

The Use of Commercial Databases for National Security: Privacy, Evaluation, and Accuracy

Rebecca Wright

*Computer Science Department
Stevens Institute of Technology
www.cs.stevens.edu/~rwright*

National Academy of Sciences
Science, Technology, and Law Panel
18 March, 2005



Surveillance and National Security

- A major current focus is to analyze large amounts of data from diverse sources in order to:
 - detect and thwart possible terrorist incidents before they occur
 - recognize that an incident is underway
 - identify and prosecute terrorists after incidents occur
- Commercial databases are a significant source of data that may prove useful in these tasks.

Concerns with Data Usage

- Accuracy of the results
- Privacy of sensitive information:
 - Insider abuse
 - Security breach to outsiders
 - Mission creep
- Evaluating the tradeoffs of a given system to determine whether or not it should be used.

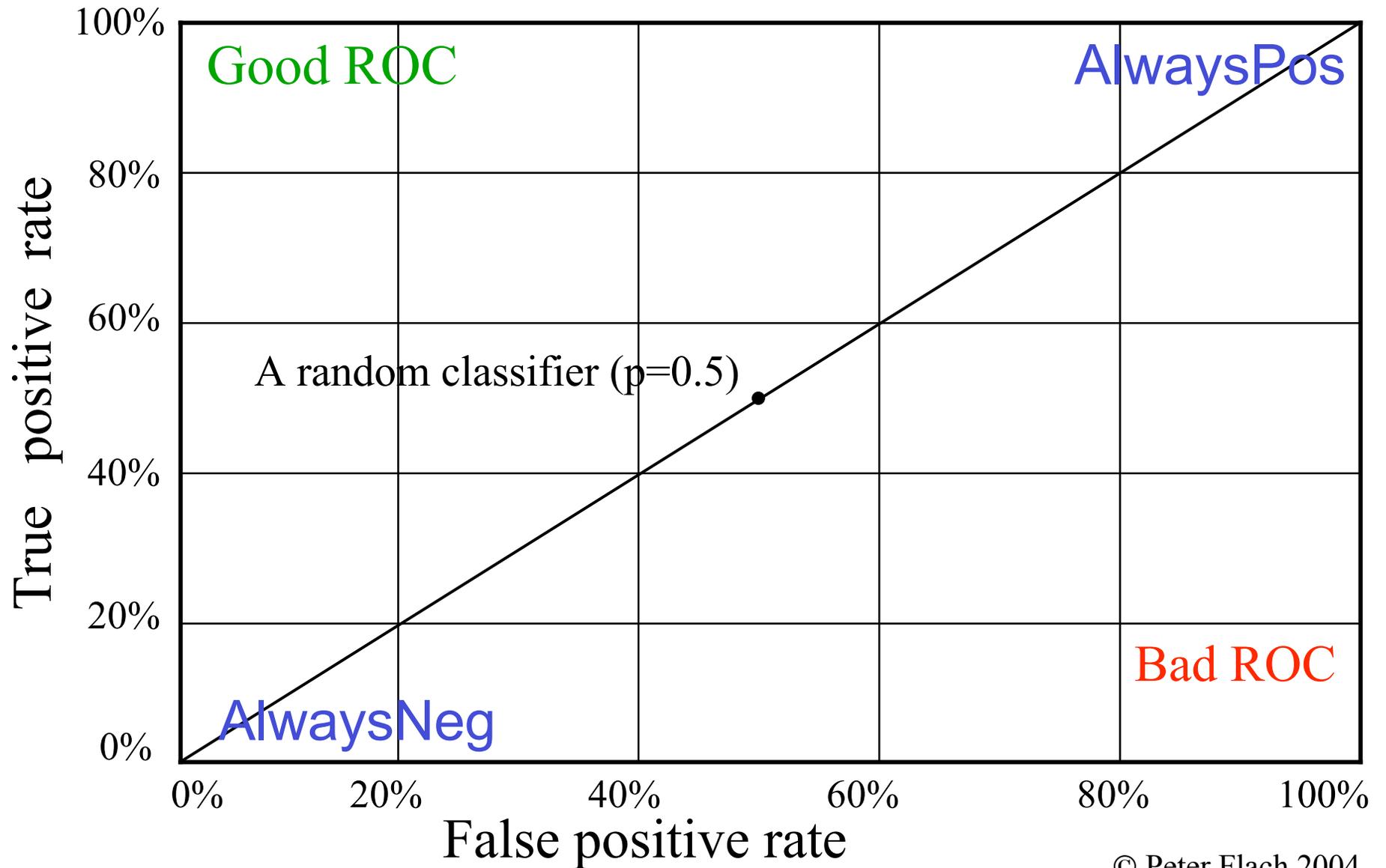
Accuracy: Positives and Negatives

	Predicted Positive	Predicted Negative
Positive Example	True Positive	False Negative
Negative Example	False Positive	True Negative

Measuring Accuracy

- False positives and negatives cannot be considered in isolation.
- The machine learning research community has various established metrics for measuring accuracy:
 - error rate: a single number, but misses important information
