10 Challenging Problems in Data Mining Research prepared for ICDM 2005

Edited by
Qiang Yang, Hong Kong Univ. of Sci. & Tech.
and
Xindong Wu, University of Vermont

Contributors

- Pedro Domingos, Charles Elkan, Johannes Gehrke, Jiawei Han, David Heckerman, Daniel Keim,Jiming Liu, David Madigan, Gregory Piatetsky-Shapiro, Vijay V. Raghavan and associates, Rajeev Rastogi, Salvatore J. Stolfo, Alexander Tuzhilin, and Benjamin W. Wah
- A companion document is upcoming…

A New Feature at ICDM 2005

- What are the 10 most challenging problems in data mining, today?
- Different people have different views, a function of time as well
- What do the experts think?
  - Experts we consulted:
    - Previous organizers of IEEE ICDM and ACM KDD
  - We asked them to list their 10 problems (requests sent out in Oct 05, and replies Obtained in Nov 05)
  - Replies
    - Edited into an article: hopefully be useful for young researchers
    - Not in any particular importance order

1. Developing a Unifying Theory of Data Mining

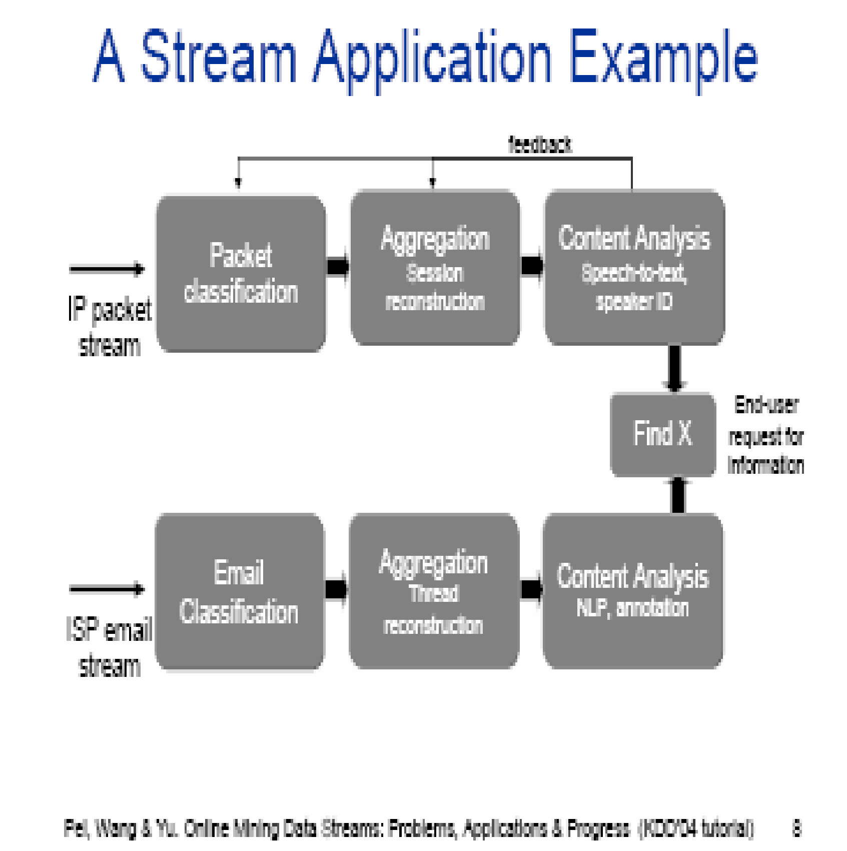
- The current state of the art of data-mining research is too ``ad-hoc“
  - techniques are designed for individual problems
  - no unifying theory
- Needs unifying research
  - Exploration vs explanation
- Long standing theoretical issues
  - How to avoid spurious correlations?
- Deep research
  - Knowledge discovery on hidden causes?
  - Similar to discovery of Newton’s Law?

An Example (from Tutorial Slides by Andrew Moore):

- **VC dimension.** If you've got a learning algorithm in one hand and a dataset in the other hand, to what extent can you decide whether the learning algorithm is in danger of overfitting or underfitting?
  - formal analysis into the fascinating question of how overfitting can happen,
  - estimating how well an algorithm will perform on future data that is solely based on its training set error,
  - a property (VC dimension) of the learning algorithm. VC-dimension thus gives an alternative to cross-validation, called Structural Risk Minimization (SRM), for choosing classifiers.
  - CV,SRM, AIC and BIC.

