

Reprinted from

# The New York Times

## Agencies Look for More Ways to Mine Data

By JOHN MARKOFF; Scott Shane contributed reporting from Washington for this article.  
Published February 25, 2006

PALO ALTO, Calif., Feb. 23 – A small group of National Security Agency officials slipped into Silicon Valley on one of the agency's periodic technology shopping expeditions this month.

On the wish list, according to several venture capitalists who met with the officials, were an array of technologies that underlie the fierce debate over the Bush administration's anti-terrorist eavesdropping program: computerized systems that reveal connections between seemingly innocuous and unrelated pieces of information.

The tools they were looking for are new, but their application would fall under the well-established practice of data mining: using mathematical and statistical techniques to scan for hidden relationships in streams of digital data or large databases.

Supercomputer companies looking for commercial markets have used the practice for decades. Now intelligence agencies, hardly newcomers to data mining, are using new technologies to take the practice to another level.

But by fundamentally changing the nature of surveillance, high-tech data mining raises privacy concerns that are only beginning to be debated widely. That is because to find illicit activities it is necessary to turn loose software sentinels to examine all digital behavior whether it is innocent or not.

"The theory is that the automated tool that is conducting the search is not violating the law," said Mark D. Rasch, the former head of computer-crime investigations for

the Justice Department and now the senior vice president of Solutionary, a computer security company. But "anytime a tool or a human is looking at the content of your communication, it invades your privacy."

When asked for comment about the meetings in Silicon Valley, Jane Hudgins, a National Security Agency spokeswoman, said, "We have no information to provide."

Data mining is already being used in a diverse array of commercial applications -- whether by credit card companies detecting and stopping fraud as it happens, or by insurance companies that predict health risks. As a result, millions of Americans have become enmeshed in a vast and growing data web that is constantly being examined by a legion of Internet-era software snoops.

Technology industry executives and government officials said that the intelligence agency systems take such techniques further, applying software analysis tools now routinely used by law enforcement agencies to identify criminal activities and political terrorist organizations that would otherwise be missed by human eavesdroppers.

One such tool is **Analyst's Notebook**, a crime investigation "spreadsheet" and visualization tool developed by **i2 Inc.**, a software firm based in McLean, Va.

The software, which ranges in price from as little as \$3,000 for a sheriff's department to millions of dollars for a large government agency like the Federal Bureau of Investigation, allows investigators to organize and view telephone and financial transaction records. It was used in 2001 by Joyce Knowlton, an investigator at the Stillwater State Correctional Facility in Minnesota, to detect a prison drug-smuggling ring that ultimately implicated 30 offenders who were linked to Supreme White Power, a gang active in the prison.

Ms. Knowlton began her investigation by importing telephone call records into her software and was immediately led to a pattern of calls between prisoners and a recent parolee. She overlaid the calling data with records of prisoners' financial accounts, and based on patterns that emerged, she began monitoring phone calls of particular inmates. That led her to coded messages being exchanged in the calls that revealed that seemingly innocuous wood blocks were being used to smuggle drugs into the prison.

"Once we added the money and saw how it was flowing from addresses that were connected to phone numbers, it created a very clear picture of the smuggling ring," she said.

Privacy, of course, is hardly an expectation for prisoners. And credit card customers and insurance policyholders give up a certain amount of privacy to the issuers and carriers. It is the power of such software tools applied to broad, covert governmental uses that has led to the deepening controversy over data mining.

In the wake of 9/11, the potential for mining immense databases of digital information gave rise to a program called Total Information Awareness, developed by Adm. John M. Poindexter, the former national security adviser, while he was a program manager at the Defense Advanced Research Projects Agency.

Although Congress abruptly canceled the program in October 2003, the legislation provided a specific exemption for "processing, analysis and collaboration tools for counterterrorism foreign intelligence."

At the time, Admiral Poindexter, who declined to be interviewed for this article because he said he had knowledge of current classified intelligence activities, argued that his program had achieved a tenfold increase in the speed of the searching databases for foreign threats.

While agreeing that data mining has a tremendous power for fighting a new kind of warfare, John Arquilla, a professor of defense analysis at the Naval Postgraduate School in Monterey, Calif., said that intelligence agencies had missed an opportunity by misapplying the technologies.

