Ted Senator, SAIC

Disclaimer: Views are my own, not those of SAIC or any Government agency

NGDM '07

Panel on Future Research Challenges and Needed Resources for Data Mining in Security, Surveillance, and Privacy Protection

11 October 2007

A Better Title:

Future Research Challenges and Needed Resources for Data Mining for Security with Privacy Protection

Cancelled Data Mining Programs

Program	Agency	Date	$ Spent	Cited Reasons
Total Information Awareness (TIA)	DARPA	September 2003	Various amounts reported	**Multiple** **(a R&D program, NOT a program to mine data)**
Computer Assisted Passenger Prescreening System (CAPPS II)	TSA	August 2004	$100M	Privacy concerns
Multi-State Anti-Terrorism Information Exchange (MATRIX)	states	April 2005	$8M	Lack of privacy safeguards
Secure Flight	TSA	February 2006	$140M ($80M more needed for privacy & security)	GAO discovers 144 known security vulnerabilities
Analysis Dissemination Visualization Insight and Semantic Enhancement (ADVISE)	DHS	September 2007	$42M	Privacy concerns

Sources: http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9037319 (all but TIA)

Data Mining Definitions: Technical

- Fayyad et. al.: the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.
- Jensen: a process that uses algorithms to discover predictive patterns in datasets
- Jonas and Harper: the process of searching data for previously unknown patterns and often using these patterns to predict future outcomes
- etc.

Data Mining Definitions: Political

- "the collection and monitoring of large volumes of sensitive personal data to identify patterns or relationships"
  (Opening Statement of Senator Patrick Leahy, Senate Judiciary Committee Hearing on "Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs" January 10, 2007)

- *DATA-MINING.-The term "data-mining" means a query or search or other analysis of 1 or more electronic databases, whereas-*
  - *(A) at least 1 of the databases was obtained from or remains under the control of a non-Federal entity, or the information was acquired initially by another department or agency of the Federal Government for purposes other than intelligence or law enforcement;*
  - *(B) a department or agency of the Federal Government or a non-Federal entity acting on behalf of the Federal Government is conducting the query or search or other analysis to find a predictive pattern indicating terrorist or criminal activity; and*
  - *(C) the search does not use a specific individual's personal identifiers to acquire information concerning that individual.*

  (Senator Feingold amendment to HR 5441)

- "searches of one or more electronic databases of information concerning U.S. person by or on behalf of an agency or employee of the government" (DoD Technology and Privacy Advisory Committee, March 2004)

