Small Data

(Panel)

Oliver Kennedy
University at Buffalo
okennedy@buffalo.edu

D. Richard Hipp
Hipp, Wyrick & Company, Inc.
drh@sqlite.org

Stratos Idreos
Harvard
stratos@seas.harvard.edu

Amélie Marian
Rutgers
amelie@cs.rutgers.edu

Arnab Nandi
Ohio State University
arnab@cse.ohio-state.edu

Carmela Troncoso
IMDEA Software Institute
carmela.troncoso@imdea.org

Eugene Wu
Columbia University
ew2493@columbia.edu

***Abstract*—Data is becoming increasingly personal. Individuals regularly interact with a wide variety of structured data, from SQLite databases on phones, to HR spreadsheets, to personal sensors, to open government data appearing in news articles. Although these workloads are important, many of the classical challenges associated with scale and Big Data do not apply. This panel brings together experts in a variety of fields to explore the new opportunities and challenges presented by "Small Data"**

I. Overview

Over a decade ago, challenges to assumptions like "Distributed systems failures are outliers", "We can't collect everything", and "There isn't enough data to distinguish signal from noise" led us into the big data era. Now, fundamental assumptions are changing again. Smart devices are making data more personal. Intelligence is moving closer to the edge with low-cost embedded computing platforms. Tools like D3 are making interactive visualizations a key part of news reporting. Interfaces like Wolfram Alpha and Siri are putting complex question-answering within easy reach. In short, we are transitioning to an era where where the data management bottleneck is personal and per-device interactions, rather than scalability — an era of "Small Data". This panel will facilitate a discussion of small data and encourage participants to challenge long-held data management assumptions. After brief 2-3 minute self-introductions, this panel will encourage audience engagement through an open debate and discussion format. Topics for discussion will include: (1) What is small data and why should the database community care? (2) How do human factors affect data management systems and how data is accessed? (3) As edge computing devices like smart sensors, embedded linux, phones, and watches become pervasive, what bottlenecks will DBMSes have to contend with? (4) Is SQL the right language for a landscape dominated by imperative programming? (5) What tools are required to help individuals leverage open public data? (6) How should new small data technologies be evaluated? (7) What resources are available for new research on small data?

II. Moderator and Panel

Oliver Kennedy (Moderator) is an assistant professor at the University at Buffalo, working on uncertain data, database usability, query optimization, and data structures. Oliver's work includes DBToaster (as seen in the best of VLDB 2012 issue of VLDBJ), Mimir (a user-friendly probabilistic ETL tool), and PocketData (an embedded database benchmark based on usage patterns from Android smartphones in the wild).

D. Richard Hipp got his PhD at Duke University in 1992, and is the original author and principle maintainer for the SQLite database engine, the most widely used database engine in the world. Richard is also the founder and a co-owner of Hipp, Wyrick & Company, Inc., a North Carolina company that provides advanced software design and implementation services for an international clientele.

Stratos Idreos is an assistant professor of Computer Science at Harvard University where he leads DASlab, the Data Systems Laboratory@Harvard SEAS. Stratos' work emphasizes making it easy to design efficient data systems as applications and hardware keep evolving and on ease-of-use for non-experts. Stratos is the recipient of numerous awards including the 2011 ACM SIGMOD Jim Gray Doctoral Dissertation award, the NSF CAREER award, and an IEEE TCDE Early Career award.

Amélie Marian is an Associate Professor in the Computer Science Department at Rutgers University, where she leads the DigitalSelf project, which aims at providing users with tools to regain control of and exploit their digital data traces. Her research interests include personal information management, semi-structured data processing and web data management.

Arnab Nandi is an assistant professor at Ohio state. Arnab's research is in the area of database systems, focusing on exploiting user behavior to address challenges in large-scale data analytics and interactive data exploration. Arnab is a founder of The STEAM Factory and faculty director of the OHI/O Hackathon Program. Arnab is a recipient of the NSF CAREER Award, an IEEE TCDE Early Career award, and the Ohio State College of Engineering Lumley Research award.

Carmela Troncoso is a researcher at the IMDEA Software Institute where she leads the research line on privacy enhancing technologies. Her research focuses on developing systematic means to analyze and design robust privacy-preserving systems. She has over 30 articles in top Security and Privacy venues, and is part of the board of the Privacy Enhancing Technologies Symposium, that she will chair in 2018.

Eugene Wu is an assistant professor of Computer Science at Columbia University focusing on accelerating the democratization of data. He interests include algorithms to explain data analysis results, data cleaning and preparation, and data visualization management systems. He believes that small and medium data sets are the life-blood of our country, and that they are struggling for resources. Please vote for small data.