2. Scaling Up for High Dimensional Data and High Speed Streams

- Scaling up is needed
  - ultra-high dimensional classification problems (millions or billions of features, e.g., bio data)
  - Ultra-high speed data streams
- Streams
  - continuous, online process
  - e.g. how to monitor network packets for intruders?
  - concept drift and environment drift?
  - RFID network and sensor network data



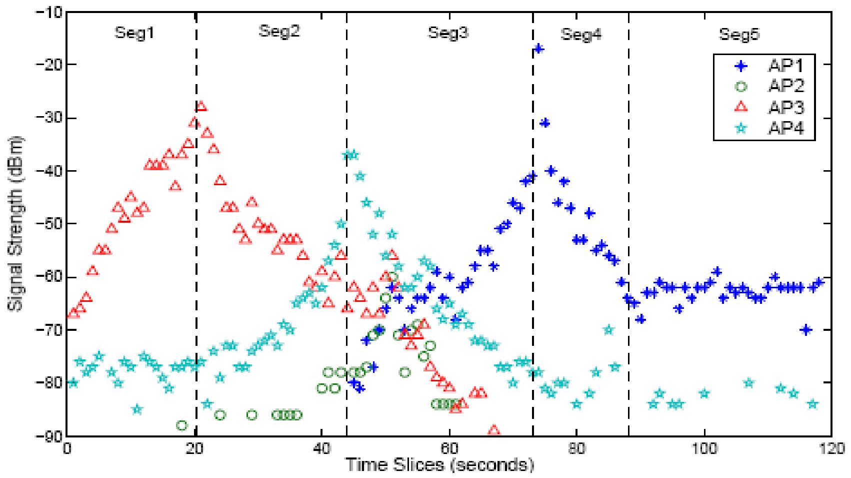
A Stream Application Example

Pei, Wang & Yu. Online Mining Data Streams: Problems, Applications & Progress (KDD'04 tutorial) 8

Excerpt from Jian Pei's Tutorial
http://www.cs.sfu.ca/~jpei/

3. Sequential and Time Series Data

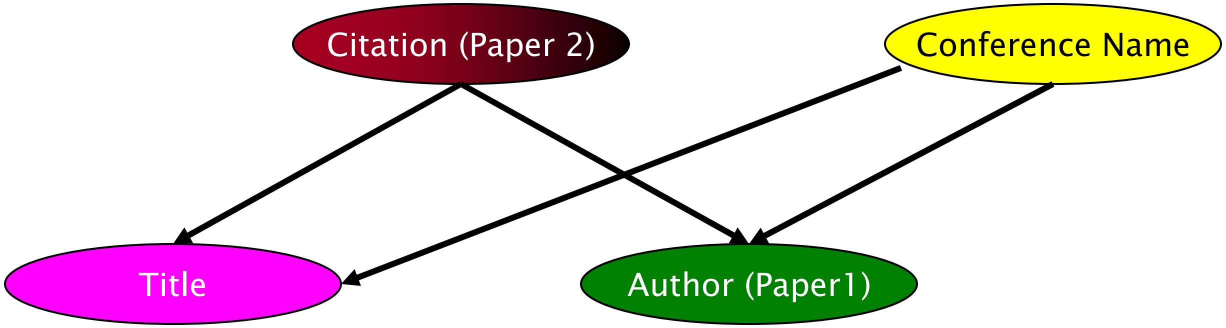
- How to efficiently and accurately cluster, classify and predict the trends ?
- Time series data used for predictions are contaminated by noise
  - How to do accurate short-term and long-term predictions?
  - Signal processing techniques introduce lags in the filtered data, which reduces accuracy
  - Key in source selection, domain knowledge in rules, and optimization methods



Real time series data obtained from Wireless sensors in Hong Kong UST CS department hallway

