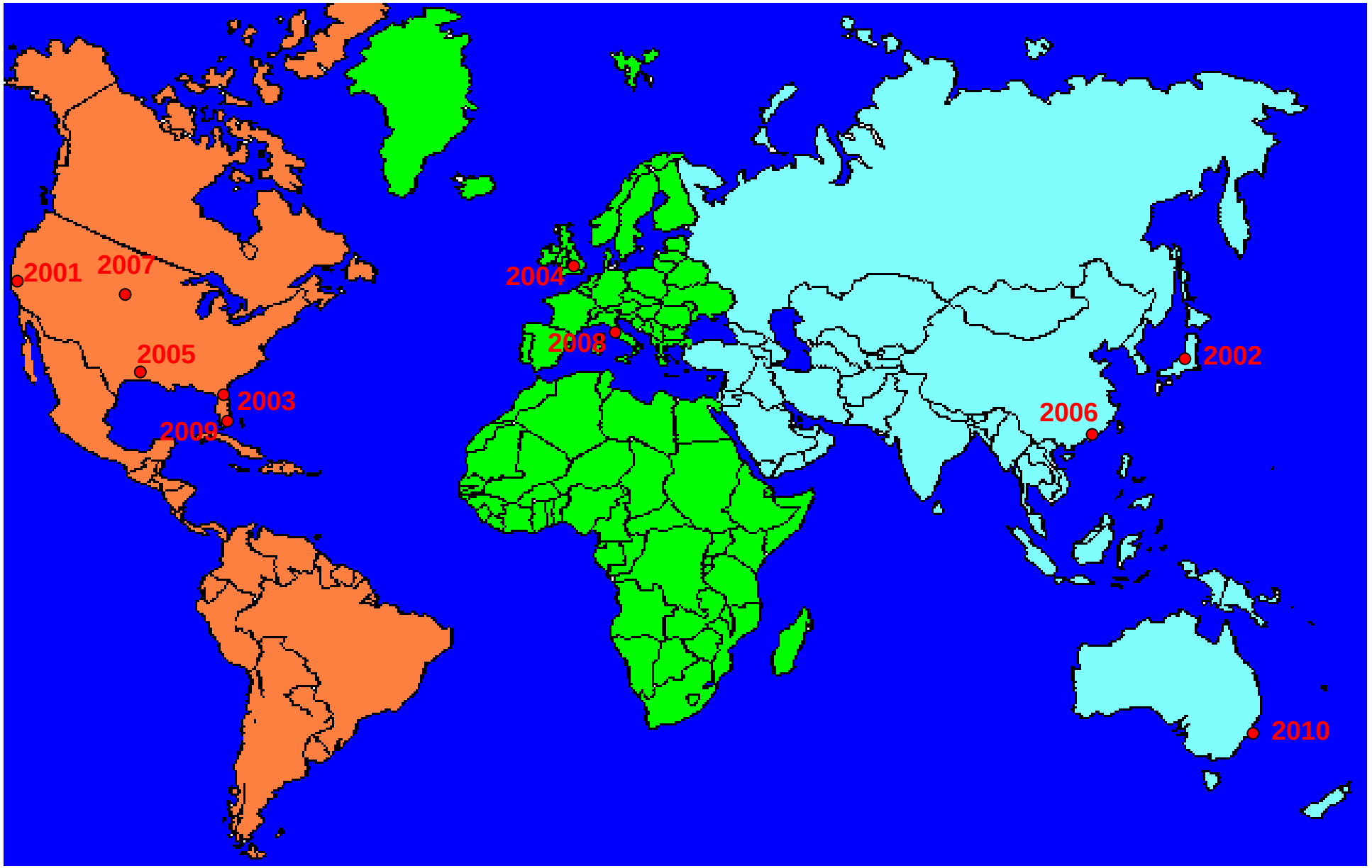
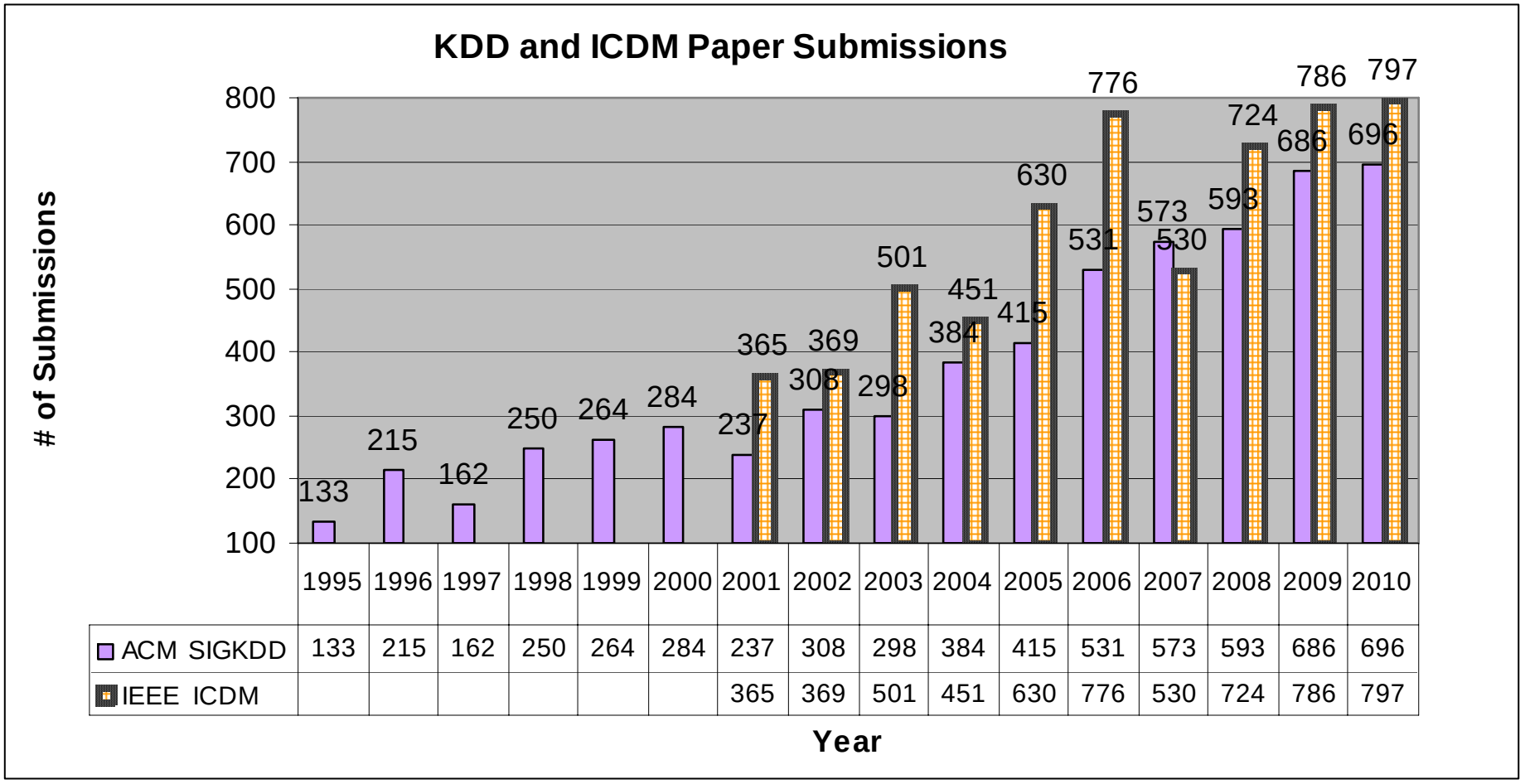
10 Years of Data Mining Research: Retrospect and Prospect

