Panel:
What are the key applications and challenge problems for privacy-preserving data mining?

Chris Clifton
Vladimir Estivill-Castro
Maximilian Teltzrow
And all the participants

What Research Is Needed?

Panel Goal: Fill This Slide

To ensure more research: To ensure real impact:
Challenge: Separating Perception from Reality

• Interesting Observation in Gunther Schadow’s paper:
  – National ID number viewed as privacy invasion (witness protests here last summer)
  – Cross-database matching will be done
  – Doing it without an ID number makes protecting privacy harder!

  We need to lay the groundwork

What can we do?

• Demonstrate that data mining does not violate privacy
  – Techniques for learning without sharing data
  – Proof of what is (and is not) revealed by results

• Develop measures to quantify privacy
  – Risk of release of data
  – Partial knowledge of data
  – ?

• Ensure the world knows about this
  – Otherwise we’ll get bad laws instead of technical solutions
### Improving Measures: Hiding Rules

- **Does setting disclosure to 0 hide rule?**
  - Will algorithm give 0 support itemset with all sub-itemsets having high support?
  - Is this unusual?
- **Better: “Hidden” when actual support matches**
  - Expected support if no correlation?
  - Expected support if no correlation given known correlation of sub-itemsets after hiding data?
- **We need formal “optimally private” definition**
  - For every type of information we want to hide

### What Research Is Needed?

**To ensure more research**
- Defining “optimally private”
- Ensure that privacy preserving algorithm doesn’t reveal preservation parameters
- Terminology for PPDM
- What information needs to be released to determine impact on validity
- Causality mining
- Advertise workshop results:
  - Send proceedings pointer to Michael Ley (DBLP)
  - add talks to proceedings
  - Workshop writeup to SIGKDD Explorations, SIGMOD Record
- Convince people it is a real problem
  - Privacy makes research hard in
    - U.S. medical community
    - Any individually identifiable data in E.U., Australia
  - Corporate liability/PR issues
    - Corporation loses if someone complains
    - Protecting corporate secrets
  - Big companies take it seriously
- Show researchers we can preserve privacy
  - So they can get data for research
  - So companies won’t give them data unless protected

**To ensure real impact**
- Quantify acceptable privacy in “human” terms
  - k-anonymity is good example
- Tradeoff between releasing details of how privacy achieved and validity of results
  - Algorithms must not be invertible
- Impact on validity vs. impact on privacy
- Terminology for PPDM
  - Privacy definition
  - What is PPDM
- Manage expectation of privacy
- Defining acceptable data mining results given real-world task
- Pave the way
  - Before data mining outlawed (e.g., EU rules about not keeping data after hiding done)
- Make this transparent
  - Access control models
  - Given security policy / access control, derive appropriate mining access controls