

BusinessWeek		Archives				Click for Feb. 3, '97 Issue
REGISTER	BW HOME	BW CONTENTS	BW PLUS!	BW DAILY	SEARCH	CONTACT US

Information Processing

COAXING MEANING OUT OF RAW DATA

Software can now find patterns never seen before

It's the bane of modern business: too many data, not enough information. Computers are everywhere, accumulating gigabytes galore. Yet it only seems to get harder to find the forest for the trees--to extract significance from the blizzard of numbers, facts, and stats.

But help is on the way in the form of a new class of software technology known broadly as data mining. First developed to help scientists make sense of experimental data, this software has enough smarts to "see" meaningful patterns and relationships on its own--to see patterns that might otherwise take tens of man-years to find. That's a huge leap beyond conventional computer databases, which are powerful but unimaginative: They must be told precisely what to look for. Data-mining tools can sift through immense collections of customer, marketing, production, and financial data and, using statistical and artificial-intelligence techniques, identify what's worth noting and what's not.

The payoffs can be huge, as MCI Communications Corp. is learning. Like other phone companies, MCI wants to keep its best customers. One way is to identify early those who might be considering jumping to a rival. If it can do that, the carrier can try to keep the customer with offers of special rates and services, for example.

How to find the customers you want to keep from among the millions? MCI's answer has been to comb marketing data on 140 million households, each evaluated on as many as 10,000 attributes--characteristics such as income, lifestyle, and details about past calling habits. But which set of those attributes is the most important to monitor, and within what range of values? A rapidly declining monthly bill may seem like a dead giveaway, but is there a subtler pattern in international calling to be looking for, too? Or in the number of calls made to MCI's customer-service lines?

To find out, MCI regularly fires up its IBM SP/2 supercomputer--its "data

warehouse"--which identifies the most telling variables to keep an eye on. So far, the SP/2 has compiled a set of 22 detailed--and highly secret--statistical profiles based on repeated crunching of historical facts. None of these could have been developed without data-mining programs, says Lance B. Boxer, MCI's chief information officer.

Data mining in itself is a relatively tiny market: Sales of such programs will grow to maybe \$500 million by 2000, from \$50 million this year, estimates Two Crows Corp., a Potomac (Md.) market research firm. But the technology is critical in getting a big payoff from what information-technology executives think will be an immensely important growth area in coming years: data warehousing. These are the enormous collections of data--sometimes trillions of bytes--compiled by mass marketers, retailers, or service companies as they monitor transactions from millions of customers. Data warehouses, running on ultrafast computers with specialized software, are the basis on which companies hope to operate in real time--instantly adjusting product mix, inventory levels, cash reserves, marketing programs, or other factors to changing business conditions. The market for data-warehousing hardware, software, and services will grow from \$2 billion in 1996 to \$12 billion in 2000, according to Meta Group, a consulting firm.

DRAGNET. And data mining will help make that investment pay off. How? By, for example, catching crooks. Telephone companies, credit-card issuers, and insurers are mining their data warehouses for subtle patterns within thousands of customer transactions to identify fraud, often just as it's happening. One unidentified U.S. cellular phone company is using Silicon Graphics Inc.'s MineSet software to dig through mountains of call data and pinpoint illegally cloned cell-phone ID numbers.

Manufacturers can mine data collected from factory-floor sensors and learn just where an intermittent assembly error is causing a defect that shows up only months after an appliance goes into use. And once shopping on the Web takes off, reams of data about customers' behavior, tastes, and interests will be available for merchandisers to mine and react to nearly instantly.

The list goes on. "A huge opportunity is opening up," says Usama M. Fayyad, who helped create one of the earliest data-mining systems at the Jet Propulsion Laboratories in Pasadena, Calif. It helped identify quasars from trillions of bytes of satellite data. Now, he's working at Microsoft Corp.'s research lab, where he's looking for new uses of the technology. He predicts that within a year or two, when computers everywhere maintain sizable databases, data mining will aid even small businesses such as restaurants and local accounting firms. Because of Microsoft's aggressive plans for its SQL Server database software, Fayyad says that "Microsoft is a good place to bring this technology to the millions."

For now, though, data mining is serving only a few giants, such as U S West Inc. Like other phone companies, it's enjoying strong demand for second and third residential phone lines, which customers want for their teenagers, fax machines, and PCs. But the carriers don't want to sink the money into new network switches and trunk lines in a particular area unless they can be

fairly certain that the orders for extra lines will really materialize. Furthermore,

U S West says it wants to pinpoint customers who will not only respond to introductory offers but also will keep their second lines long enough for the carrier to make a profit.