Xindong Wu (吴信东)
Department of Computer Science
University of Vermont, USA
【中国】合肥工业大学计算机与信息学院



2001 (11/29-12/02) San Jose, California, USA
2003 (11/19-11/22) Melbourne, Florida, USA
2005 (11/27-11/30) Houston, Texas, USA
2007 (10/28-10/31) Omaha, Nebraska, USA
2009 (12/06-12/09) Miami, Florida, USA

2002 (12/09-12/12) Maebashi City, Japan
2004 (11/01-11/04) Brighton, United Kingdom
2006 (12/18-12/22) Hong Kong, China
2008 (12/15-12/19) Pisa, Italy
2010 (12/14-12/17) Sydney, Australia


KDD and ICDM Paper Submissions
# of Submissions
800
700
600
500
400
300
200
100
133
215
162
250
264
284
237
308
298
384
415
531
573
593
686
696
365
369
501
451
630
776
530
724
786
797
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
ACM SIGKDD 133 215 162 250 264 284 237 308 298 384 415 531 573 593 686 696
IEEE ICDM 365 369 501 451 630 776 530 724 786 797
Year

Outline

1. **Top 10 Activities in the Past 10 Years**
2. **10 Data Mining Topics**
3. **The Top 10 Algorithms**
4. **10 Challenging Problems in Data Mining Research**

Top 10 ICDM Activities in the Past 10 Years

1. KAIS journal publications of the best papers every year since 2001
2. IEEE ICDM Research Contributions Award and IEEE ICDM Outstanding Service Award since 2001
3. A panel discussion on "how research meets practical development" in 2001
4. Data mining on ICDM submissions since 2003
5. Identifying "10 challenging problems in data mining research" in 2005
6. A panel discussion on "top 10 algorithms in data mining" in 2006
7. Workshops proceedings by IEEE CPS since 2006
8. Double blind reviewing since 2007 (likely triple blind since 2011)
9. A panel discussion on "top 10 data mining case studies" in 2010
10. Publishing both Regular Papers and Short Papers since 2001

Outline

1. Top 10 Activities in the Past 10 Years
2. **10 Data Mining Topics**
3. The Top 10 Algorithms
4. 10 Challenging Problems in Data Mining Research

10 Data Mining Topics

- Classification
  - #1. **C4.5**: Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.
  - #2. **CART**: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
  - #3. ***K* Nearest Neighbours (*k*NN):** Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.
  - #4. **Naive Bayes**: Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
  - #5. **SVM:** Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.
  - #6. **EM**: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.
- Association Analysis
  - #7. **Apriori**: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - #8. **FP-Tree**: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
- Link Mining
  - #9. **PageRank**: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - #10. **HITS**: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

10 Data Mining Topics (2)

- **Clustering**
  - #11. ***K*-Means**: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - #12. **BIRCH**: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- **Bagging and Boosting**
  - #13. **AdaBoost**: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.
- **Sequential Patterns**
  - #14. **GSP**: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #15. **PrefixSpan**: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- **Integrated Mining**
  - #16. **CBA**: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- **Rough Sets**
  - #17. **Finding reduct**: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992.
- **Graph Mining**
  - #18. **gSpan**: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Outline

1. Top 10 Activities in the Past 10 Years
2. 10 Data Mining Topics
3. **The Top 10 Algorithms**
   (joint efforts with Vipin Kumar)
4. 10 Challenging Problems in Data Mining Research

The Top 10 Algorithms

- #1: C4.5
- #2: *K*-Means
- #3: SVM
- #4: Apriori
- #5: EM
- #6: PageRank
- #7: AdaBoost
- #7: *k*NN
- #7: Naive Bayes
- #10: CART

4. 10 Challenging Problems in Data Mining Research

Joint Efforts with Qiang Yang (Hong Kong Univ. of Sci. & Tech.)
With Contributions from ICDM & KDD Organizers

1. Developing a Unifying Theory of Data Mining

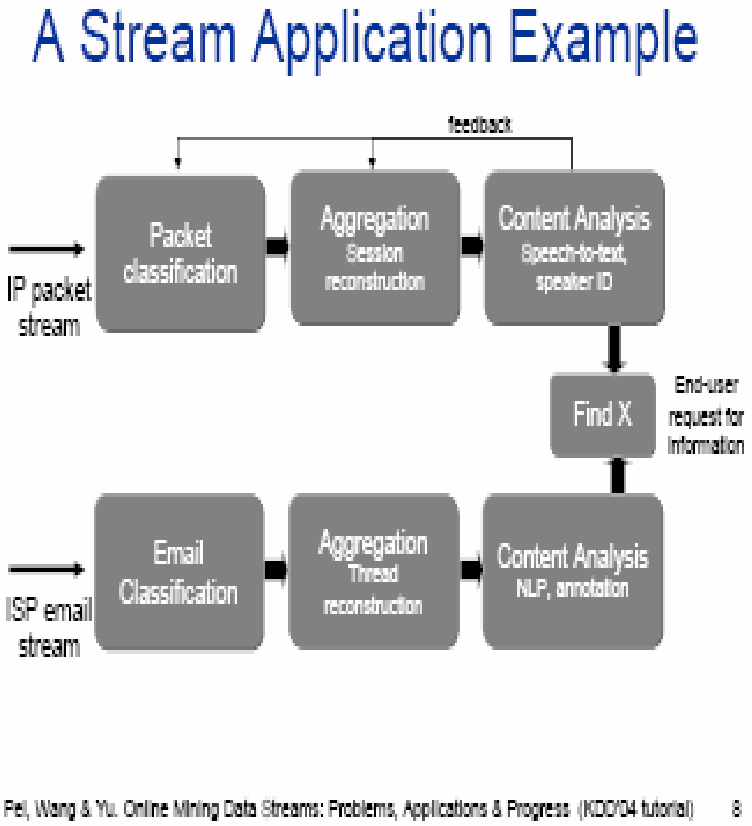
- The current state of the art of data-mining research is too ``ad-hoc“
  - techniques are designed for individual problems
  - no unifying theory
- Needs unifying research
  - Exploration vs. explanation
- Long standing theoretical issues
  - How to avoid spurious correlations?
- Deep research.
  -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

