system for integrating hundreds or thousands of disparate data sources.

**H.V. Jagadish** Dr. Jagadish obtained his PhD from Stanford in 1985, and worked many years for AT&T, where he eventually headed the database department. After a brief detour through the University of Illinois, he joined the University of Michigan in the fall of 1999, where he currently is the Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science. He is the founding Editor-in-Chief of the Proceedings of the Very Large Database Endowment (PVLDB). His SIGMOD 2007 keynote on database

usability gave the database community a significant impetus to think about usability issues. He recently coordinated a community white paper on Big Data.

**Jun Yang** Jun Yang is an Associate Professor of Computer Science at Duke University in USA. He is broadly interested in research on databases and data-intensive systems. He received his BA from University of California at Berkeley in 1995, and his PhD from Stanford University in 2001. He has been a recipient of the US National Science Foundation CAREER Award, IBM Faculty Award, HP Labs Innovation Research Award, and David and Janet Vaughan Brooks Teaching Award. His recent interests include continuous querying systems, scalable statistical computing, and computational journalism.