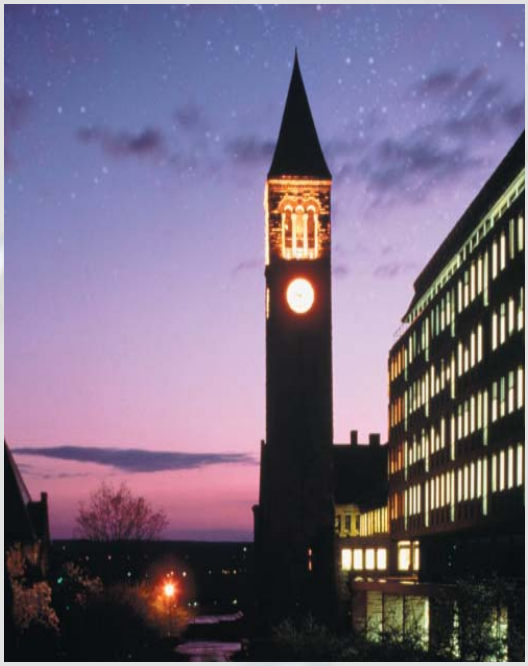
Data mining, today and tomorrow

We leave digital puddles wherever we go. Buy something at a supermarket, and your market basket gets added to the grocery chain's data warehouse for purchase-behavior analysis. Visit a Web site, and your interactions may be used to personalize future interactions. The amount of stored data grows about 30% every year. How can useful information be extracted from this ever-growing ocean of data?

Data mining is the science of discovering new information, such as unknown patterns or hidden relationships, from huge databases; it is the science of finding knowledge that you were unaware of before. "Think of it as database queries on steroids," says CS professor Johannes Gehrke, whose group has developed some of the fastest data-mining algorithms. "Traditional database queries let you specify exactly what records to retrieve. In data mining, the computer finds interesting gold nuggets without you pointing a specific query at them."



Cornell's Olin Library and McGraw Tower.

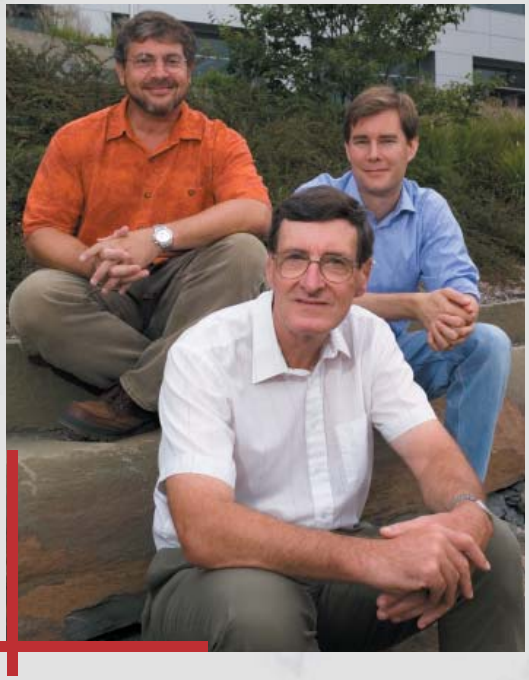
Data mining is the science of finding knowledge that you were unaware of before.

Three ongoing projects illustrate different research challenges in data mining. First, data-mining algorithms most often search humongous combinations of possibilities. For example, consider finding out what items shoppers frequently purchase together in Wal-Mart. Assuming that Wal-Mart stocks a few 10,000 items, there are about 10^{3000} possible combinations of items to investigate! Also, a 100-terabyte database does not fit into memory, and access to data on disk can be five orders of magnitude slower. Fast search and scalability of algorithms are needed; they have been the focus of research in the last decade, leading to much progress on scalable data mining algorithms.

Scalability today presents an enormous challenge. CS researcher Alan Demers and Gehrke are working with Jim Cordes of the Astronomy Department on the design and implementation of an analysis infrastructure for a new census of pulsars in the Milky Way Galaxy. The data will be collected at the Arecibo Observatory in Puerto Rico. "The data rates and processing requirements for the pulsar survey are truly astronomical," says Gehrke. The total raw data, which will take three to five years to acquire, will be about one petabyte —14 terabytes of data will arrive every two weeks via "Fed-Ex-Net" on USB disk packs, requiring the processing of one TB of data per day. A recent $2M research infrastructure award has allowed the team to build the necessary computing infrastructure at the Cornell Theory Center.

A second challenge is to mine data with missing or wrong entries. CS professor Rich Caruana and researcher Mirek Riedewald are working with scientists from the Cornell Lab of Ornithology on analyzing large citizen-produced datasets. Every year, tens of thousands of volunteers report sightings of birds to the Cornell Lab of Ornithology, creating one of the largest and longest-running resources of environmental time-series data in existence. Its analysis could reveal long-term changes in ecosystems due to human intervention; for example changes in farming practices have been shown to affect bird abundance over time. But mining the data is challenging. Volunteers often leave some entries in bird report forms empty, novice observers may confuse bird species, and other variables such as habitat, weather, human population, climate, and geography have to be considered when estimating the true abundance of a species. "Compensating for bias in the collected data is a major challenge, and each observation could be differently biased," says Caruana.