An Example (from VC Dimension tutorial slides by Andrew Moore):

- 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

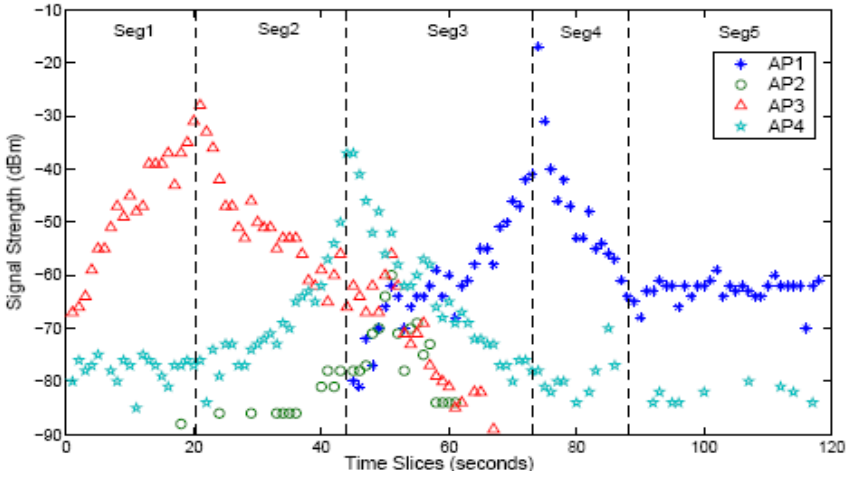
- Scaling up is needed
  - Ultra-high dimensional classification problems (millions or billions of features, e.g., bio data)
  - Ultra-high speed data streams
- Streams
  - Continuous, online process
  - e.g. how to monitor network packets for intruders?
  - Concept drift and environment drift?
  - RFID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



Excerpt from Jian Pei's Tutorial

3. Sequential and Time Series Data

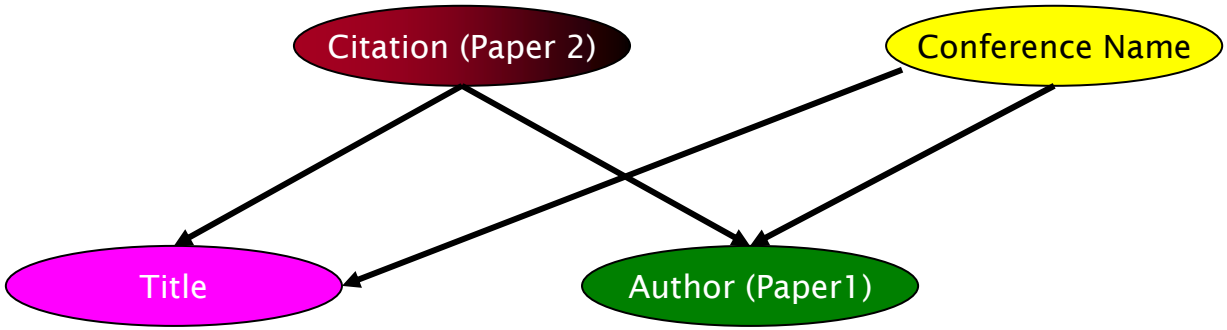
- How to efficiently and accurately cluster, classify and predict the trends ?
- Time series data used for predictions are contaminated by noise.
  - How to do accurate short-term and long-term predictions?
  - Signal processing techniques introduce lags in the filtered data, which reduces accuracy
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

