



Careers in DATA MINING: The Next 10 Years

Richard D. DeVeaux and Lyle H. Ungar

Where is the field of data mining going in the next 10 years, and what kind of opportunities will it provide statisticians? These are great questions. But, we all know that prediction is risky business. As Nils Bohr said, "Prediction is very difficult, especially if it's about the future." So, we won't speculate on which class of algorithms might win out or which type of modeling will be most common. Besides being an impossible task, we think it's the wrong question. Instead, we'll focus on the aspects of data mining that are unlikely to change and why data mining will continue to provide great opportunities for statisticians, especially those just starting out in their careers.

By the early 1990s, the term “data mining” had begun to shed its pejorative connotation in statistics and had come to mean the process of finding information in large data sets. Everyone seems to have their favorite definition of exactly what data mining means to them, but Wikipedia currently defines it as “nontrivial extraction of implicit, previously unknown, and potentially useful information from data,” and that’s good enough for our purposes. We’ve witnessed the gain in legitimacy of this field as measured by the number of academic conferences and journals either relevant to or wholly dedicated to data mining. This has been great for statistics, as it’s brought bright minds in other fields to focus on problems traditionally thought to be statistical issues.

Statisticians and computer scientists discovered some of the many algorithms and methods used in data mining, like decision trees, independently at roughly the same time. Two of the big ideas of the past decade, bagging and boosting, come from a statistician and a pair of computer scientists, with further contributions from both areas. Large-scale Bayesian models are so firmly intertwined between the two fields that one cannot say in which area they are. It is exactly this interplay between computer science, data management, machine learning, and statistics that has contributed to the vitality of data mining.

It is an obvious cliché, but still a key fact, that the amount of available data to analyze and the power of the computers to analyze it are both growing exponentially. The consequence for statisticians is a rapidly growing set of interesting and important areas in which to work. And that’s true not only methodologically, but practically. The areas where data mining is used are unbelievably broad. So, why do we say “data mining” and not “statistics”? Because non-statistical aspects, such as the storage and retrieval of data, are a key part of many projects, and the most important part of solving the problem may be in how users access and present the data. The statistical algorithms are sometimes obvious and relatively unimportant.

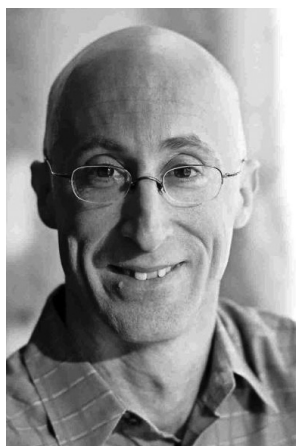
The common thread to taking advantage of data mining’s breadth is being open to ideas from different fields. How do I predict which of millions of possible search results or ads will be of most interest to a user? How would I determine what single

nucleotide polymorphisms in your genome might cause a drug to give you side effects? What could I usefully (and legally) predict if I had the text messages and locations over time of every cell phone used at my university? Answering each of these ques-

Authors



Richard D. DeVeaux is a professor of statistics at Williams College. Last year, he served as the William R. Kenan, Jr., Visiting Professor for Distinguished Teaching at Princeton University. He holds degrees in civil engineering (BSE, Princeton), mathematics (AB, Princeton), dance education (MA, Stanford) and statistics (PhD, Stanford).



Lyle H. Ungar is an associate professor of computer and information science (CIS) at the University of Pennsylvania. He also holds appointments in several other departments in the engineering, medicine, and Wharton schools and serves as associate director of the Penn Center for Bioinformatics (PCBI).

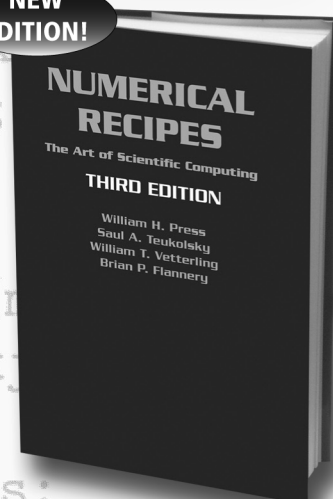
CAMBRIDGE

**Coming this Summer...
Order Today!**

“This monumental and classic work is beautifully produced and of literary as well as mathematical quality.”

