CNET News
News - Politics and Law
October 7, 2008 9:30 AM PDT

Government report: Data mining doesn't work well

Posted by Declan McCullagh

The most extensive government report to date on whether terrorists can be identified through data mining has yielded an important conclusion: It doesn't really work.

A National Research Council report, years in the making and scheduled to be released Tuesday, concludes that automated identification of terrorists through data mining or any other mechanism "is neither feasible as an objective nor desirable as a goal of technology development efforts." Inevitable false positives will result in "ordinary, law-abiding citizens and businesses" being incorrectly flagged as suspects.

The whopping 352-page report, called "Protecting Individual Privacy in the Struggle Against Terrorists," amounts to at least a partial repudiation of the Defense Department's controversial data-mining program called Total Information Awareness, which was **limited by Congress** in 2003.

But the ambition of the report's authors is far broader than just revisiting the problems of the TIA program and its successors. Instead, they aim to produce a scholarly evaluation of the current technologies that exist for data mining, their effectiveness, and how government agencies should use them to limit false positives--of the sort that can result in situations like heavily-armed SWAT teams raiding someone's home and **shooting their dogs** based on the **false belief** that they were part of a drug ring.

The report was written by a committee whose members include William Perry, a professor at Stanford University; Charles Vest, the former president of MIT; W. Earl Boebert, a retired senior scientist at Sandia National Laboratories; Cynthia Dwork of Microsoft Research; R. Gil Kerlikowske, Seattle's police chief; and Daryl Pregibon, a research scientist at Google.

They admit that far more Americans live their lives online, using everything from VoIP phones to Facebook to RFID tags in automobiles, than a decade ago, and the databases created by those activities are tempting targets for federal agencies. And they draw a distinction between subject-based data mining (starting with one individual and looking for

connections) compared with pattern-based data mining (looking for anomalous activities that could show illegal activities).

But the authors conclude the type of data mining that government bureaucrats would like to do--perhaps inspired by watching too many episodes of the Fox series *24*--can't work. "If it were possible to automatically find the digital tracks of terrorists and automatically monitor only the communications of terrorists, public policy choices in this domain would be much simpler. But it is not possible to do so."

A summary of the recommendations:

> * U.S. government agencies should be required to follow a systematic process to evaluate the effectiveness, lawfulness, and consistency with U.S. values of every information-based program, whether classified or unclassified, for detecting and countering terrorists before it can be deployed, and periodically thereafter.
>
> * Periodically after a program has been operationally deployed, and in particular before a program enters a new phase in its life cycle, policy makers should (carefully review) the program before allowing it to continue operations or to proceed to the next phase.
>
> * To protect the privacy of innocent people, the research and development of any information-based counterterrorism program should be conducted with synthetic population data... At all stages of a phased deployment, data about individuals should be rigorously subjected to the full safeguards of the framework.
>
> * Any information-based counterterrorism program of the U.S. government should be subjected to robust, independent oversight of the operations of that program, a part of which would entail a practice of using the same data mining technologies to "mine the miners and track the trackers."
>
> * Counterterrorism programs should provide meaningful redress to any individuals inappropriately harmed by their operation.
>
> * The U.S. government should periodically review the nation's laws, policies, and procedures that protect individuals' private information for relevance and effectiveness in light of changing technologies and

> circumstances. In particular, Congress should re-examine existing law to consider how privacy should be protected in the context of information-based programs (e.g., data mining) for counterterrorism.

By itself, of course, this is merely a report with non-binding recommendations that Congress and the executive branch could ignore. But NRC reports are not radical treatises written by an advocacy group; they tend to represent a working consensus of technologists and lawyers.

The great encryption debate of the 1990s was one example. The NRC's so-called **CRISIS report** on encryption in 1996 concluded export controls--that treated software like Web browsers and PGP as munitions--were a failure and should be relaxed. That **eventually happened** two years later.