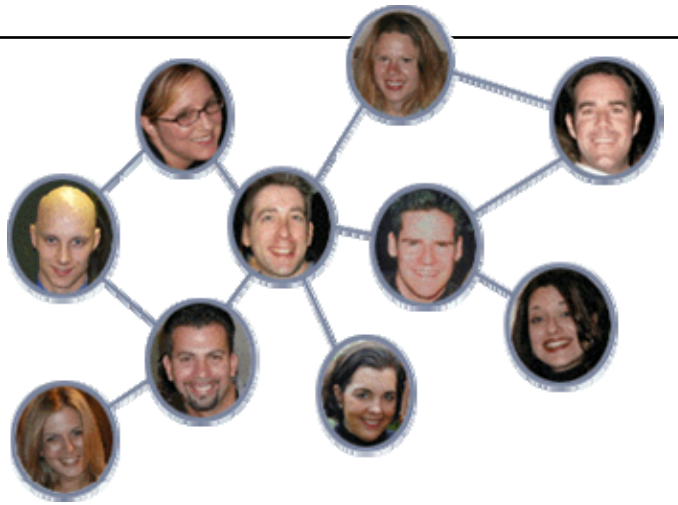
4. Mining Complex Knowledge from Complex Data

- 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.

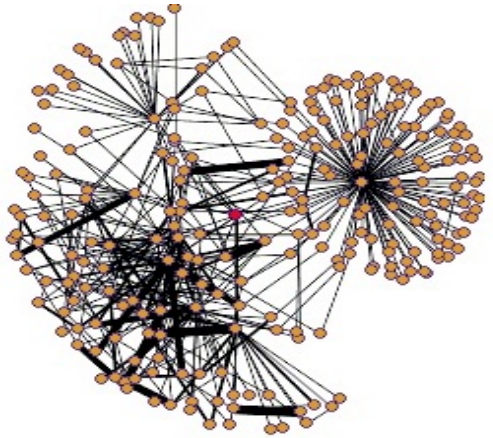


5. Data Mining in Graph Structured Data

- Community and Social Networks
  - Linked data between emails, Web pages, blogs, citations, sequences and people
  - Static and dynamic structural behavior
- Pattern Discovery and Modeling for Complex Networks (such as protein interaction networks)
  - Network backbone structure (or micro-structure) discovery
    - E.g., Frequent Subgraph Discovery (Kuramochi & Karypis, ICDM-01)
    - gSpan: Graph-Based Substructure Pattern Mining (Yan & Han, ICDM-02).