 - receiver operating characteristic (ROC): a curve rich in information, but harder to compare two curves
 - area under ROC: a single number determined from the ROC

ROC Curves



Dealing with Inaccuracy

- Even the best algorithms will sometimes make some errors (e.g., due to unexpected or incorrect data).
- Must also create procedures and culture recognizing that:
The computer is NOT always right.

Evaluation

- How well does a given system perform a desired task?
- What drawbacks does it have (cost, inaccuracy, loss of privacy)?
- Are there other solutions that could perform the task at least as well, with fewer drawbacks?
- Is there a completely different approach to the larger problem avoiding the need for the particular task?

Privacy

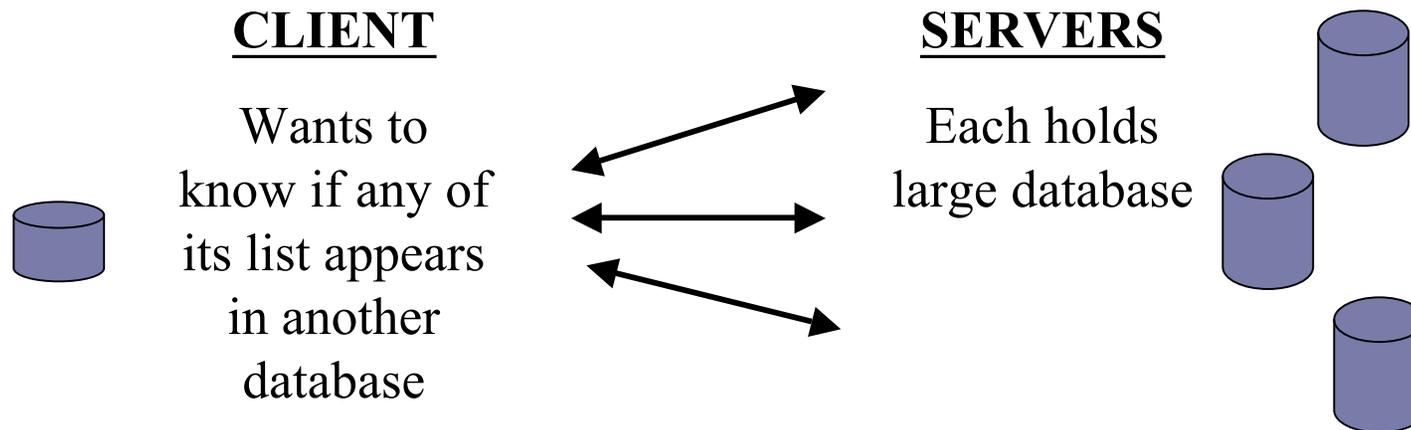
- protection of personal information: protects innocent individuals against loss of seclusion, autonomy, trust in government
- protection of proprietary or security-sensitive information
- enables collaboration between different data owners (since they may be more willing or able to collaborate if they need not reveal their information)
- compliance with legislative policies

Privacy-Preserving Computation on Large Databases

Allow multiple data holders to collaborate to compute important information while protecting the privacy of other information.

