Should we all be teaching "Intro to Data Science" instead of "Intro to Databases"?

Bill Howe
Computer Science & Engineering
University of Washington
Seattle, WA
billhowe@cs.washington.edu

Michael J. Franklin
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA
franklin@cs.berkeley.edu

Juliana Freire
Department of Computer Science and Engineering
New York University, New York
juliana.freire@nyu.edu

James Frew
University of California, Santa Barbara
Bren School of Environmental Science and Management
Santa Barbara, CA
frew@bren.ucsb.edu

Tim Kraska
Computer Science Department
Brown University
tim_kraska@brown.edu

Raghu Ramakrishnan
Microsoft Cloud Information Services Laboratory
Microsoft Corporation
raghu@microsoft.com

1. ABSTRACT

The Database Community has a unique perspective on the challenges and solutions of long-term management of data and the value of data as a resource. In current computer science curricula, however, these insights are typically locked up in the context of the traditional Intro to Databases class that was developed years (or in some cases, decades) before the modern concept of Data Science arose and embedded in the discussion of legacy data management systems. We consider how to bring these concepts front and center into the emerging wave of Data Science courses, degree programs and even departments.

Categories and Subject Descriptors

K.3.2.b [**Computer and Education**]: Computer Science Education

2. INTRODUCTION

The term *data science* has recently enjoyed explosive popularity in industry: thousands of data scientist positions are open, and universities are scrambling to train students to fill them. Masters programs, certificate programs, new courses, and workshops and bootcamps have sprung up nationally and internationally, and several new research institutes have also been launched in the last 12 months alone. These programs all tend to involve some combination of topics in data management, visualization, and statistics/ML, with statistics and machine learning topics usually receiving disproportionate attention from students and the popular media. But data science practitioners report that the data management challenges — the forte of the database community — are the real bottlenecks. The acquisition, cleaning, integration, manipulation, and sharing of data — flexibly and scalably — are the issues that "keep them up at night." Moreover, we hear them quip "80% of analytics is just sums and averages" — the implication being that simple statistical methods are more than sufficient for many applications. So while the database community has the potential to take a leadership role in the education and training of the next generation of successful data scientists, we are at risk of allowing other communities to drive the conversation.

The result of these trends is that we are in the process of losing mindshare for data science and big data among upcoming students. Entire courses in data science are taught with R, implicitly conveying to students that datasets larger than main memory aren't all that important to a data scientist. When scale is discussed, NoSQL systems (despite being occasionally softened to mean "Not Only SQL") are positioned as the only viable solution for "web scale" despite the obvious tradeoffs.

Overall, our core strength is about treating data as a resource that is a key part of any information system, one that is not associated with any one particular application or program. This idea is important to teach in any data science curriculum and seize the opportunity to put database techniques and technologies front and center in the data science discussion. For example, we should unapologetically argue that the relational algebra is as important as the linear algebra in the data scientist's toolbox, and that transactions and schemas are important systems-level features that cannot be easily left up to the application programmer.

3. QUESTIONS TO CONSIDER

We consider how the database community can influence data science education nationally and internationally, and how data science should influence how we should teach databases. We ask the following questions:

- Is "Intro to Data Science" the new "Intro to Computer Science"? For instance, the Intro to CS class on Udacity (`https://www.udacity.com/course/cs101`) builds a search engine. Doesn't that say it all?

SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.
ACM 978-1-4503-2376-5/14/06.
http://dx.doi.org/10.1145/2588555.2600092.

- At what level (junior, senior,...) should Data Science be offered and what are the appropriate pre-requisites? Who are we targeting? Majors, non-majors, working professionals seeking masters degrees, researchers in various fields seeking depth?
- If Intro to Data Science doesn't replace Intro to Computer Science, should it replace our traditional Intro to Database courses?
- Do we all now need to learn about all these statistics and machine learning algorithms in order to teach such a course?
- Is Data Science just another class or two, a reorganization of exsting material, or is it an emerging field on its own?
- What are the core set of skills, tools, theories, etc. that Data Scientists need to know and how should we teach them? Are there "principles of data science" that are distinct from existing material?
- As database researchers, how should we be teaching/positioning our core contributions in the context of data science? What is the role of NoSQL and other not-exactly-database technologies? How much expertise in systems and architectures do data scientists need?
- Which communities in CS (visualization, ML) and other disciplines (stats, sciences, etc) should we engage with on this and in what way?
- How broadly can DB formalisms be applied to data science problems? Can we "jailbreak" the relational algebra from conventional RDBMS systems and used broadly as a reasoning tool?
- How do we better integrate our expertise in other relevant topics — visualization and machine learning in particular?
- What formats should be exploring for the material? Undergrad, masters, PhD programs, certificates, online? What about non-majors? Physicists and biologists are starting successful big data software companies — should we embrace this breadth?
- How do we get industry involvement in the classroom (or should we)?
- Big data resources are important for big data education: How do we partner with industry to provide a shared infrastructure?