Data Mining At the Crossroads: Successes, Failures, and Learning from Them

Haym Hirsh

Department of Computer Science
Rutgers University

Division of Information and Intelligent Systems
U.S. National Science Foundation

What is Data Mining?

For the purposes of my presentation:

Data Mining =

The extraction of useful information from data

(I.e., Data Mining broadly construed)

Some of My Favorite Data Mining Successes

Some of My Favorite Data Mining Successes

- Web search

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering
- Recommender systems

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering
- Recommender systems
- Machine translation

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering
- Recommender systems
- Machine translation
- Massive data clusters

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering
- Recommender systems
- Machine translation
- Massive data clusters

- Conferences like this one: Participation by people in diverse, previously disjoint subfields (databases, machine learning, statistics, etc.)

Some of My Favorite Data Mining Successes

- Web search
- Spam filtering
- Recommender systems
- Machine translation
- Massive data clusters

- Conferences like this one: Participation by people in diverse, previously disjoint subfields (databases, machine learning, statistics, etc.)
- Benchmark datasets

Some Data Mining "Failures"

Some Data Mining "Failures"

- Socio-Political

The New York Times
Tuesday, August 14, 2007

Technology

Florida List for Purge of Voters Proves Flawed

By FORD FESSENDEN
Published: July 10, 2004

Florida election officials used a flawed method to come up with a listing of people believed to be convicted felons, a list that they are recommending be used to purge voter registration rolls, state officials acknowledged yesterday. As a result, voters identifying themselves as Hispanic are almost completely absent from that list.

Of nearly 48,000 Florida residents on the felon list, only 61 are Hispanic. By contrast, more than 22,000 are African-American.

About 8 percent of Florida voters describe themselves as Hispanic, and about 11 percent as black.

In a presidential-election battleground state that decided the 2000 race by giving George W. Bush a margin of only 537 votes, the effect could be significant: black voters are overwhelmingly Democratic, while Hispanics in Florida tend to vote Republican.

Elections officials of Florida's Republican administration denied any partisan motive in use of the method they adopted, and noted that it had been approved as part of a settlement of a civil rights lawsuit.

Sen. Kennedy flagged by no-fly list

'T. Kennedy' used as alias by terrorist suspect

By Sara Kehaulani Goo

washingtonpost.com

Updated: 12:10 a.m. ET Aug 20, 2004

WASHINGTON - U.S. Sen. Edward M. "Ted" Kennedy said yesterday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on the government's secret "no-fly" list.

Federal air security officials said the initial error that led to scrutiny of the Massachusetts Democrat should not have happened even though they recognize that the no-fly list is imperfect. But privately they acknowledged being embarrassed that it took the senator and his staff more than three weeks to get his name removed.

A senior administration official, who spoke on condition he not be identified, said Kennedy was stopped because the name "T. Kennedy" has been used as an alias by someone on the list of terrorist suspects.

Story continues below ↓

Spencer Platt / Getty Images file

U.S. Sen. Ted Kennedy of Massachusetts

The Washington Post

Lawsuits May Illuminate Methods of Spy Program
Nasdaq Gives High Rollers A Market Free Of Regulation
Source Disclosure Ordered in Anthrax Suit
U.S. Will Fund Plan to Charge Tolls In New York
Shuttle Tiles Perforated, But Repair Poses Dangers

POLITICS : SECURITY RSS

Pentagon Spy Office to Close

Associated Press 09.25.03 | 9:41 AM

WASHINGTON -- House and Senate negotiators have decided to close a Pentagon office that was developing a vast computerized terrorism surveillance system and bar spending that would allow those high-tech spying tools to be used against Americans on U.S. soil.