Technological tools include cryptography, data perturbation and sanitization, access control, inference control, and secure hardware or trusted platforms.

Private Database Matching [Jon03, FMP04]

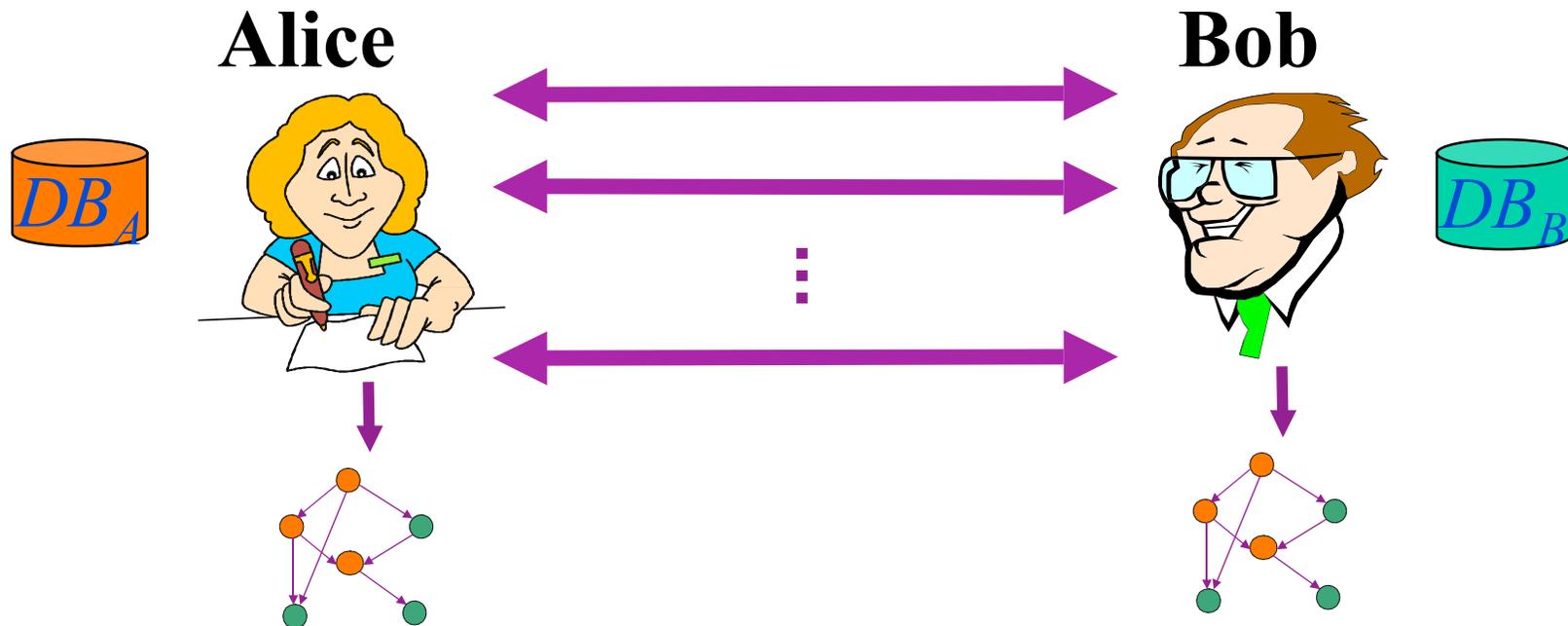


- Parties communicate so that:
 - Client learns which of its records appear in the server data, but learns nothing else about the server data
 - Servers do not learn client's data, except perhaps records that match
 - Potentially, servers do not learn which fields are queried, or any information about other servers' data
 - Computation and communication are very efficient

Privacy-Preserving Bayes Networks

[WY04, YW05]

Goal: Cooperatively learn Bayesian network structure on the combination of DB_A and DB_B , ideally without either party learning anything except the Bayesian network itself.



