

Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype?

Divy Agrawal
QCRI
dagrawal@qf.org.qa

Michael Jordan
UC Berkeley
jordan@cs.berkeley.edu

Magdalena Balazinska
University of Washington
magda@cs.washington.edu

Tim Kraska
Brown University
tim_kraska@brown.edu

Michael Cafarella
University of Michigan
michjc@michigan.edu

Raghu Ramakrishnan
Microsoft
raghu@microsoft.com

Christopher Ré
Stanford
chrismre@cs.stanford.edu

Categories and Subject Descriptors

H.2.0 [Information Systems]: Database Management

General Terms

Database Research, Machine Learning

Keywords

Database Research, Machine Learning, Panel

1. INTRODUCTION

Machine learning seems to be eating the world with a new breed of high-value data-driven applications in image analysis, search, voice recognition, mobile, and office productivity products. To paraphrase Mike Stonebraker, *machine learning is no longer a zero-billion-dollar business*. As the home of high-value, data-driven applications for over four decades, a natural question for database researchers to ask is: *what role should the database community play in these new data-driven machine-learning-based applications?*

The last few years have seen increasing crossover between database research and machine learning. But is this crossover a wise choice for database research? What are the opportunities and the costs of this approach to industry, to the future of database research, and to academics? Do database researchers have something to contribute to this trend? These two areas have dissimilar traditions in both research, intellectually, and in industry, so bridging the gap between the fields is likely to require considerable effort. *Is it worth it?*

2. QUESTIONS TO CONSIDER

We consider how, why, and in what way the database community could make contributions at the intersection of machine learning and databases.

What are the research opportunities and pitfalls for database researchers in these machine-learning applications?

- What are the most interesting research problems at this intersection? Are there core intellectual problems in machine learning that can only be solved with researchers from both sides? Or Are the problems all data-janitor work? If it is data janitor work, is it sufficiently interesting janitorial work to examine in research?
- Is there anything fundamentally different about building database systems that use machine learning or are designed to support machine learning? Or are these new systems just the same old thing rebranded with sexier packaging?
- To attract partners in the machine learning side of the world, we need to be viewed as providing intellectual value. What do database people know that is useful to machine learning? At which level is our knowledge useful? Should we regard machine learning as a black box? Should we apply our ideas inside the black box? Should we build systems that make the black box happy? Where is the most bang for the buck?
- Do we need a new conference on ML+Databases? Or is SIGMOD or KDD the right place?
- What is the risk to the database community if database people build machine learning tools? Could this lead to us becoming a “me-too” community, i.e., a lagging—rather than a leading—indicator? Or is this risk higher if we don’t jump on the machine learning bandwagon like other fields, notably NLP and Computer Vision?
- Can we teach old dogs new tricks? Does working at the intersection of machine learning and databases require that database researchers learn an entirely new set of skills? In contrast, while Database research is applied to and often driven by business, there are few

MBAs in the community. Where are the key intellectual differences with machine learning that make this skills gap more or less challenging?

Database research and the database industry have benefited from a tremendously close working relationship. Where is the momentum for database research and machine learning in industry?

- Much of the excitement about machine learning has come from the large web companies, e.g., Google Brain, Twitter, Facebook’s DeepFace, or Microsoft’s platforms. However, databases have traditionally had an impact in building commodity data platforms. Will machine learning be a major part of successful commodity data platforms?
- Data mining tools like clustering and support vector machines have been part of the database stack for a decade or more. The conventional wisdom from those who built them seems to be that “no one used them.” To what extent is that true? Is there a difference this time around? If so, what is the difference? What remains the same?

- There is a push in industry to marry statistical analytic frameworks like R and Python with almost every data processing engines from classical engines like Oracle’s ORE to newer engines like Cloudera’s Impala or Spark SQL. Is this really machine learning or is this actuarial science? Is the momentum in the enterprise tools or tools for people about high-end machine learning artifacts or making SAS-style products into a commodity?
- Machine learning is generally seen as a new workload for databases. But ML has had its impact on other fields mainly as a technical solution to long-standing problems. What are the long-standing database problems where ML can be more aggressively applied?

In academia, we often see that the “best students” want to study machine learning. Do we risk losing the best minds if we do not embrace machine learning? It can seem like the academic landscape is shifting around machine learning. Informally, you seem to be more likely to interact with a student who has taken machine learning than compilers. What does this mean for how we teach, recruit, and train students?