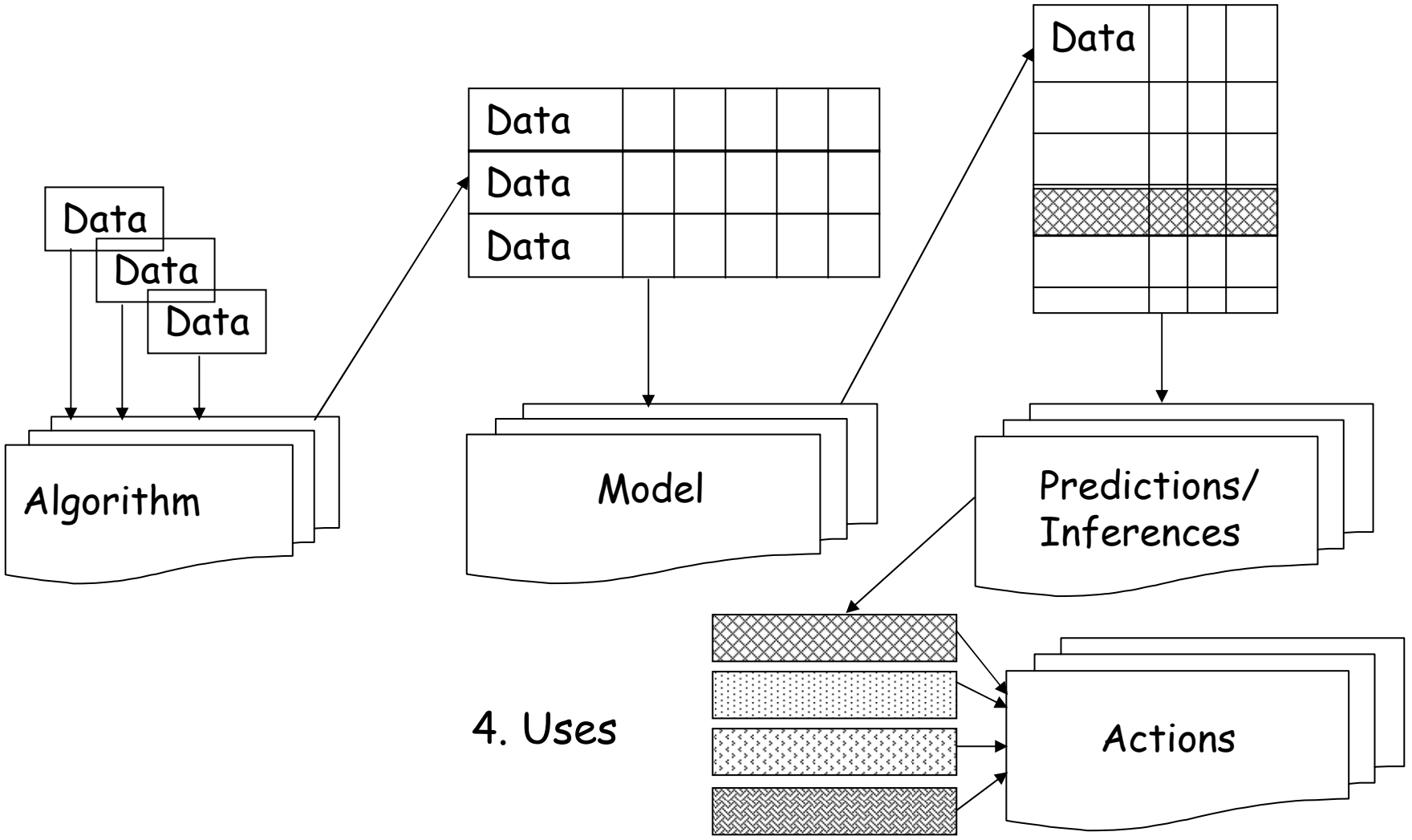
- "The untested and controversial intelligence procedure known as data-mining is capable of maintaining extensive files containing both public and private records on each and every American," Feingold said. Data-mining is a broad search of public and non-public databases in the absence of a particularized suspicion about a person, place or thing. Data mining looks for relations between things and people without any regard for particularized suspicion. January 16, 2003.

What is Data Mining, really?