—*Computing Reviews*,
on the previous edition

**NEW
EDITION!**



NUMERICAL RECIPES

The Art of Scientific Computing
Third Edition

William H. Press, Saul A. Teukolsky,
William T. Vetterling, and
Brian P. Flannery

- New edition of this groundbreaking work contains over 25% more material – including new chapters, new sections and upgrades throughout.
- Source Code CD-ROM contains all of the routines discussed in the book, in ANSI/ISO C++ source code (can be used with almost any C++ vector/matrix class library).
- Go to www.nr.com for more general information about licenses, and to www.cambridge.org/us/numericalrecipes to learn more about the book and Source Code CD-ROM.

Book/Hb/\$80.00/ 978-0-521-88068-8

Source Code CD-Rom/\$80.00/978-0-521-70685-8

Note: CD-ROM contains source code only and does not include text of the book
Numerical Recipes, Third Edition.

**BUY THE BOOK AND SOURCE CODE CD-ROM
TOGETHER AND SAVE!**

Hb with CD-ROM/\$140.00/978-0-521-88407-5



CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org/us
1-800-872-7423

tions—and many more—requires far more than mastery of random forests, conditional random fields, or least angle regressions. It also requires understanding the problem and what data are available (or could be obtained). It requires working closely with database or computer science experts to process the data and devise and implement efficient algorithms. And it requires not just talking to, but really communicating and working closely with, physicians or lawyers. In short, it requires data mining.

To contribute effectively to a data mining problem, you have to be willing to learn new ideas from other fields. You may not have to become an expert in data architecture, object-oriented programming, or algorithm design, but you will certainly have to work with people who are. These complementary skills are essential to a successful data mining team. In data mining, getting the right data out of databases, increasingly from emails and web pages, requires programming expertise. Efficiently processing large data sets requires some knowledge of algorithmic design—one of

us recently built a regression model with more than a million terms in it, and then applied it to every page on the World Wide Web. Equally important, however, is technical statistical knowledge, such as knowing how to estimate the cost of deleting terms from a model or what the geometry of high-dimensional spaces looks like. The interesting problems require large teams of experts, and statisticians need to work with all of them—whether they're computer scientists, linguists, neuroscientists, economists, or physicians.

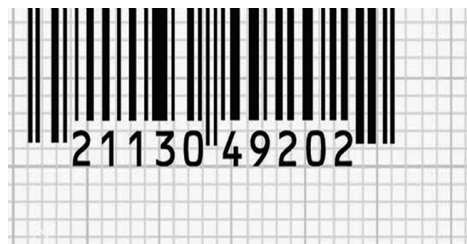
Statisticians have an enormous role to play as long as we build on our strengths (e.g., the honest analysis and interpretation of data) and work on our shortcomings (e.g., the ability to communicate effectively with both the other analysts and the client interested in the answer). Statisticians, unlike people trained in other fields, have a deep awareness of uncertainty and the role it plays in decisionmaking. It's sometimes surprising how little appreciation for variation and uncertainty experts in other fields have. One of us once worked on a team looking

“Being an effective statistician on a data mining team takes all the technical savvy of a well-trained statistician, plus effective communication and teamwork skills.”

at predicting which policies might result in the largest losses for an insurance company. To build a predictive model, we were given tens of thousands of mature, inactive policies; hundreds of descriptive variables about them; and the final cost to the insurance company. The computer scientist on the team couldn't understand what the 'problem' was because he, by sorting on various criteria in this large relational database, could easily find the highest loss cases and didn't understand why that wasn't the answer. Getting across the basic idea of generalizing from one specific data set to a general population turned out to be the biggest challenge of the entire data mining effort.

Statistical ideas that seem so natural to us are not necessarily natural to these other experts, and the diplomatic communication of some of the dangers of overfitting, type I errors, and violation of model assumptions is essential. Often, incorrectly learned statistical ideas need to be overcome. Sometimes, these beliefs are deeply ingrained. A chemist we worked with learned that $n=22$ is the 'correct' sample size. A top notch research physician spent months trying to get around the fact that if one protein out of 20 tested was significant at the 0.05 level, it doesn't mean much. And often, we're asked such questions as how we know whether the data set at hand is a "statistically significant" sample.