A third challenge is the enormous complexity of today's databases. For example, consider the Web. CS professors Bill Arms, Gehrke, Dan Huttenlocher, Jon Kleinberg, and Jai Shanmugasundaram are building a testbed that will enable the study of temporal dynamics of the Web over time. The team

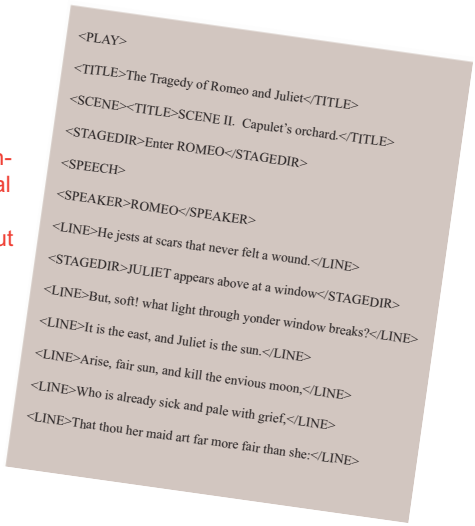


Interests in data mining and the Web —itself a humungous database— bring Rich Caruana (left), Bill Arms (center), and Johannes Gehrke together.

will obtain the 40 billion Web pages archived by the Wayback Machine, the time machine of the Internet. The team will also receive new 20–terabyte snapshots of Web crawls every two months. This collection will enable the research community, for the first time, to evaluate models of Web growth and evolution at a wide range of different time scales. "The combination of content, link structure, and temporal evolution creates an immensely complex dataset," says Arms. "With this data and associated data-mining tools, we will be able to tackle really big questions, for example how new technologies, opinions, fads, fashions, norms, and urban legends spread over time."

"The beauty of working in this area is that you have discovery at two levels," says Gehrke. "You develop interesting new computer science methods, and you find nuggets by applying these to real datasets."

The marriage of structured and unstructured data



Most digital documents contain a mixture of structured and unstructured data. For example, online versions of congressional bills in the Library of Congress database contain not only the names, dates, and sponsors of these bills (structured data) but also the text of the bills and hyperlinks to related documents (unstructured data). The same is true for the Internet, which contains database-backed Web sites (structured) and static HTML pages (unstructured). Similarly, online versions of Shakespeare's plays contain information about acts, scenes, and names of persona (structured) and the text of the play (unstructured).

Current data management systems such as relational database systems and information retrieval systems do not provide a unified way to handle both structured and unstructured data.

Recent advances led mostly by CS researchers are bringing together the separate worlds of structured and unstructured data. CS professor Jayavel Shanmugasundaram and grad student Chavdar Botev, in collaboration with Sihem Amer-Yahia at AT&T Research Labs, have developed a new language and system architecture for managing both kinds of data. The core of their contribution is a new query language, TeXQuery, which enables users to seamlessly query XML documents that contain a mix of structured and unstructured data.

TeXQuery has had a rapid and profound influence on the data management industry. The World Wide Web Consortium (W3C), the body that developed fundamental standards such as HTML and XML, has adopted TeXQuery as the precursor to their standard *XQuery Full-Text* language for querying structured and unstructured XML data. Shanmugasundaram and Botev serve as invited experts at the W3C to help shape the evolution of the language. Rarely has a research idea been transferred to a standards body in such a short time.

Jai Shanmugasundaram and his colleagues are the matchmakers for structured and unstructured data.

The potential impact of TeXQuery is enormous. The Library of Congress has recently started an effort to convert its data into XML to make it searchable using XQuery Full-Text. All major data management vendors have announced plans to implement the language in future releases of their systems.

Structured and unstructured data are engaged, and their marriage date will be set soon. And CS researchers are the matchmakers!

Ken Birman starts a company based on Isis. Isis is used extensively on Wall Street and in telecommunications and VLSI FAB systems. Today, Isis is still the core technology in the New York Stock Exchange (every trade and every quote since 1993 ...), the Swiss Exchange, the French Air Traffic Control system, the US Navy's AEGIS warship, and the Florida Electric and Gas SCADA system.

Tom Coleman and Charlie Van Loan publish the *Handbook for Matrix Computations* (SIAM).

Tim Teitelbaum and former student Tom Reps publish two books on the Synthesizer Generator, with Springer-Verlag.

Bard Bloom, Steve Vavasis join.

1989

The Computer Science Board, chaired by Gries, changes its name to the Computing Research Association (CRA), opens an office in Washington, and works to represent the national interests of computing research.

John Hopcroft authors a report for the NSF Advisory Committee for Computer Research (with Ken Kennedy). "Computer Science: Achievements and Opportunities" helps set the direction of the NSF computing research funding.

Gerry Salton is Chair-Elect of Section T of the AAAS (American Association for the Advancement of Science). Section T concerns Information, Computing, and Communication.

Tom Coleman becomes Director of the Cornell Advanced Computing Research Institute, a unit of the Cornell Theory Center. The interdisciplinary institute is concerned with scientific computation research and its application to engineering and scientific problems.

Gerry Salton receives the ASIS Award of Merit, the American Society of Information Science and Technology's highest honor, bestowed annually to an individual who has made a noteworthy contribution to the field of information science.

John Hopcroft receives an honorary doctorate from Seattle University.

Bob Constable and student Doug Howe publish *Implementing Metamathematics as an Approach to Automatic Theorem Proving* (Elsevier Science).

Gerry Salton publishes *Automatic Text Processing* (Addison Wesley).