

## Data Mining and the Hunt for Terrorists

By Stephen E. Fienberg: Maurice Falk University Professor of Statistics and social Science in the Department of Statistics, the Center for Automated Learning and discovery, and Cylab, and former Dean of the College of Humanities and Social Sciences.

The events of September 11, 2001 changed our lives in irrevocable ways. You would have to have been in hibernation for the past four years to miss the fact that the federal government is engaged in a hunt for terrorists. The government has always used data collected for a variety of purposes for security purposes. Indeed, such activities lie at the heart of organizations such as the National Security Agency. But after 9/11 the effort has intensified and shifted in some surprising, ways bringing to the fore tools that we at CMU know something about.

Where and how does the government look for terrorists? Clearly there are many sources of information that government officials tap into in this regard. Some believe that if we sift through the data available from government and other sources using the latest data mining techniques we'll find the clues to prevent the next attack. For those who haven't caught up on the latest lingo from computer science, data mining is usually described as an information extraction activity, using machine learning and statistical tools for discovering hidden facts contained in databases. Using "An Algorithm As a Pickaxe" was the data mining metaphor in a recent *New York Times* article (October 10, 2005). Data mining is what most of us in CMU's Center for Automated Learning and Discovery (CALD) do, in one form or another, and we now offer a Ph.D. in Data Mining and Knowledge Discovery. So it must be a good thing, right? But then why are civil libertarian and public watch-dog groups like the ACLU so upset over the use of data mining? One of the problems is that many of the databases that are targets for data mining contain personal information that we have long believed was not subject to such data snooping. Thus there is a public issue as to whether we should trade our privacy for the protection against terrorists that data mining might afford us. This is another topic that those of us associated with CALD have examined, i.e., the risk-utility tradeoff associated with access to confidential data.

There's one further problem: We don't know what form the next attack will take, but we can look for clues about potential terrorists we have seen in the past. This is why Congress, in its wisdom as part of the USA Patriot Act, rescinded the confidentiality provision for data from the National Center of Educational Statistics (NCES) to allow government access to identifiable individual information. This still might sound strange to you until you realize that NCES data files include information on individuals enrolling in flight schools and we know that several of the 9/11 hijackers spent time at a flight school in Florida prior to their coordinated attack! Moreover, the 20 attackers showed up in a variety of other "unprotected" non-confidential databases once they had been identified. And the price for this kind of "retrodiction" is the potential loss of privacy, especially when we look at the sources of the data the advocates of data mining want to exploit. As the Patriot Act came up for renewal this past summer, some in Congress

attempted to force the government to disclose its use of data mining techniques in tracking suspects in terrorism cases, an effort that the Bush administration strongly resisted. Advocates for controls believe the public has the right to know what databases are being searched and with what data mining algorithms.

As pernicious as the USA Patriot Act provision rescinding the confidentiality provision of NCEC was, perhaps the biggest threat to privacy in the hunt for terrorists comes from data warehouse operations that draw on all databases, public and private, confidential and not. Companies such as Acxiom, ChoicePoint and LexisNexis use their data to perform background check on prospective applicants to employers, insurers, and credit providers. The recent security lapses at these are other large organizations such at Bank of America have been regular news items this past year and they point to serious security vulnerabilities. Someone recently observed that the security lapses at these companies confirm the maxim that “a company can have information security without privacy but not privacy without information security.”

The data warehouses also participate in the hunt for terrorists. They merge and match data on individuals, impervious to error and inaccuracies, and then repackage the data for sale to other private enterprises and government programs, at both the federal and state level, such as the *Multi-state Anti-Terrorism Information eXchange* system (MATRIX), in which Pennsylvania has been participating. MATRIX was terminated in April of this year as a result wide scale criticism, but if the past is any guide, the databases will continue in some form and re-emerge in a related form as part of yet another such program. Such data warehousing efforts are ticking privacy-invading time bombs, both because of the widely heralded unauthorized releases of information that we read about almost daily in our newspapers, and because of the ways there error-filled individual records are likely to be used in ways no one originally envisioned.

Should you worry about these data warehouses? With very high probability they contain data on you and your household, but you will never quite know what data or how accurate the information is. And soon the data may be matched into a government-sponsored terrorist search systems such as the one being set up by the Transportation Security Administration (TSA) to match passenger lists into a consolidated watch list of suspected terrorists. On September 19, 2005, the “Secure Flight” Working Group to the Transportation Security Administration (TSA) submitted a report questioning TSA’s secrecy regarding what data it plans to use and how:

“The TSA is under a Congressional mandate to match domestic airline passenger lists against the consolidated terrorist watch list. TSA has failed to specify with consistency whether watch list matching is the only goal of Secure Flight at this stage....

“Will Secure Flight be linked to other TSA applications? ...