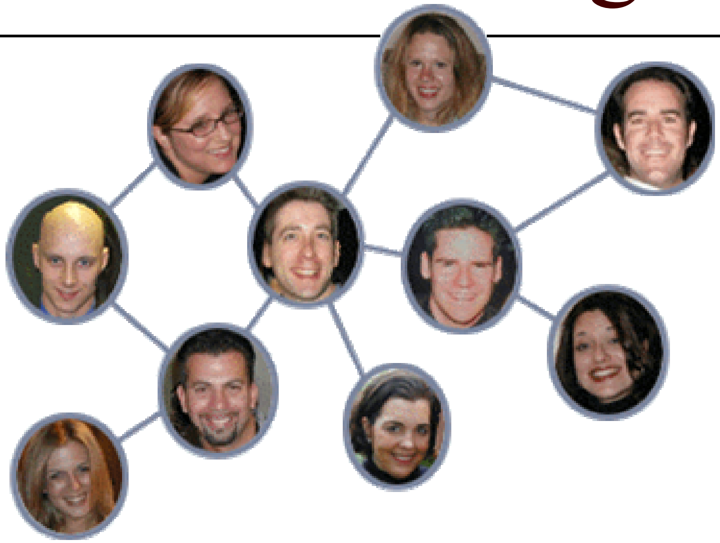
4. Mining Complex Knowledge from Complex Data

- Mining graphs
- Data that are not i.i.d. (independent and identically distributed)
  - many objects are not independent of each other, and are not of a single type.
  - mine the rich structure of relations among objects,
  - E.g.: interlinked Web pages, social networks, metabolic networks in the cell
- Integration of data mining and knowledge inference
  - The biggest gap: unable to relate the results of mining to the real-world decisions they affect - all they can do is hand the results back to the user.
- More research on *interestingness of* knowledge



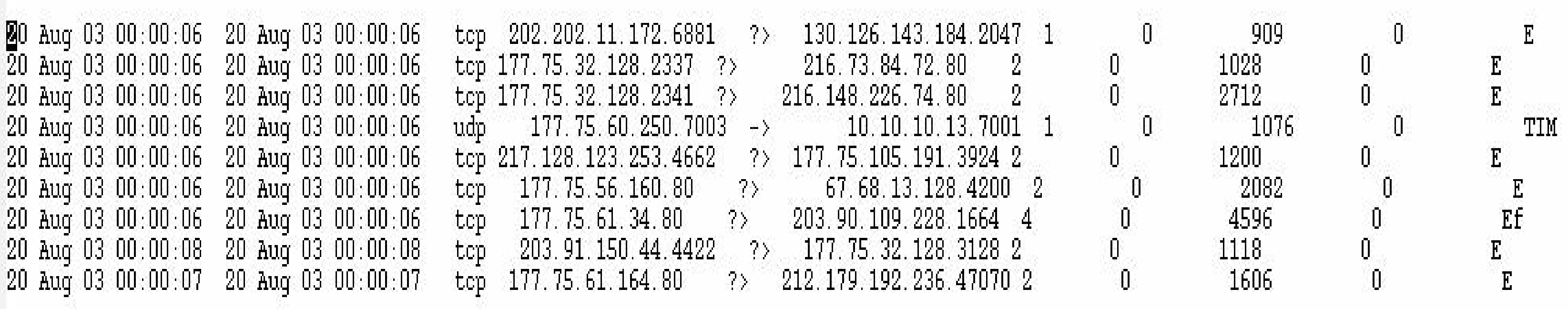
5. Data Mining in a Network Setting

- Community and Social Networks
  - Linked data between emails, Web pages, blogs, citations, sequences and people
  - Static and dynamic structural behavior
- Mining in and for Computer Networks
  - detect anomalies (e.g., sudden traffic spikes due to a DoS (Denial of Service) attacks
  - Need to handle 10Gig Ethernet links (a) detect (b) trace back (c) drop packet



Picture from Matthew Pirretti's slides,penn state

An Example of packet streams (data courtesy of NCSA, UIUC)