"In many respects, we're fighting the last intelligence war," Mr. Arquilla said. "We have not pursued data mining in the way we should."

Mr. Arquilla, who was a consultant on Admiral Poindexter's Total Information Awareness project, said that the \$40 billion spent each year by intelligence agencies had failed to exploit the power of data mining in correlating information readily available from public sources, like monitoring Internet chat rooms used by Al Qaeda. Instead, he said, the government has been investing huge sums in surveillance of phone calls of American citizens.

"Checking every phone call ever made is an example of old think," he said.

He was alluding to databases maintained at an AT&T data center in Kansas, which now contain electronic records of 1.92 trillion telephone calls, going back decades. The Electronic Frontier Foundation, a digital-rights advocacy group, has asserted in a lawsuit that the AT&T Daytona system, a giant storehouse of calling records and Internet message routing information, was the foundation of the N.S.A.'s effort to mine telephone records without a warrant.

An AT&T spokeswoman said the company would not comment on the claim, or generally on matters of national security or customer privacy.

But the mining of the databases in other law enforcement investigations is well established, with documented results. One application of the database technology, called Security Call Analysis and Monitoring Platform, or Scamp, offers access to about nine weeks of calling information. It currently handles about 70,000 queries a month from fraud and law enforcement investigators, according to AT&T documents.

A former AT&T official who had detailed knowledge of the call-record database said the Daytona system takes great care to make certain that anyone using the database -- whether AT&T employee or law enforcement official with a subpoena -- sees only information he or she is authorized to see, and that an audit trail keeps track of all users. Such information is frequently used to build models of suspects' social networks.

The official, speaking on condition of anonymity because he was discussing sensitive corporate matters, said every telephone call generated a record: number called, time of call, duration of call, billing category and other details. While the database does not contain such billing data as names, addresses and credit card numbers, those records are in a linked database that can be tapped by authorized users.

New calls are entered into the database immediately after they end, the official said, adding, "I would characterize it as near real time."

According to a current AT&T employee, whose identity is being withheld to avoid jeopardizing his job, the mining of the AT&T databases had a notable success in

helping investigators find the perpetrators of what was known as the Moldovan porn scam.

In 1997 a shadowy group in Moldova, a former Soviet republic, was tricking Internet users by enticing them to a pornography Web site that would download a piece of software that disconnected the computer user from his local telephone line and redialed a costly 900 number in Moldova.

While another long-distance carrier simply cut off the entire nation of Moldova from its network, AT&T and the Moldovan authorities were able to mine the database to track the culprits.

Much of the recent work on data mining has been aimed at even more sophisticated applications. The National Security Agency has invested billions in computerized tools for monitoring phone calls around the world -- not only logging them, but also determining content -- and more recently in trying to design digital vacuum cleaners to sweep up information from the Internet.

Last September, the N.S.A. was granted a patent for a technique that could be used to determine the physical location of an Internet address -- another potential category of data to be mined. The technique, which exploits the tiny time delays in the transmission of Internet data, suggests the agency's interest in sophisticated surveillance tasks like trying to determine where a message sent from an Internet address in a cybercafe might have originated.

An earlier N.S.A. patent, in 1999, focused on a software solution for generating a list of topics from computer-generated text. Such a capacity hints at the ability to extract the content of telephone conversations automatically. That might permit the agency to mine millions of phone conversations and then select a handful for human inspection.

As the N.S.A. visit to the Silicon Valley venture capitalists this month indicates, the actual development of such technologies often comes from private companies.

In 2003, Virage, a Silicon Valley company, began supplying a voice transcription product that recognized and logged the text of television programming for government and commercial customers. Under perfect conditions, the system could

be 95 percent accurate in capturing spoken text. Such technology has potential applications in monitoring phone conversations as well.

And several Silicon Valley executives say one side effect of the 2003 decision to cancel the Total Information Awareness project was that it killed funds for a research project at the Palo Alto Research Center, a subsidiary of Xerox, exploring technologies that could protect privacy while permitting data mining.

The aim was to allow an intelligence analyst to conduct extensive data mining without getting access to identifying information about individuals. If the results suggested that, for instance, someone might be a terrorist, the intelligence agency could seek a court warrant authorizing it to penetrate the privacy technology and identify the person involved.

With Xerox funds, the Palo Alto researchers are continuing to explore the technology.