IDEAL FAMILY. To find those people, U S West uses a program called PALMS. It designed the program with AT&T's NCR computer unit and Sabre Decision Technologies, a unit of AMR, which owns American Airlines. Running on a powerful NCR parallel-processing computer, PALMS first spent hours sifting through a sample of a few thousand customer records from the Phoenix area. Each record contains as many as 250 items about a household: income bracket, monthly phone bill, number of repair calls in the past year, and its history of trying and keeping such services as call-waiting, for instance. The result is a statistical model of the ideal prospect.

Then PALMS used that model to search through millions more customer records--almost 1 trillion bytes of data. By correlating data about the location of each home, the location of U S West's trunk lines, and the capacity of local switches, the program identified clusters of prospects--households that fit the model and that U S West could provide service without significant expense. From the first direct-mail campaign, which ran from Nov. 4 to early January, U S West has enjoyed a response rate equal to that of a broadcast campaign costing "several million dollars" more, says Gloria A. Farler, executive director of marketing intelligence. PALMS even calculates when a direct-mail campaign will peak so the carrier can cut back before the response rate craters.

That's a giant leap forward from what conventional database setups do. Leading ones, from companies such as IBM, Oracle, Informix, and Sybase, are good at swiftly locating and updating any specific item--your savings balance when you make a withdrawal, for instance. Or they can retrieve in a flash all items that meet specific criteria--New Jersey males who leased red Fords in 1993, say. The key is first cross-indexing all of the records according to a selected set of attributes.

BORN TO CRUNCH. In contrast, data-mining programs rely much less on indexes and may take hours or even days to return an answer. By repeatedly sorting records according to varying sets of attributes--sometimes even randomized sets--these programs attempt to categorize the records and identify subtle correlations between their many variables. "Humans are good with a small number of variables," says Fayyad--no more than about eight. "But now, we're seeing databases with hundreds or thousands of variables. As a human, that's a lost cause. But it's what machines were born to do."

Still, it's important to keep in mind that a machine is still a machine--even if it embodies this new data-mining technology, warns Herb Edelstein, president of Two Crows. "Data mining is not a miracle," he says. The

technology, still in its early stages, most often comes back with flakes of informational gold, not nuggets.

Companies that are experimenting with data mining are quickly discovering that it's important to understand what they're looking for and what type of tool will work best. Data-mining software comes in many different forms: Some look for clusters of like items, for example, while others search for anomalies. Without a proper match between tool and data, though, a program may come up with useless insights--that senior citizens don't buy rap music, for instance--or overlook those that really matter. To help customers cover all bases, IBM, Silicon Graphics, and Thinking Machines have assembled suites of different mining tools. Smaller UltraGem and Information Discovery Inc. sell just one approach.

LOAN RANGER. Already, a wide range of companies are putting on their data-mining helmets to find the information that can make a difference to their bottom lines.

UltraGem, a San Francisco startup, has been working with an unidentified bank to predict the profitability of adjustable-rate mortgages, for example. UltraGem's software first analyzed information concerning more than 100,000 loans. The data ranged from the age and zip code of the customers to the source of their loan and whether or not it was converted from a previous loan. The result: a set of rules for identifying loans likely to yield the highest profits--rules that assembled combinations of variables "that were beyond what any human mind could figure out," says UltraGem President Steven A. Vere. Now, the bank can predict such things as who might prepay or who might become delinquent and then adjust rates and fees accordingly.

One sign that data mining has arrived: Wal-Mart Stores Inc., whose pioneering use of massive transaction databases revolutionized retailing, is plunging in. Since the 1980s, Wal-Mart has collected volumes of cash-register data from its stores each night. But despite running one of the most powerful computers around--a specialized machine from NCR Corp., Wal-Mart has been unable to use all those data. Faced with a mind-boggling 700 million potential forecasts to calculate--one for each item in 2,700 stores--it was forced to lump stores into regions and products into categories.

In the past year, Wal-Mart has turned to a data-mining system from NeoVista Solutions Inc., formerly called MassPar. Harnessing hundreds of processors to the task, it's helping Wal-Mart predict demand for individual items in specific stores. And it's improving the accuracy of Wal-Mart's market-basket analyses, which look at the combinations of items that consumers tend to buy during one visit. Wal-Mart officials decline to discuss the technology. But as NeoVista CEO John M. Harte puts it, "the devil really is in the details." And data mining is what's helping dig that devil out.

By John W. Verity in New York

BusinessWeek

Updated June 15, 1997 by bwwebmaster

Copyright 1997, Bloomberg L.P.

[Terms of Use](#)