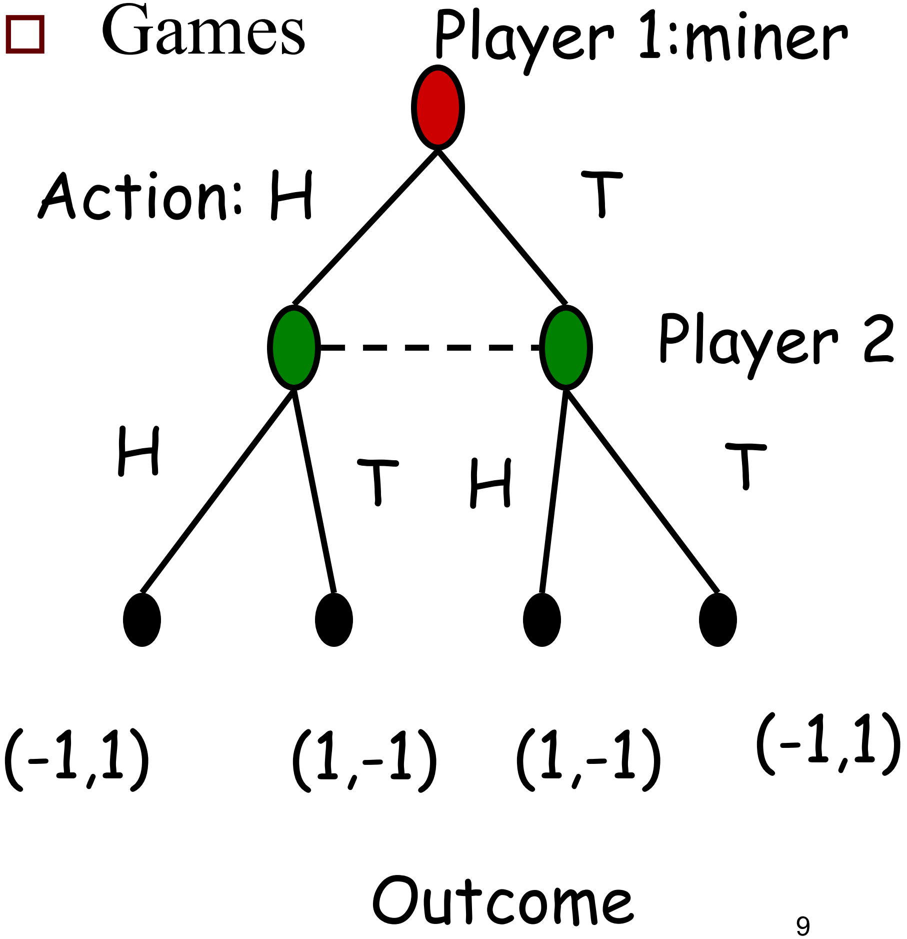


```
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp  202.202.11.172.6881   ?>   130.126.143.184.2047  1        0         909        0         E
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp 177.75.32.128.2337  ?>      216.73.84.72.80    2       0        1028       0        E
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp 177.75.32.128.2341  ?>    216.148.226.74.80    2       0        2712       0        E
20 Aug 03 00:00:06  20 Aug 03 00:00:06   udp    177.75.60.250.7003  ->       10.10.10.13.7001  1        0         1076       0         TIM
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp 217.128.123.253.4662   ?>  177.75.105.191.3924 2       0        1200       0        E
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp   177.75.56.160.80     ?>     67.68.13.128.4200  2       0        2082       0        E
20 Aug 03 00:00:06  20 Aug 03 00:00:06   tcp   177.75.61.34.80     ?>   203.90.109.228.1664  4       0        4596       0        Ef
20 Aug 03 00:00:08  20 Aug 03 00:00:08   tcp   203.91.150.44.4422   ?>   177.75.32.128.3128 2       0        1118       0        E
20 Aug 03 00:00:07  20 Aug 03 00:00:07   tcp  177.75.61.164.80     ?>  212.179.192.236.47070 2       0        1606       0        E
```