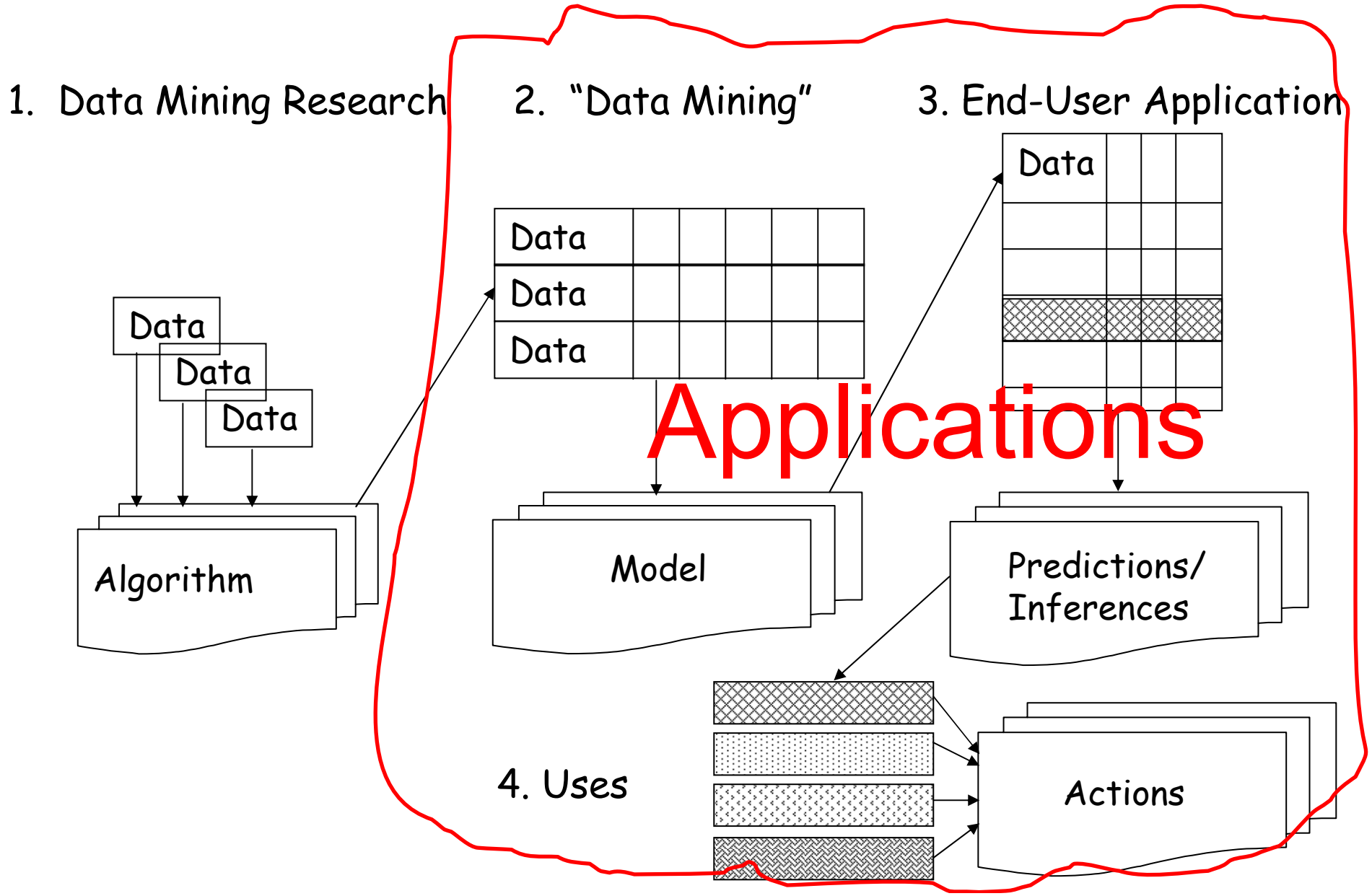
- Data Mining is not data collection
- Data Mining is not data querying
- Data Mining is not data aggregation or linking
- BUT Data Mining *Programs* may include the above
- Data Mining is a set of methods and techniques, not a particular application domain
- Data Mining is building models/finding patterns that are useful for prediction
- Interpreting the models or predictions is beyond the ability of today's data mining techniques
- Data Mining research is the development of methods and techniques for data mining
- Is Prediction part of data mining or not?

Distinct Activities (with different data needs)