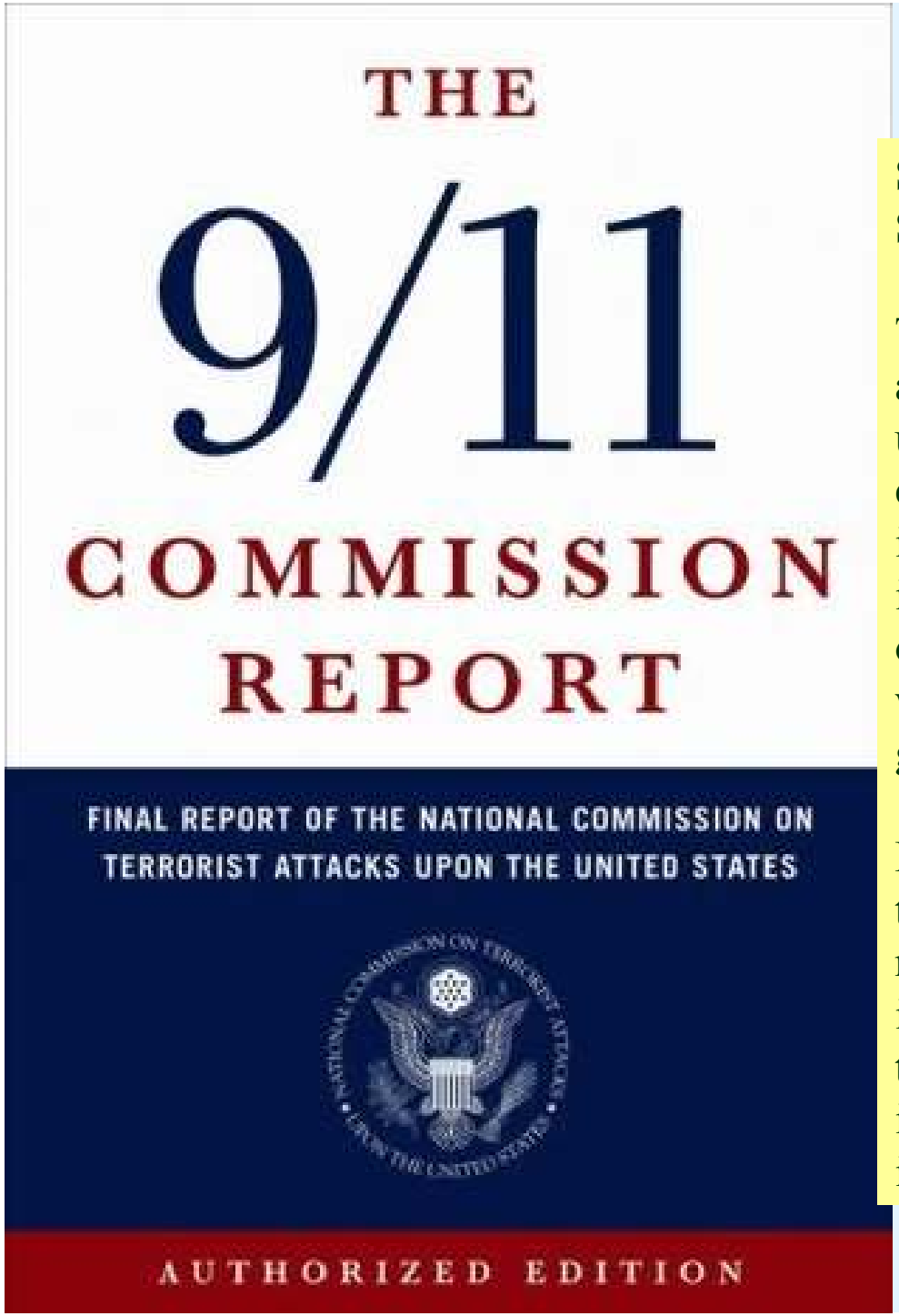
But they left open the possibility that some or all of the high-powered software tools under development might be employed by different government offices to gather foreign intelligence from foreigners, U.S. citizens aboard or foreigners in this country.

The controversial Terrorism Information Awareness program was conceived by retired Adm. John Poindexter and was run by the Information Awareness Office that he headed inside the Defense Advanced Research Projects Agency. It was developing software that could examine the computerized travel, credit, medical and other records of Americans and others around the world to search for telltale activities that might reveal preparations for a terrorist attack.

Sen. Ron Wyden (D-Oregon), who has led a campaign against the program, hailed the result Wednesday. "Americans on American soil are not going to be targets of TIA surveillance that would have violated their privacy and civil liberties. The government is not going to be able to pick Americans up by their ankles and shake them to see if anything funny falls out," Wyden said in an interview.

Most Popular Most

1. June 13, 1944: V-1 R
New Kind of Warfare
2. PlayStation 2 Games



Section 13.3: UNITY OF EFFORT IN SHARING INFORMATION

The U.S. government has access to a vast amount of information. When databases not usually thought of as "intelligence," such as customs or immigration information, are included, the storehouse is immense. … In interviews around the government, official after official urged us to call attention to frustrations with the unglamorous "back office" side of government operations. …

Recommendation: The president should lead the government-wide effort to bring the major national security institutions into the information revolution. He should coordinate the resolution of the legal, policy, and technical issues across agencies to create a "trusted information network."

Some Data Mining "Failures"

- Socio-Political

Some Data Mining "Failures"

- Socio-Political
  - Bad data mining

Some Data Mining "Failures"

- Socio-Political
  - Bad data mining
  - Misused data mining

Some Data Mining "Failures"

- Socio-Political
  - Bad data mining
  - Misused data mining
  - Ignorant decision-making

Some Data Mining "Failures"

- Socio-Political
  - Bad data mining
  - Misused data mining
  - Ignorant decision-making
  - Ramifications of data mining

Some Data Mining "Failures"

- Socio-Political
  - Bad data mining
  - Misused data mining
  - Ignorant decision-making
  - Ramifications of data mining
  - Presuming fixed technology

Some Data Mining "Failures"

- Data Mining is about Real Data:
  Benchmark data sets are a means to an end
  - Data sets are supposed to be representative of the sorts of problems our algorithms will see in practice
  - Data sets must stay timely as technological and scientific advances allow our ambitions to grow
  - A data set from some domain is not an application -- Who do you personally know that cares about your results?

Some Data Mining "Failures"

- How do we ensure reproducible results?
  - Many of the applications of data mining are in the commercial sector -- How do we handle research results that reflect proprietary or otherwise restricted data?
  - How do we make sure academic research results address problems that are important in practice?
  - How do we handle inherent resource differentials between industry and academic research?
    - Access to data
    - Massive data centers
  - What new models of publication are particularly suited to data mining – "Executable articles" (Mark Liberman)