

Education and Career Paths for Data Scientists

Magdalena Balazinska
Computer Science & Engineering
University of Washington, USA
magda@cs.washington.edu

Susan B. Davidson
Computer and Information Science
University of Pennsylvania, USA
susan@cis.upenn.edu

Bill Howe
Computer Science & Engineering
University of Washington, USA
billhowe@cs.uw.edu

Alexandros Labrinidis
Computer Science
University of Pittsburgh, USA
labrinid@cs.pitt.edu

ABSTRACT

MOTIVATION: As industry and science are increasingly data-driven, the need for skilled data scientists is exceeding what our universities are producing. According to a McKinsey report: "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills". Similarly, the ability to extract knowledge from scientific data is accelerating discovery and we need the next generation of domain scientists to be experts not only in their domain but also in data management. At the same time, however, researchers in academia who focus on building instruments or data management tools are often less recognized for their contributions than researchers focusing purely on the actual science.

OVERVIEW: The goal of this panel will be to discuss all these challenges. We will discuss various aspects of how we should be educating both the emerging "data science" experts and the next generation of database and domain science experts. The panel will also discuss career paths for researchers who choose to specialize in developing new methods and tools for Big Data management in domain sciences, with recommendations for how we should better support these less traditional career paths.

1. BIOGRAPHICAL SKETCHES

1.1 Organizer

Magdalena Balazinska is an Associate Professor in the department of Computer Science and Engineering at the University of Washington. Magdalena's research interests are in the field of database management systems. Her current research focuses on big data management, sensor and scientific data management, and cloud computing. Magdalena holds a Ph.D. from the Massachusetts Institute of



Permission to make digital or hard copies of part or all 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. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

SSDBM'13, July 29–31 2013, Baltimore, MD, USA
ACM 978-1-4503-1921-8/13/07 \$15.00

Technology (2006). She is a Microsoft Research New Faculty Fellow (2007), received an NSF CAREER Award (2009), a 10-year most influential paper award (2010), an HP Labs Research Innovation Award (2009 and 2010), a Rogel Faculty Support Award (2006), a Microsoft Research Graduate Fellowship (2003-2005), and multiple best-paper awards.

1.2 Panelists

Susan B. Davidson received the B.A. degree in Mathematics from Cornell University, Ithaca, NY, in 1978, and the M.A. and Ph.D. degrees in Electrical Engineering and Computer Science from Princeton University, Princeton NJ, in 1980 and 1982. Dr. Davidson is the Weiss Professor and Chair of Computer and Information Science at the University of Pennsylvania, where she has been since 1982. She also served as Deputy Dean of the School of Engineering and Applied Science from 2005-2007.



Dr. Davidson's research interests include database and web-based systems, scientific data management, and extra-large databases. She co-developed the Kleisli data integration system (with Drs. Buneman, Tannen and Overton), featuring a complex value model of data that was amenable to optimizations, which was used for on-the-fly data integration of large genomic datasets. Kleisli was subsequently commercialized in the company GeneticXChange. She has also developed techniques for provenance management in scientific workflow systems, including support for search and query, techniques for focusing user attention on "relevant" provenance information, and marrying database-style and workflow-style provenance management using Pig-Latin to elucidate the function of black-box modules. More recently, she has focused on privacy concerns surrounding the capture and use of provenance information in both databases and workflow systems.

Dr. Davidson was the founding co-director of the Penn Center for Bioinformatics from 1997-2003, and the founding co-director of the Greater Philadelphia Bioinformatics Alliance. She holds a secondary appointment in the Department of Genetics, is an ACM Fellow, received the Lenore Rowe Williams Award (2002), and was a Fulbright Scholar and recipient of a Hitachi Chair (2004).

Bill Howe is the Director of Research for Scalable Data Analytics at the UW eScience Institute and holds an Affiliate Assistant Professor appointment in Computer Science & Engineering, where he studies data management, analytics, and visualization systems for science applications. Howe has received two Jim Gray Seed Grant awards from Microsoft Research for work on managing environmental data, has had two papers elected to VLDB Journal's "Best of Conference" issues (2004 and 2010), and co-authored what are currently the most-cited papers from both VLDB 2010 and SIGMOD 2012. Howe serves on the program and organizing committees for a number of conferences in the area of databases and scientific data management, and serves on the Science Advisory Board of the SciDB project. He has a Ph.D. in Computer Science from Portland State University and a Bachelor's degree in Industrial & Systems Engineering from Georgia Tech.



Alexandros Labrinidis received his Ph.D degree in Computer Science from the University of Maryland, College Park in 2002. He is currently an associate professor at the Department of Computer Science of the University of Pittsburgh and co-director of the Advanced Data Management Technologies Lab. He is also an adjunct associate professor at Carnegie Mellon University (CS Dept).



Dr. Labrinidis' research focuses on user-centric data management for network-centric applications, including web-databases, data stream management systems, sensor networks, and scientific data management (with an emphasis on big data). He has published over 65 papers at peer-reviewed journals, conferences, and workshops; he is the recipient of an NSF CAREER award in 2008. Dr. Labrinidis served as the Secretary/Treasurer for ACM SIGMOD (July 2009 - June 2013), and has previously served as the Editor of SIGMOD Record. He has also served on numerous program committees of international conferences/workshops.