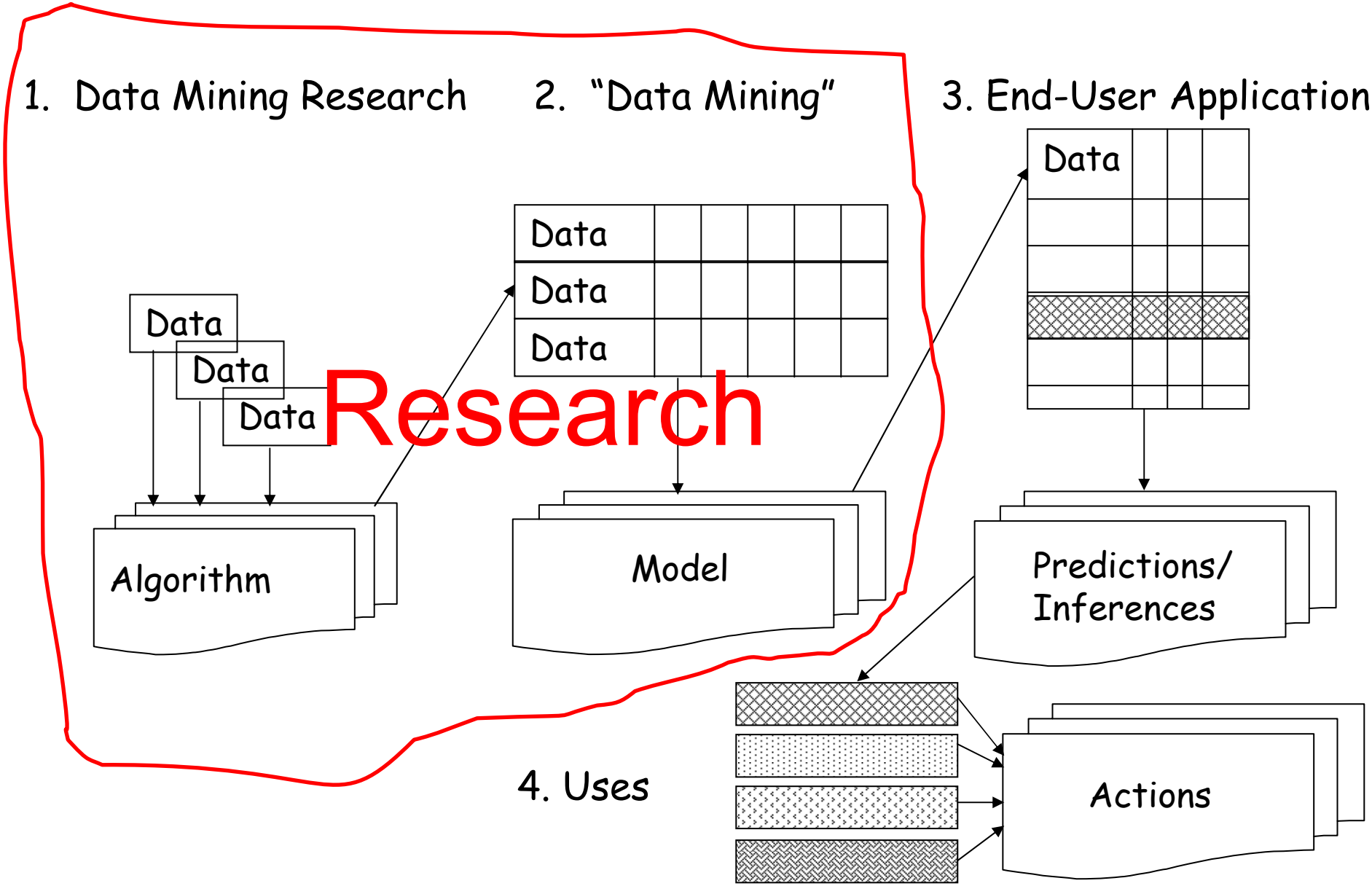
1. Data Mining Research
2. "Data Mining"
3. End-User Application