“How will commercial data sources be used? One of the most controversial elements of Secure Flight has been the possible uses of commercial data. TSA has

never clearly defined two threshold issues: what it means by “commercial data,” and how it might use commercial data sources in the implementation of Secure Flight. TSA has never clearly distinguished among various possible uses of commercial data, which all have different implications.

"Possible uses of commercial data sometimes described by TSA include: (1) identity verification or authentication; (2) reducing false positives by augmenting passenger records indicating a possible match with data that could help distinguish an innocent passenger from someone on a watch list; (3) reducing false negatives by augmenting all passenger records with data that could suggest a match that would otherwise have been missed; (4) identifying sleepers, which itself includes: (a) identifying false identities; and (b) identifying behaviors indicative of terrorist activity. A fifth possibility has not been discussed by TSA: using commercial data to augment watch list entries to improve their fidelity. Assuming that identity verification is part of Secure Flight, what are the consequences if an identity cannot be verified with a certain level of assurance?"

If TSA won't tell its advisory group charged with evaluating the privacy and security of its system what it plans to do, it's hard for the public to relax about the potential uses of commercial data whose quality the vendors will not and cannot guarantee.

So are we destined to hunt for terrorists in this willy-nilly and privacy-invasive fashion, always searching for those that resemble the few we have caught or those involved in the attacks we have experienced? Even then the hunt for terrorists using interconnected databases would be like the proverbial search for a needle in a haystack, and if we were not concerned with lots of false positives, that is people identified as terrorists who are not, then data mining tools as we know them today might well be perfectly fine tools for the job. The problem is that to do the job well we need to look for people who are going to commit an act that we have not seen before. Two quotes about predicting the future are especially apt in this context:

Niels Bohr: “Prediction is very difficult, especially about the future.”

Yogi Berra: “The future ain't what it used to be.”

It's one thing to develop data mining tools that detect events that have already happened and it is another to use such tools to predict aspects of events that have not yet occurred and who will participate in them. We have no way to validate the latter.

The hype about technological tools such as data mining often makes it difficult to assess the claims. For example, *Nature* ran a news article on September 22 with the headline: “Brain imaging ready to detect terrorists, say neuroscientists.” Imagine, functional magnetic resonance imaging (fMRI) to the rescue! What *Nature* did not report was that this conclusion was based on the neuroscientists' study of only 26 subjects in the laboratory using a “guilty knowledge” type test which all experts on the detection of deception acknowledge could hardly use to identify terrorists of whose plans

we are unaware! And a careful examination of the paper summarizing the study suggests that the accuracy of fMRI for this purpose was hardly distinguishable from the polygraph.

The analogy with polygraphs is an apt one. We know that the use of polygraphs to detect spies leads to two kinds of errors: false positives and false negatives. False positives are the thing we usually focus on, honest individuals labeled incorrectly as deceptive or upstanding citizens labeled as terrorists. But false negatives are at least as bad: real spies or real terrorists who go undetected. Believe in the accuracy of the prediction can breed complacency, under the expectation that all of the terrorist have been discovered. When it comes to polygraphs there is evidence suggesting that both types of errors are too high and in particular that using the polygraph as a screening device to detect spies simply doesn't work! Yet those who administer polygraphs in our national laboratories still believe in their accuracy and effectiveness. Should we view data mining for terrorists in a similar fashion?

The arguments over data mining extend to surveillance in public spaces, whether it is done by TSA at airports or by cameras located throughout our metropolitan areas. How is information gathered and utilized? Where do watch-lists used at airports come from and how are they being used? Does matching names on airline tickets against IDs protect anyone? Can we accurately and effectively match facial images from passengers in airports or on subways against ones of known terrorists? What are the costs associated with surveillance techniques, including privacy loss, and what are the real gains? The list of questions goes on and on. At least some of these questions are amenable to empirical examination.

The claims for data mining usually rest on its value in automatically sifting through large amounts of data and detecting undiscovered patterns and relationships. We often hear that the more data the better. But data mining algorithms in the end are built on a foundation involving statistical models and if the models are aren't correct or the data aren't appropriate then more is not better. Are the databases the government wants to use in the hunt for terrorists appropriate? What do we know about the data fusion techniques that will be used to combine data from different sources? Can we really use data mining tools effectively to protect us from terrorist attacks in the future without seriously compromising our privacy? A new committee at the National Research Council is charged with investigating answers to these questions. We can expect a report in a couple of years, so stay tuned. But in the meantime be wary, both about what you read and hear, and about what some are trying to do with your data and mine.



Source: [http://www.claybennett.com/images/archivetoons/peeping\\_john.gif](http://www.claybennett.com/images/archivetoons/peeping_john.gif)  
claybennett@earthlink.net