Statisticians, unfortunately, often make equally egregious mistakes by failing to understand the real problems in data mining. In a regression model one of us helped build, we used tens of millions of observa-



PASS


**Power Analysis
and Sample Size**

GESS


Microarray Analysis

NCSS

**Statistical Analysis
and Graphics**



sales@ncss.com
1-800-898-6109



Order Today at:
www.ncss.com

tions and millions of predictors (features, as the team called them), plus interactions between those features. Asking whether a feature was statistically significant was irrelevant; the question was: Given the constraints posed by the amount of computational power available, what is the best subset of features to retain in the model? Computing standard errors on the predictions of the model were irrelevant; the key question was how well the models would do on a data set (drawn from a different distribution) for which we did not know the true responses. Refusing to say anything by hiding behind distributional assumptions and lack of asymptotics wasn't going to win us any friends. New tools also were needed to examine the model. It is hard to look at each of half a million coefficients (or other regression diagnostics) to refine and improve the model. New problems demand new approaches. These approaches require teams that understand both statistics and efficient algorithm design. Companies such as Google, Yahoo!, and Microsoft have hundreds of computer scientists building regression and Bayesian models, yet these same companies have only dozens of statisticians. Many data mining projects have no statisticians on them. We have to step outside the statistics box and show flexibility in what we're willing to learn in order to ameliorate this situation.

Statisticians have a lot to bring to the data mining effort and a lot to benefit from it. A disgruntled reporter, stuck with covering the JSM in Chicago once wrote, "Now I know who statisticians are. They're people with the skills to be actuaries, but not the charisma." We have to make sure we can't replace actuary with data miner in that sentence. Let's not lose time by arguing whether the problem is purely statistical enough to be interesting. Instead, let's use the interest in data mining to get everyone to think about the information in large data sets. We can't afford to lose data mining completely to nonstatisticians. We need to be part of the data mining effort. Being an effective statistician on a data mining team takes all the technical savvy of a well-trained statistician, plus effective communication and teamwork skills. No matter what algorithms are developed in the next 10 years, this aspect of effective data mining won't change. And the demand will only continue to grow. It's a wonderful opportunity for statisticians. ■

WWW Resources for Statisticians

STATpages.net

This award-winning site was put together by volunteers and compiled by John C. Pezzullo, a retired associate professor in the departments of pharmacology and biostatistics at Georgetown University in Washington, DC. This page offers links to free software packages you can download and a description of each program. Other sections contain links to statistical books and manuals, demos and tutorials, and information about the site itself.

r-project.org

This site includes information about how to download R, a free software for statistical computing and graphics that runs on a variety of UNIX platforms.

bettycjung.net/Statpgms.htm

This site contains links to general statistical software sites, data sets, and links for specific data management statistical software packages.

www.ourworld.compuserve.com/homepages/Rainer_Wuerlaender/statwww.htm

This page was written by Rainer Würländer, a statistical consultant and IT manager. It offers links to resources in statistics, newsgroups and mailing lists, statistical associations and departments, statistical software, statistical quotes, and a directory of professional statisticians worldwide.

http://directory.google.com/Top/Science/Math/Statistics

This Google directory lists categories such as Bayesian analysis, people, statistical consulting, job opportunities, and software. Each category includes a list of web pages that are ranked in order of times the site is linked.

The New Jersey Chapter of the American Statistical Association and the Mathematics Department of Kean University are sponsoring a half-day workshop focusing on potential careers for statistics and mathematics majors. This short workshop intends to provide useful information to current and prospective students, as well as to professors about the career opportunities in statistics and mathematics disciplines.

Mathematics and Statistics Career Day at Kean University, Downs Hall

Friday, September 21, 2007

2:00–5:00 p.m.

Presented by

Tom Capizzi, VP of statistics/programming/data management
(Schering Plough) • Steve Ascher, senior director (Johnson & Johnson) • Eswar Phadia, professor (William Paterson University)