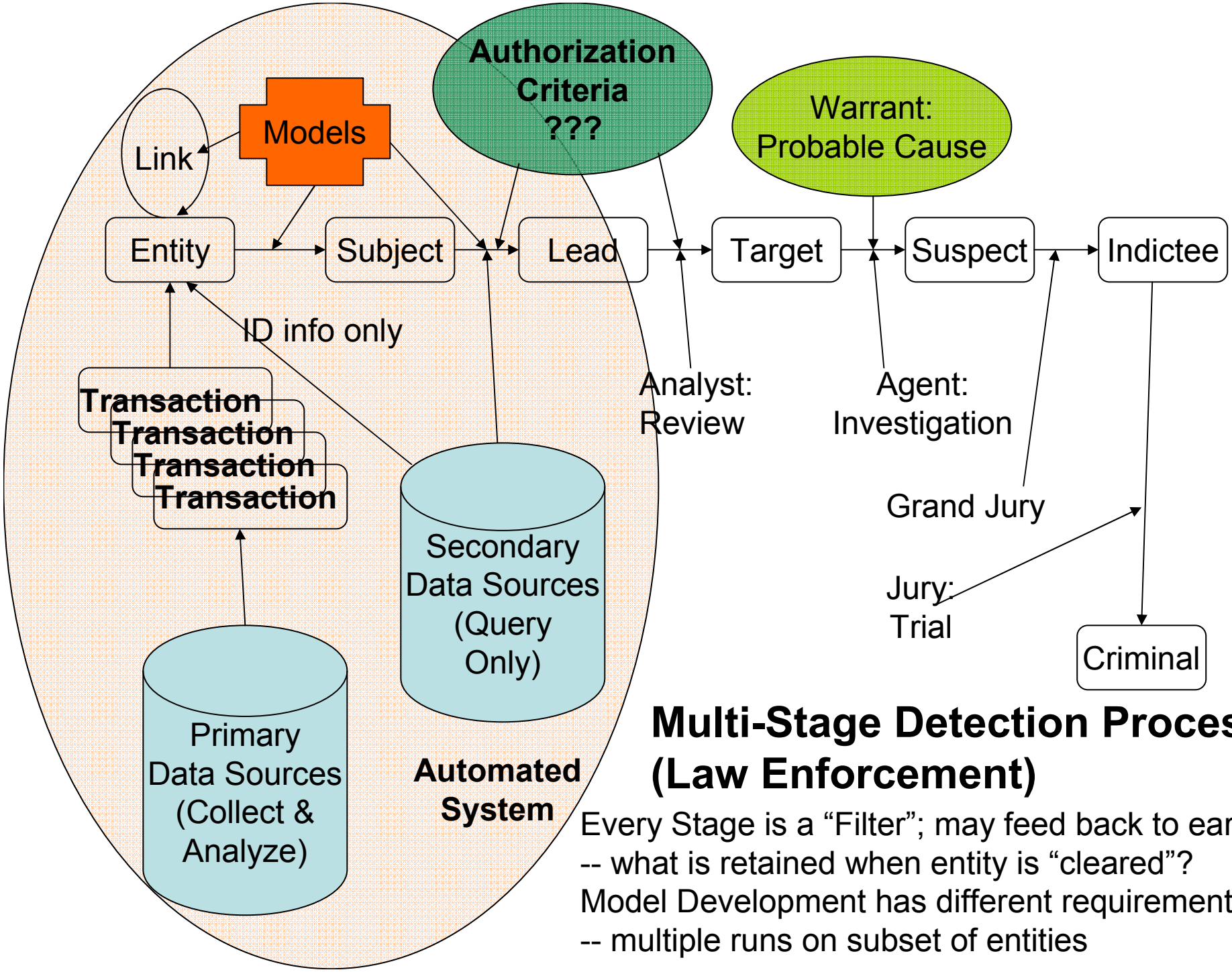


Researchers' View



Practioners' View





Multi-Stage Detection Process (Law Enforcement)

Every Stage is a “Filter”; may feed back to earlier
-- what is retained when entity is “cleared”?
Model Development has different requirements
-- multiple runs on subset of entities

What Is Identity ?
What is Privacy?