6. Distributed Data Mining and Mining Multi-agent Data

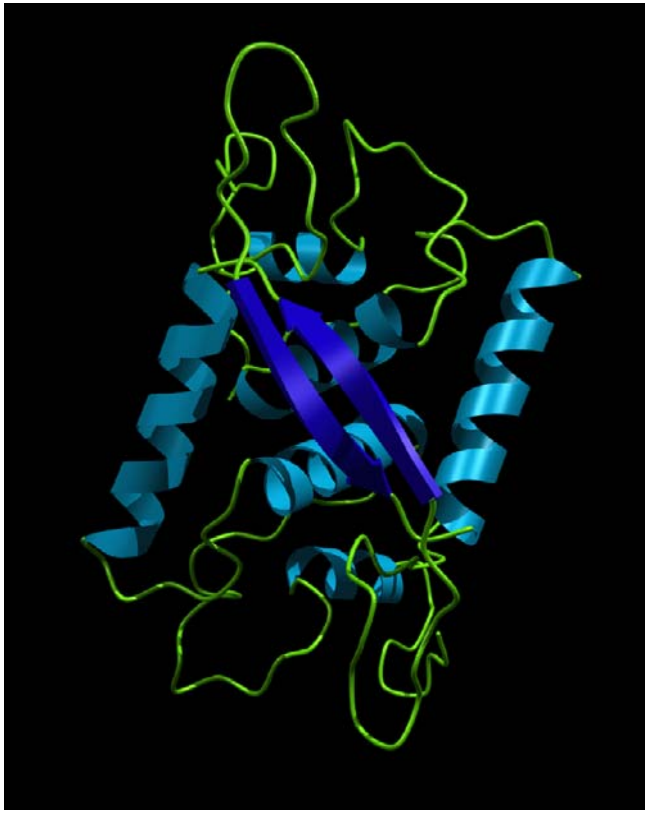
- Need to correlate the data seen at the various probes (such as in a sensor network)
- Adversary data mining: deliberately manipulate the data to sabotage them (e.g., make them produce false negatives)
- Game theory may be needed for help

□ Games



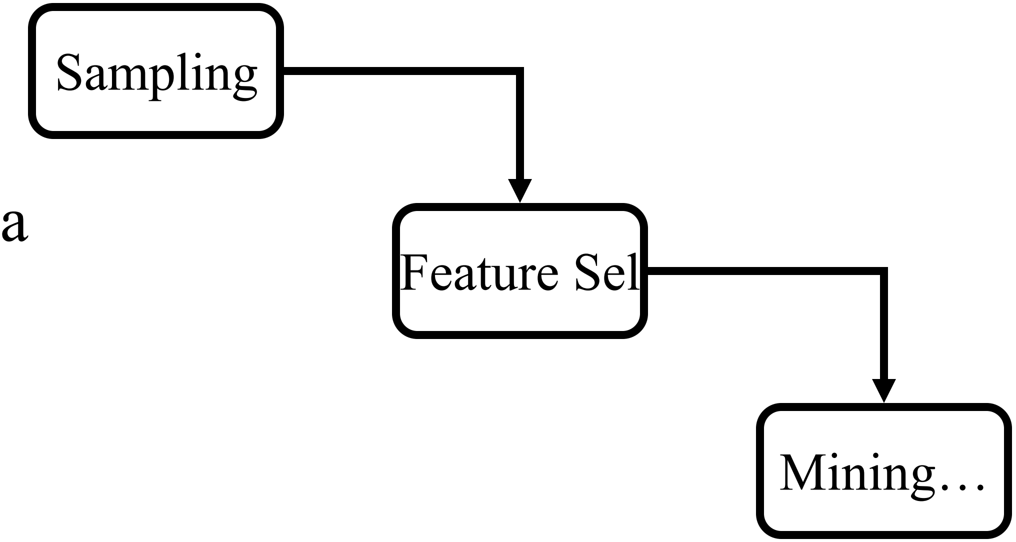
7. Data Mining for Biological and Environmental Problems

- New problems raise new questions
- Large scale problems especially so
  - Biological data mining, such as HIV vaccine design
  - DNA, chemical properties, 3D structures, and functional properties →need to be fused
  - Environmental data mining
  - Mining for solving the energy crisis



8. Data-mining-Process Related Problems

- How to automate mining process?
  - the composition of data mining operations
  - Data cleaning, with logging capabilities
  - Visualization and mining automation



Sampling

Feature Sel

Mining…

- Need a methodology: help users avoid many data mining mistakes
  - What is a canonical set of data mining operations?

