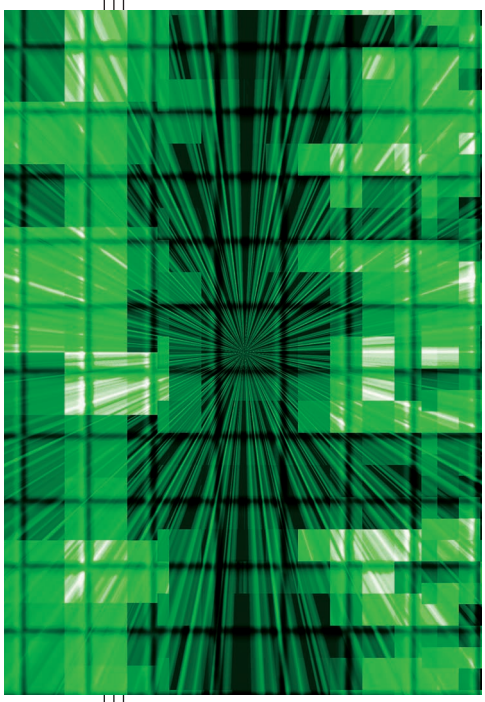

# THE PERVASIVENESS OF DATA MINING AND MACHINE LEARNING

Naren Ramakrishnan
*Virginia Tech*

**A collection of four articles highlights the ever-growing importance of data mining and machine learning in multiple fields.**

Extracting inferences and knowledge from data is the objective of data mining and machine-learning research. Different researchers draw distinctions between what constitutes data mining versus machine learning, and these distinctions are typically ones of emphasis and lineage of research. But in the recent past there has been a healthy cross-fertilization and spillover of authors between the respective communities.

We collect here some of the most promising research, with a view toward highlighting the ever-growing importance of data mining and machine learning in multiple fields. The four feature articles in this issue span applications in recommender systems, sensor networks, automated scientific discovery, and software engineering.

## RECOMMENDER SYSTEMS

We love to complain about our recommender systems. Everyone has a favorite example of how some online site gave a poor recommendation.

About three years ago, Netflix, the movie rental company, announced a $1 million prize competition (www.netflixprize.com) in recommender systems. The charge was to design algorithms that improve upon Netflix's proprietary system by at least 10 percent. Some have hailed the competition as a clever example of crowdsourcing, but there is no denying that Netflix's release of a more than 100-million rating dataset is a bonanza for data-mining researchers. In the past couple of years the competition has created quite a buzz in machine learning/data mining blogs.

As this article goes to press, a multinational team of researchers has reportedly cracked the 10 percent barrier and, barring any last-minute challengers, is expected to win the grand prize. Yehuda Koren, Robert Bell, and Chris Volinsky are part of this winning team. There are many ingredients to their solution, but in "Matrix Factorization Techniques for Recommender Systems," they reveal one of their secret sauces, namely *matrix factorization techniques*. These techniques provide a formal basis for incorporating many preference models, and the authors show how this approach has helped them make steady progress in the Netflix competition. *Computer* wishes the authors and their team continued success in the days ahead.

 Published by the IEEE Computer Society 0018-9162/09/$26.00 © 2009 IEEE

## SENSOR NETWORKS

We are overloaded with data—so goes the saying. But in some domains data collection is actually costly, and we must judiciously place "probes" to intelligently gather information for subsequent data interpretation and data mining.

Since we have control over data collection, the pertinent questions are: Where should we collect data? and What type of data should we collect? In "Optimizing Sensing: From Water to the Web," Andreas Krause and Carlos Guestrin present an approach for optimizing sensor placement and, as the article title suggests, apply it to several domains. Where should we measure contaminations in a water distribution network? Which blogs should we read to keep track of the news?

The authors show how the concept of *submodular functions* captures many characteristics of these applications and design their algorithms using this observation. Interestingly, they are also record holders in a sensor placement competition called Battle of the Water Sensor Networks.

## AUTOMATED SCIENTIFIC DISCOVERY

The third piece will sound a bit futuristic. Ross King and colleagues describe a "robot scientist" tasked with planning and executing functional genomics experiments. One of the current questions in biology is how to understand the functions of all the genes in "model" organisms, such as baker's yeast (*Saccharomyces cerevisiae*). A typical way to approach this question is to "knock out" the gene from the organism and then grow the mutant and observe its physical characteristics (phenotype).

The system described in "The Robot Scientist Adam" undertakes the full cycle of planning these experiments, directing laboratory equipment to perform the experiment, analyzing data, forming/revising a hypothesis, and suggesting new experiments. This work thus shares with the previous article the emphasis on integrating data collection and learning from experiments. In this and their publication in *Science* earlier this year, the authors also present some of the new discoveries made by their system.

## SOFTWARE ENGINEERING

The fourth article focuses on an emerging hotbed of data—software engineering projects. Tao Xie and colleagues describe how practically everything a software engineer deals with—code bases, program call graphs, bug reports, code documentation—is now so plentiful that it can be mined. The applications also span the life cycle of software engineering activities from specification, programming, and debugging, to software maintenance. It nicely reinforces this issue's pervasiveness theme of data mining/machine learning.

***Naren Ramakrishnan*** *is a professor and the associate head for graduate studies in the Department of Computer Science at Virginia Tech. His research interests include computational science, especially computational biology, mining scientific data, and information personalization. Ramakrishnan received a PhD in computer sciences from Purdue University. He is a member of the IEEE Computer Society, the ACM, and the AAAI, and serves on* Computer*'s editorial board. Contact him at naren@cs.vt.edu.*