- Traditional/Intuitive Fields
  - Name, SSN, etc.
- Unique, and Interpretable, Signatures
  - Fingerprint, DNA
- Behaviors
  - Who You Call (AT&T)
  - Medical Diagnosis (e.g., Quintuplets)
- Combination of Features
  - Sex, Age, Zip Code
- Identity = Behavior + Recognition
  - Specialized versus Public Information for Recognition
- Privacy = Ability to prevent linkage of identity to information
- Intuitions
  - “Oh, so you are <name>”
  - Name versus Picture (NY Times example)
  - “But that’s private mom.”

Some Reports

- Data Mining and Homeland Security: An Overview, Jeffrey W Seifert, Congressional Research Service Report for Congress RL31798, Updated June 5, 2007
(also versions of May 21, 2003; May 3, 2004; December 16, 2004; June 7, 2005; January 27, 2006; January 18, 2007)
- Report to Congress on the Impact of Data Mining Technologies on Privacy and Civil Liberties, Maureen Cooney, Acting Chief Privacy Officer, US Department of Homeland Security, July 6, 2006
- Think Before You Dig: Privacy Implications of Data Mining and Aggregation, NASCIO Research Brief, September 2004
- Terrorism Information Awareness Program (D-2004-033), Department of Defense Inspector General, December 12, 2003
- Safeguarding Privacy in the Fight Against Terrorism: Report of the Technology and Privacy Advisory Committee, March 2004

Multiple Issues

- Who owns specific information?
- For what purposes can information be used?
  - For what purposes can individually identifiable information be used?
- When can general information be used to identify specific individuals?
- Under what conditions can individual information be revealed? To Whom? When can information from multiple sources be combined?
- How can information be corrected?
- Where in a system should “privacy” be protected?
- When does privacy need to be considered?
- Who decides? Who checks?
- When is pattern-based prediction justified?
- What actions are justified based on predictions?
- How can predictions be challenged?
- What accuracy is necessary? (Cost of false positives)

Lessons and Recommendations

- Consider Privacy Implications Before Beginning Project
- Be completely transparent regarding purpose, reason data are collected, how they will be used, who will have access, how it will be secured, where/for how long data are retained, whether individuals can access and correct their personal information, etc.
- Technology is only part of the solution; data, processes; policies, authorities, laws, etc. are at least as important and difficult

The Debate: Trends

- Advocacy groups and lawyers, but few scientists
  - Where are the Data Miners?
  - Corporations getting involved
- ACM SIGKDD Letter, "Data Mining is NOT Against Civil Liberties" June 30, 2003 (revised July 28 2003) http://www.sigkdd.org/civil-liberties.pdf
- S. 236 "Federal Agency Data Mining Reporting Act of 2007"

Role of Data Mining Researchers/Experts

- Invent Good/Useful Technology
  - Move the tradeoff curve
  - More security and more privacy
- Inform policy debate
  - Based on real science
  - What is known and what is not
  - Don’t claim expertise outside of area of competence
- Recognize societal implications of work
- No special role with respect to societal choices

Trends & Approaches

- Only 1 paper in KDD 2007
- 15 page on-line bibliography: www.csee.umbc.edu/kunliul/research/privacy_review_html

- Anonymization Techniques
- Blurring Techniques
- Hiding Techniques

- Guarantees vs Practicality: built-into algorithms or system?

- BUT, do these address
  - Scalability ?
  - Network Effects ? (Backstrom-Dwork-Kleinberg 2007)
  - Social and legal issues (e.g., use/consequences) ?

Needed Research & Resources

- Identity-Free Pattern Discovery
  - Entity Resolution without identification
  - Linking without identification
- Multi-Stage Detection in Multiple Relational Databases
- Maintaining Networked Anonymity
- Provably Auditable Data Mining/Predictions/Systems
- Privacy aware/allowing data mining algorithms
- Privacy policies, formalizations, etc.
- Privacy enforcing mechanisms, limitations
- Relationship-Preserving Anonymization
- Privacy Officers who understand Technology
- Scientists, managers, users who understand privacy