9. Security, Privacy and Data Integrity

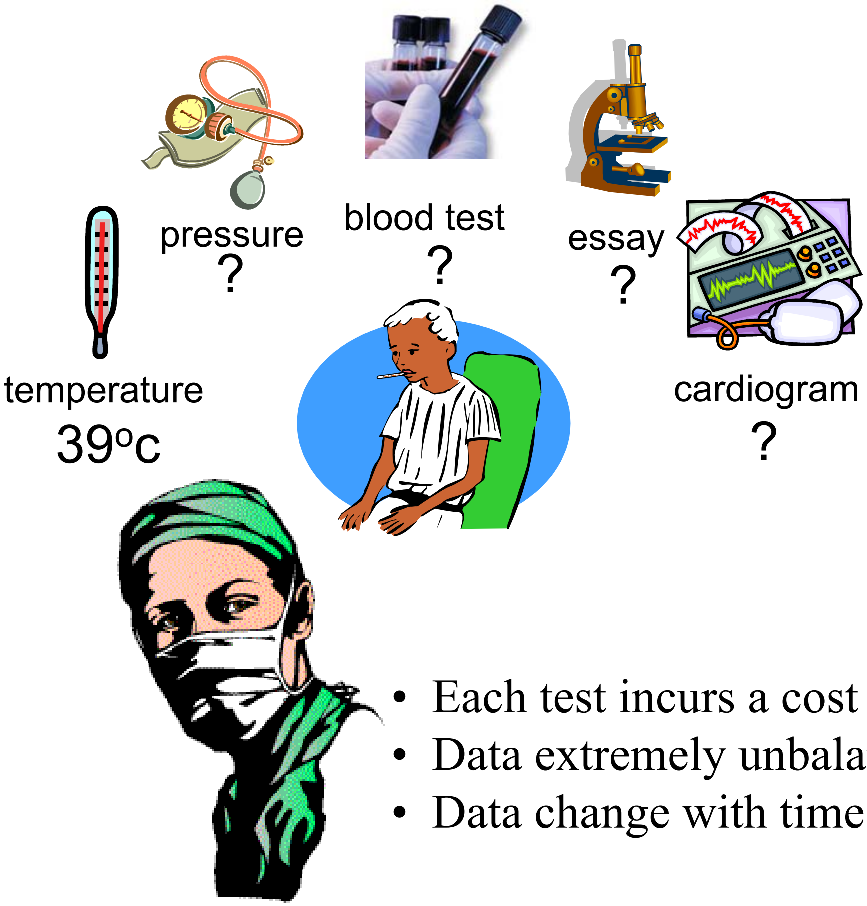
- How to ensure the users privacy while their data are being mined?
- How to do data mining for protection of security and privacy?
- Knowledge integrity assessment
  - Data are intentionally modified from their original version, in order to misinform the recipients or for privacy and security
  - Development of measures to evaluate the knowledge integrity of a collection of
    - Data
    - Knowledge and patterns

http://www.cdt.org/privacy/

Headlines (Nov 21 2005)
Senate Panel Approves Data Security Bill - The Senate Judiciary Committee on Thursday passed legislation designed to protect consumers against data security failures by, among other things, requiring companies to notify consumers when their personal information has been compromised. While several other committees in both the House and Senate have their own versions of data security legislation, S. 1789 breaks new ground by including provisions permitting consumers to access their personal files …

10. Dealing with Non-static, Unbalanced and Cost-sensitive Data

- The UCI datasets are small and not highly unbalanced
- Real world data are large (10^5 features) but only < 1% of the useful classes (+'ve)
- There is much information on costs and benefits, but no overall model of profit and loss
- Data may evolve with a bias introduced by sampling



- Each test incurs a cost
- Data extremely unbalanced
- Data change with time

Summary

1. Developing a Unifying Theory of Data Mining
2. Scaling Up for High Dimensional Data/High Speed Streams
3. Mining Sequence Data and Time Series Data
4. Mining Complex Knowledge from Complex Data
5. Data Mining in a Network Setting
6. Distributed Data Mining and Mining Multi-agent Data
7. Data Mining for Biological and Environmental Problems
8. Data-Mining-Process Related Problems
9. Security, Privacy and Data Integrity
10. Dealing with Non-static, Unbalanced and Cost-sensitive Data