Technologies for Protecting Privacy

- Secure multiparty computation [Yao82, BGW88,...]
- Private information retrieval [CKGS95,...]
- Search on encrypted data [FIPR04, CM04,...]
- Private matching [Jon03, FMP04, ...]
- Private inference control [SW04, JW05]
- Privacy-preserving data mining [LP00, AS00, FIMNSW01, ESAG02, KC02, VC03, WY04, YW05, ...]
- k-anonymity [Swe02, MW03, ZYW05]
- ...

The PORTIA Project

Privacy, Obligations, and Rights in Technologies of Information Assessment

A five-year multidisciplinary project focusing on the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners, and data users.

Funded by the National Science Foundation as a Large ITR (Information Technology Research) grant, Oct 2003 - Sept 2008.

PORTIA Personnel

- Academic investigators:
 - **Dan Boneh**, Hector Garcia-Molina, John Mitchell, Rajeev Motwani, *Stanford*
 - **Joan Feigenbaum**, Ravi Kannan, Avi Silberschatz, *Yale*
 - **Stephanie Forrest**, *University of New Mexico*
 - **Helen Nissenbaum**, *NYU*
 - **Rebecca Wright**, *Stevens Institute of Technology*

PORTIA Personnel

- Research partners
 - Jack Balkin, *Yale Law School*
 - Greg Crabb, *Secret Service*
 - Cynthia Dwork, Brian LaMacchia, *Microsoft*
 - Sam Hawala, *US Census Bureau*
 - Kevin McCurley, *IBM Research*
 - Perry Miller, *Yale Center for Medical Informatics*
 - John Morris, *Center for Democracy and Technology*
 - Benny Pinkas, *HP Labs*
 - Marc Rotenberg, *Electronic Privacy Information Center*
 - Alejandro Schaffer, *DHHS/National Institutes of Health*
 - Dan Schutzer, *Citigroup*

PORTIA Goals

- Produce a next generation of technology for handling sensitive information that is qualitatively better than the current generation's.
- Enable end-to-end handling of sensitive information over the course of its lifetime.
- Formulate an effective conceptual framework for policy making and philosophical inquiry into the rights and responsibilities of data subjects, data owners, and data users.

Major Technical Themes

- privacy-preserving data mining
- identity theft and identity privacy
- database policy enforcement tools
- managing sensitive information in P2P systems
- using trusted platforms to provide trusted privacy-preserving services
- contextual integrity

Barriers to Deployment

A number of real and/or perceived barriers prevent these technologies from enjoying widespread usage.

- Efficiency concerns
- Too complicated and difficult to use
- Misalignment of incentives
- Not ready for prime time

Efficiency

- Solutions that require a significant number of public key cryptography operations suffer from slower performance, but can provide very strong privacy.
- Solutions that use more efficient cryptographic mechanisms such as hashing, make use of (partially) trusted third parties, or randomization and statistical masking techniques may have less privacy, but can be much more efficient.
- Further research may yield solutions that provide both strong privacy and high efficiency.

Ease of Use

- Configure systems with a layered approach:
 - Reasonable defaults
 - Extensive customization possible
- There has been growing interaction between usability experts and security/privacy experts.
 - E.g., DIMACS workshop and working group on Usable Privacy and Security Software, July 7-9, 2004.

Misalignment of Incentives

- Often, those who deploy and use systems are not the entities who are directly affected by privacy breaches and inaccuracies of data. No incentive to take on the costs of better privacy and accuracy.
- Legislation can help align incentives.
- Individuals may push for such legislation if they perceive sufficient risk without it.

Readiness for Deployment

- Some of the technologies are fairly mature and close to deployable.
- Others are less mature. Core ideas are there, but more research is needed to create deployable systems.
- In both cases, significant software development, systems integration, systems engineering may be required.
- With appropriate resources (research funding, technology transfer), these can be done.

Summary

- Commercial databases contain vast amounts of data too important to be ignored in the quest for national security.
- Many technologies exist that can help address correctness, evaluation, and privacy of information systems, but some barriers must be overcome.
- Technology, policy, and education must work together.