INIT

Big Data

Big data is everywhere. In just about every part of the modern world, scientists and engineers are developing new ways to measure events. Whether it's sensors, traffic cameras, sales data, Web usage, gene expression, or just about anything else, we have entered an age of truly massive data. Why do we collect this data? It's simple—to learn. We want to make predictions, quantify reality, or understand the past to optimize the decisions we make.

Massive data leads to many challenges for computer scientists. We're recording petabytes of data every day. Before we even think about learning from it, how and where do we store it? What kinds of systems do we build to retrieve and analyze the data? Can we develop theoretical guarantees about the optimality of these systems and strategies? Even if we can store the data, how do we learn from data sets that we cannot hold on a single computer or even in many computers? Can we learn from data on the fly? Moreover, our data is heterogeneous: We are observing social networks, ad click-throughs, gene sequences, protein concentrations from cells, as well as confidential personal data that must be kept secret. How do we adapt our systems and algorithms for all kinds of data? These are just some of the exciting challenges facing the big data community.

For such a diverse topic like big data, it is nearly impossible to provide a comprehensive picture. Instead, in this issue we try to highlight some recent developments organized into three main themes: the theoretical foundation providing models and algorithms for reasoning about various data processing tasks, the large-scale computer systems for handling big data, and the range of applications and analyses enabled by big data from a variety of scientific domains. It has been an interesting time for big data with innovations coming simultaneously from theorists, system builders, and scientists or application designers. We hope to provide readers with an idea of the interplay between developments in these three different communities, how ideas and priorities in different communities interact, and together drive forward the development of big data analysis.

THEORY

Opening the issue is an introduction to the theo-

Interest in big data has given rise to a lot of recent interest in building systems to support queries and transactions over massive quantities of data.

The U.N. Global Pulse project is researching the use big data can have in understanding global development.

One of the datasets they are currently providing to researchers is anonymized cell phone records from Cote d'Ivoire, to explore what these might reveal about society and development there.

retical work for modeling and studying challenges in big data. Jelani Nelson introduces us to the world of streaming algorithms where there is a voluminous stream of data passing by. One can only examine each piece rudimentarily and yet is still able to report meaningful statistics about the whole stream at the end.

Next, Ashwin Machanavajjhala and Jerome P. Reiter describe a principled approach to privacy when dealing with big data. They provide examples of common pitfalls and general methods in both statistics and computer science for protecting privacy while still providing the enormous utility of big data.

Ronitt Rubinfeld follows with a task seeming even more incredible: Computing the answer without even looking at the whole input. She focuses on the problem of understanding distributions just from a few samples, in fact, much fewer than the domain size.

To wrap up the theory foundation, Jeff Ullman provides a gentle introduction to designing algorithms for the map-reduce framework for parallel processing of big data, a hugely successful approach for distributed computing in computer clusters with many practical applications.

SYSTEMS

Interest in big data has given rise to a lot of recent interest in building systems to support queries and transactions over massive quantities of data. A number of important technical developments in this arena have happened outside of academia. We have chosen to present three different perspectives from industry: one from a mature company, one from a small startup, and one from a company that is somewhere in between.

From Cloudera, Yanpei Chen and his coauthors—Andrew Ferguson, Brian Martin, Andrew Wang, Patrick Wendell—provide lessons that can be learned from a small startup on big data and why it makes sense for students to intern in a big data startup.

Our interview with Surajit Chaudhuri from Microsoft Research provides a lens into big data systems design from a company that has been designing database systems for decades.

Raghotham and Rajat from Facebook, which has been in the forefront of the NoSQL big data movement (as it is called), tell us how Facebook designs systems used internally to support queries over the massive quantities of data.

If the industry perspective on building systems wasn't enough, we present an article from Mike Carey, Chen Li, and Vinayak Borkar from UC Irvine, who have been rethinking the design of these big data systems from first principles, and have been making some exciting progress.

It has been an interesting time for big data with innovations coming simultaneously from theorists, system builders, and scientists or application designers.

APPLICATIONS

In such massive data contexts, getting data into a form amenable to analysis and visualization is challenging. Jeff Heer and Sean Kandel write about cutting-edge work that enables data analysts to quickly gain valuable insights from their data.

Social network analysts have been using massive graph data to understand social interactions and behavior. B. Aditya Prakash presents some of the challenges and strategies for studying propagation and immunization in the realm of large social networks.

John Langford from Microsoft Research gives an overview of the challenges in machine learning on big data. He addresses approaches to thinking about learning from data in parallel and some interesting applications.

We also have massive datasets on the cell-by-cell and genome levels. Cliburn Chan reviews the current issues facing the computational biology community and current computational strategies for tackling these problems.

SUMMARY

Overall, through this issue, we are providing a peek into the exciting world of big data—through the lens of theorists, systems designers, scientists and application developers. Indeed, it is undeniable that big data is going to grow in importance in all fields, and will need our expertise. The expertise of an educated bunch of young researchers, scientists, and engineers.

—Andrew Cron, Huy L. Nguyen, and Aditya Parameswaran, Issue Editors