Data, Technology and the Challenges of Abundance

Joseph M. Hellerstein
UC Berkeley
hellerstein@berkeley.edu

1. ABSTRACT

It is no secret that we live in good times for experts in data-centric technologies. There are ample resources for research in our area from both federal and private sources. There are unprecedented opportunities for researchers to interact with and influence companies of all sizes, from startups to major Internet services. Skilled human resources are flocking to our area from across the globe, flooding our courses and applying for research opportunities. Entire new areas of study and profession are being planned around Data Science.

The database research community has never had it so good, but the opportunities before us raise significant questions that we need to tackle in short order.

A short list of concerns includes the following:

Relationship to Computing. With the notable exception of IBM, most of the pioneering academic and engineering institutions in computing history viewed data management as a side dish to the main course of Computer Science. The tables are finally turning thanks to Big Data and the Cloud, with a general perception that computers are incidental vehicles enabling us to achieve the key task of gathering and manipulating data. But how should experts in data management approach the Computer Science establishment and canon? Suppose we had control over things like school curricula, research funding, company strategy and government policy—what would we want to achieve? In order to get a modicum of such control, how should we proceed? Should we be rethinking elementary courses in computing to emphasize data-centric topics and viewpoints that will prepare students for data-centric jobs? Should we be designing the next generation of programming languages, to attack data-centric problems in modern software development? Should we be agitating for control in national and international bodies like NSF and CCC to drive these goals? How can we leverage our influence in industry, and improve our influence in national funding agencies?

Educational leadership. The number of people wanting to learn about data technologies has grown at an unprecedented pace. At the same time, the incentives for joining teaching professions have shrunk significantly, as rewarding opportunities at startups and big companies have grown. Meanwhile, most want-to-be practitioners learn about data management technology from relatively unsophisticated—or at best doggedly pragmatic—sources on the internet. Who is best qualified to teach the next generation of practitioners and experts in database technologies? How do we best reach the right audiences of students, lifelong learners and would-be mentors?

Concentration of expertise and its effects. There is a very small number of companies that have expertise with massive data services. The expertise in these companies—regarding the design of data systems, the content they manage, and the usage they observe—is essentially impossible to replicate elsewhere. Meanwhile, when these services do publish details or even source code for their systems, the following in the broader community can often be powerfully influential and yet technically ill-suited: what is well-designed for a major Internet service may be a terrible fit for 99% of other companies. How do we reckon with this consolidation of knowledge and influence? Is this a winners-take-all ecosystem that will eventually squelch outside innovation? What role do technologists outside those organizations play, and how can they continue to learn and improve? How can engines of innovation like universities and venture capital be harnessed to useful ends?

Diversity of use. For decades, data management systems have been designed for two categories of users: software developers and professional data analysts. Today, every consumer uses sophisticated data products like recommender services, fraud detection, and even interactive data visualizations in outlets like media and personal health and finance services. The "consumerization of work" means that users expect equally intuitive experiences working with data technologies in their chosen work. Can we develop science, engineering and design principles for data systems targeting these new categories of users? Is there a middle ground of usability between programmer APIs and consumer recommendations that supports rich yet "self-service" user experiences?

Relationship outside computing. Database research has long been tied to use cases in Enterprise data management, with a small but persistent thread of work in Scientific data management. Relative to other parts of computer science, however, database researchers have been slow to tackle issues outside of enterprise data management and Internet services. Yet Data Science is increasingly viewed as a topic of broad interest across society. What can we do as a community and as individuals to be relevant to a broader range of applications, and interact meaningfully with thinkers from a wider range of disciplines? Again, if we suddenly had the ability to engage on those levels, what would we like to achieve?

7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA..