TIME
ONLINE EDITION

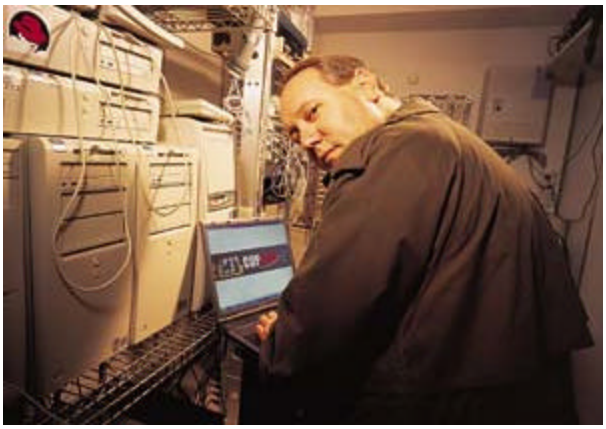
global BUSINESS

Data Miners

New software instantly connects key bits of data that once eluded teams of researchers

By DANIEL FRANKLIN/ WASHINGTON

Monday, Dec. 23, 2002



DANIEL PEEBLES FOR TIME

Americans got a glimpse of how such a system might work this fall during the Washington-sniper investigation.

Two weeks into the shootings, Knowledge Computing, an Arizona company whose COPLINK system has integrated police databases. . .

Quick: What do Hamas terrorists have in common with Martha Stewart? No, we're not talking about their public-approval ratings. Rather, both may have drawn unwanted scrutiny in part because of the same piece of software.

The data-mining algorithms of ClearForest, based in New York City, are at work within both Israeli security agencies and NASDAQ. Israel uses them to drill for hidden connections among suspected terrorists: say, a pattern of phone calls shortly before each of several suicide bombings. NASDAQ uses the same software to detect block trades of stock quietly placed just before the release of company news — including sales by relatives of ImClone's founder, Sam Waksal, who this fall pleaded guilty to insider-trading charges, and his friend Martha Stewart, who remains under investigation (and has denied any wrongdoing).

Both NASDAQ and Israel's security services are sprawling organizations, bombarded daily with terabytes of information, any bit of which may prevent a catastrophe, whether measured in lives or in retirement savings lost to fraud. And these days, data-mining software, combined with technologies that connect disparate computer systems and databases, is making it possible for everyone from police departments to clothing merchants to global manufacturers to search through ever expanding data warehouses and draw valuable connections that would otherwise be lost to human eyes.

Consider the British aerospace firm BAE Systems. Software developed by Autonomy, based in Cambridge, England, connected BAE's research databases and alerted civilian aircraft engineers to the fact that the wing-construction problem they were working on was also being addressed by the company's military division. Ending this duplication helped the company save millions of dollars.

Even as the application of data mining has spread throughout the retail, manufacturing, pharmaceutical and financial sectors, the lingering tech slump has slowed the technology's growth. This year will go down as the worst in history for information technology, with sales falling 2.3% after averaging a 12% annual growth rate over the past 20 years. Data-mining companies have been among the hardest hit in recent years: MicroStrategy, based in McLean, Va., rode the near messianic predictions of founder and CEO Michael Saylor from a market capitalization of $1 billion to more than $20 billion and back again in less than 18 months.

Few companies attained the heights of MicroStrategy or the ignominy of its fall after it confessed to overstating earnings. But the combination of shrinking corporate IT budgets and a lack of interest among small and midsize companies laid waste a host of firms that once promised to be the brains of the information revolution.

Today, however, companies that excel in connecting the data dots are finding a lifeline in a customer whose IT ineptitude is matched only by its means: the U.S. government, which will spend $53 billion on information technology this year. The

Federal Government's inability to share and analyze information became clear in the months after the 9/11 attacks. An FBI agent in Phoenix, Ariz., who was suspicious of Arab flight-school students was not aware that he had a colleague in Minneapolis, Minn., with the same concerns. Immigration officials didn't know that a Saudi suspected by the CIA of terrorist ties had applied for a visa. "If you look at the lessons learned from Sept. 11," says Ed Sketch, director of Ford's Learning Network, which uses Autonomy software to categorize and distribute information to all its salaried employees, "this technology is slap-bang in the middle because it sorts the relevant from the irrelevant to keep an organization from drowning in data."

It didn't take long for data-management companies to realize that if their software could find links in customer buying patterns and improve retailers' inventory decisions, perhaps it could find links among the government's vast terrorism-related intelligence warehouses and enhance the government's ability to prevent the next attack. After 9/11, many tech companies saw opportunities for both patriotism and profit. Oracle offered to donate the software to create a federal identity database. Siebel Systems CEO Thomas Siebel boasted to a House subcommittee that had his company's software been used by law-enforcement and intelligence organizations before 9/11, "there may have been a different outcome."

Promises of quick fixes have faded as the scale of the government's challenge has become clear. But the early setbacks have not deterred companies from setting up shop in or near Washington to get closer to the action. In the past year, Raytheon and EMC launched a joint government IT unit, and SAP relocated its global public-sector headquarters to the Washington area. And PeopleSoft launched two new products designed for the Homeland Security market.

The biggest winner so far appears to be Autonomy, which recently won a contract initially worth about $3 million to provide the software for information collection, analysis and routing for the 22 agencies that fall under the new Department of Homeland Security (DHS). But few other contracts have been signed thus far, as the feds slog through the swampy process of budgeting, appropriating and procuring. "We've all been waiting for the wheelbarrows of money to show up," says Leonard Pomata, president of the government division of webMethods, based in Fairfax, Va. "Aside from the emergency funding that has been spent on guards and gates and guns ... there hasn't been a significant amount for new initiatives."

That should change in the coming months, once DHS Secretary Tom Ridge has time to survey his new dominion. As details of the government's actions before 9/11 continue to unfold, two immediate needs become clear. The U.S. needs better ways of uncovering hidden connections within the masses of information it collects from different sources. And it needs to make sure that information stored within one agency's database can be shared with the appropriate officials elsewhere.

Just a few years ago, such goals would have been laughable. But innovations in artificial intelligence let the government recognize and quantify links between disparate forms of data. New software languages automatically translate information from different systems into a common tongue. Both advances have vastly expanded the tools available for the war against terrorism.

Throughout the '90s, data mining spread from one industry to the next, enabling companies to know more about customers' needs and to zero in on the characteristics that distinguish the customers they want from those they do not. A credit-card company using a system designed by Teradata, a division of NCR, found that customers who fill out applications in pencil rather than pen are more likely to default. A major hotel chain discovered that guests who opted for X-rated flicks spent more money and were less likely to make demands on the hotel staff, according to privacy consultant Larry Ponemon. These low-maintenance customers were rewarded with special frequent-traveler promotions. Victoria's Secret stopped uniformly stocking its stores once MicroStrategy showed that the chain sold 20 times as many size-32 bras in New York City as in other cities and that in Miami ivory was 10 times as popular as black. Aspect Communications, based in San Jose, Calif., sells a program that identifies callers by purchase history. The bigger the spender, the quicker the call gets picked up. So if you think your call is being answered in the order in which it was received, think again.

The technology that underlies these applications, known as customer-relationship management (CRM), is a subset of data mining and can be used by any organization that needs to quickly analyze massive amounts of data. "More and more often," says Michael Schiff, vice president of Current Analysis, a business-intelligence firm based in Sterling, Va., "you see it becoming 'X'RM, where the relationship being managed could be with a customer or supplier or stockholder or, to go out on a limb here, terrorist."

The potential has not been lost on the intelligence community. When CIA agents came calling on digiMine, a retail data-mining specialist based in Bellevue, Wash., they carried a list of 900 companies that were pursuing data-mining capabilities, says CEO Usama Fayyad. Through a nonprofit venture-capital fund that it finances, called In-Q-Tel, the CIA has invested in several data-mining companies that serve both the private sector and the CIA.

One such firm, Systems Research and Development (SRD), based in Las Vegas, uses an algorithm that looks for what are known as nonobvious relationships to flag a casino if one of its employees appears to have a connection to a known cheater. After cleansing databases of misspelled names and aliases, the software looks through the casino's customer databases, as well as public records it acquires, which may contain criminal charges, addresses and phone numbers. Within 90 days of implementation, a Louisiana riverboat casino caught four employees who were helping friends and family members cheat,

including a roulette-wheel spinner who had shared a phone number with someone the casino had caught placing bets after the ball had landed. The same data-sifting capabilities, says SRD founder Jeff Jonas, can help police and intelligence agencies make sense of Arabic names that may be transliterated in a dozen ways.

The CIA's investments, typically no more than a couple of million dollars a year, give the agency a chance to tailor products to its needs while boosting the survival rate among small but innovative firms. A CIA investment in Intelliseek, based in Cincinnati, Ohio, accelerated the company's ability to give its software multilingual capabilities. In addition to creating a more useful product for CIA analysts, the investment allowed Intelliseek to enter international markets two years faster than it had planned.

Most tantalizing for counterterrorism investigators are the possibilities of predictive analysis. Technology from digiMine helps the J. Crew website identify the piece of clothing a shopper is most likely to buy based on his or her previous purchases. A loan company using predictive-analysis software from Sightward, based in Bellevue, Wash., discovered that the No. 1 indicator of whether Web applicants will go through with a loan rather than merely check current quotes was whether they voluntarily identified their gender on the website.

ABM, based in Nottingham, England, has sold its predictive software to both businesses and law-enforcement agencies, primarily in its home market. Police in Hampshire, England, used the system to analyze patterns behind a rash of burglaries within an apartment complex and determined which building was likely to be the next target. Police stepped up patrols there and arrested a man carrying out a computer at 4 a.m. He confessed to the other crimes.

Even more useful is new software that links the databases of different agencies or divisions of large corporations. Such software languages as XML (short for extensible markup language) provide a universal translator that can reach into the oldest computer systems and make them comprehensible to customized search engines and data-mining applications. Autonomy has enabled employees of such megafirms as BP, General Electric and General Motors to retrieve information from their companies' mishmash of databases around the globe. By installing such software, GM was able to offer its employees easy access to each of its 700 intranets and thereby reduce the tech staff used to maintain these systems from 30 full-time positions to one half-timer.

Autonomy has also linked the databases of 56 British police forces through a project with Unisys known as HOLMES II — after Sherlock, of course. HOLMES II allows officers in different departments to search one another's crime databases and uses artificial-intelligence technology to recognize the meaning of words from their context and make links between similar clues that may have been entered differently by different people.

Americans got a glimpse of how such a system might work this fall during the Washington-sniper investigation. Two weeks into the shootings, Knowledge Computing, an Arizona company whose Coplink system has integrated police databases in Tucson, Ariz., and Phoenix, volunteered its software to help with the investigation. The system was set up in Montgomery County, Md., only a day before the arrests were made, so it did not play a role in solving the shootings. Working through the hundreds of thousands of leads that were entered into various police computer systems, however, Coplink noted that witnesses reported seeing John Muhammad's blue Chevrolet Caprice near two of the Washington-area shootings, and local police ran computer checks on his license plate at least three times during the killing spree.

Such tip-sharing systems are expected to spread quickly among police agencies. But implementing them at the federal level will be a nightmare. The size, complexity and backwardness of some government systems mock any IT company's claim of scalability. Many FBI agents still use computers without point-and-click capabilities. And the IRS still stores many of its records on reels of computer tape.

A key challenge will be better sharing of information without compromising its security and without violating laws that protect the privacy and civil rights of individuals. A CIA analyst can legally view only a small percentage of information within FBI computers. The complex system of security clearances locks down entire classes of data, often just for bureaucratic convenience. There are first steps toward solving the problem. Convera, based in Vienna, Va., is adding a new feature to its retrieval software that will automatically identify the classification level of a certain document and then distribute it to whoever is most likely to need it. Verity, based in Sunnyvale, Calif., already deploys for businesses software that hides classified documents from unauthorized employees.

Ultimately, though, the biggest hurdle will be the size of the government and the sheer number of its discrete systems and databases. "This is a 10-year project," says digiMine's Fayyad. "Anyone who thinks that you can simply apply a piece of software on top of the data and then you're in business doesn't understand it." Fayyad has decided to take a pass on the government market because of its complexities and what he describes as the Bush Administration's failure to present a clear picture of the system it wants to create. But eager to take his place are scores of software companies looking for a market, any market, that shows signs of life.