Picture from Matthew Pirretti's slides, Penn State

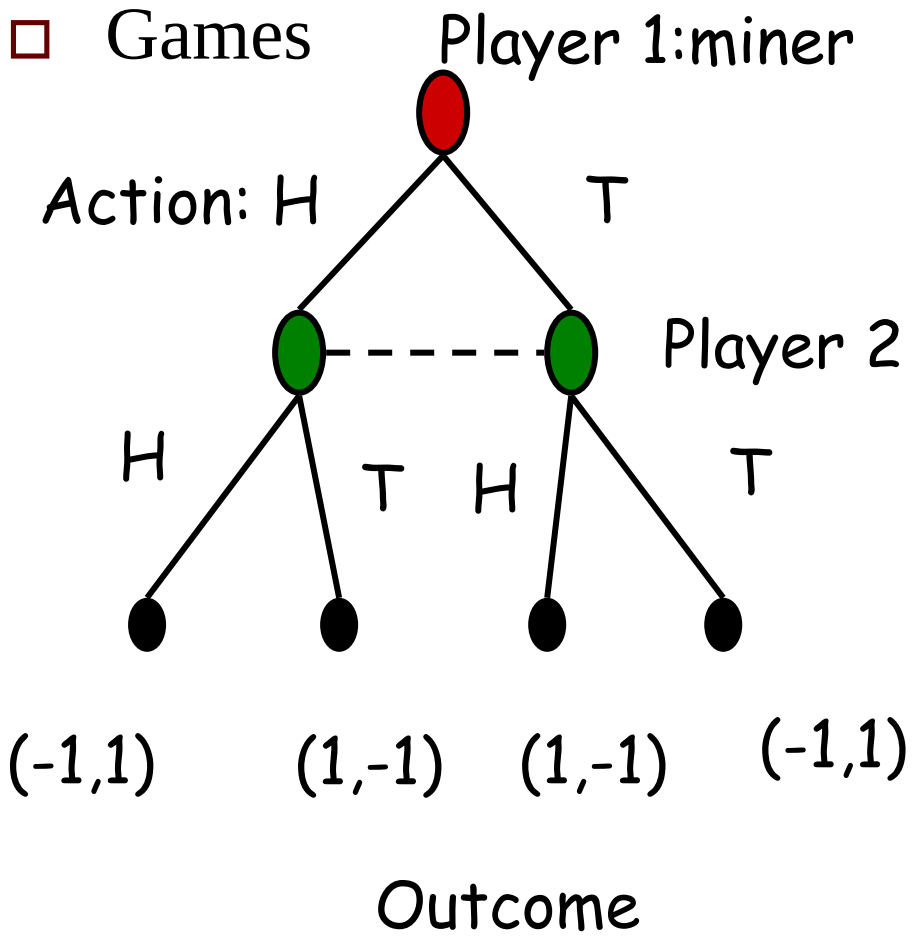


Protein Interaction Graph

http://dip.doe-mbi.ucla.edu/dip/

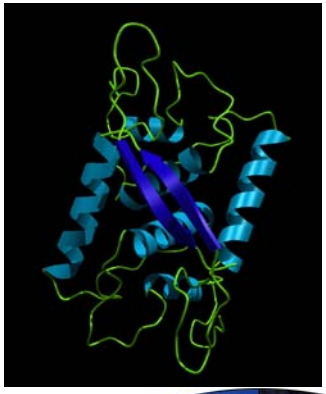
6. Distributed Data Mining and Mining Multi-agent Data

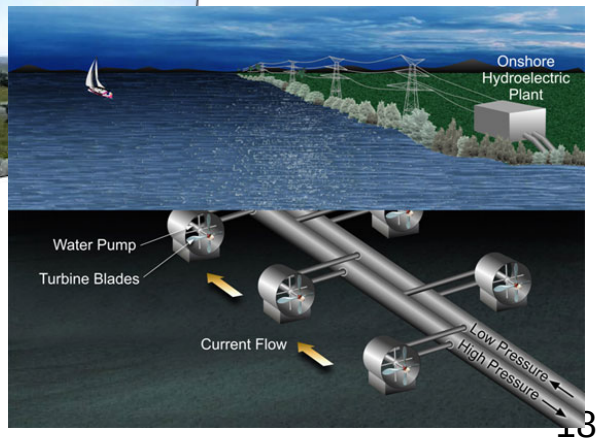
- Need to correlate the data seen at the various probes (such as in a sensor network)
- Adversary data mining: deliberately manipulate the data to sabotage them (e.g., make them produce false negatives)
- Game theory may be needed for help.



7. Data Mining for Biological and Environmental Problems

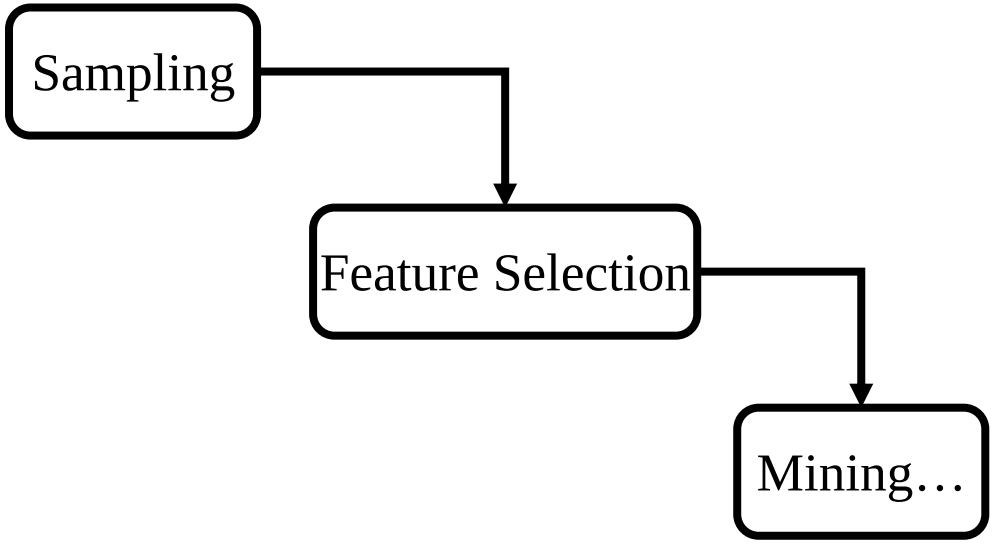
- Emerging domains raise new problems, which lead to new questions
- Large scale problems especially so
  - Biological data mining, such as HIV vaccine design
  - DNA, chemical properties, 3D structures, and functional properties → need to be fused
  - Environmental data mining
  - Mining for solving 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.



- 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

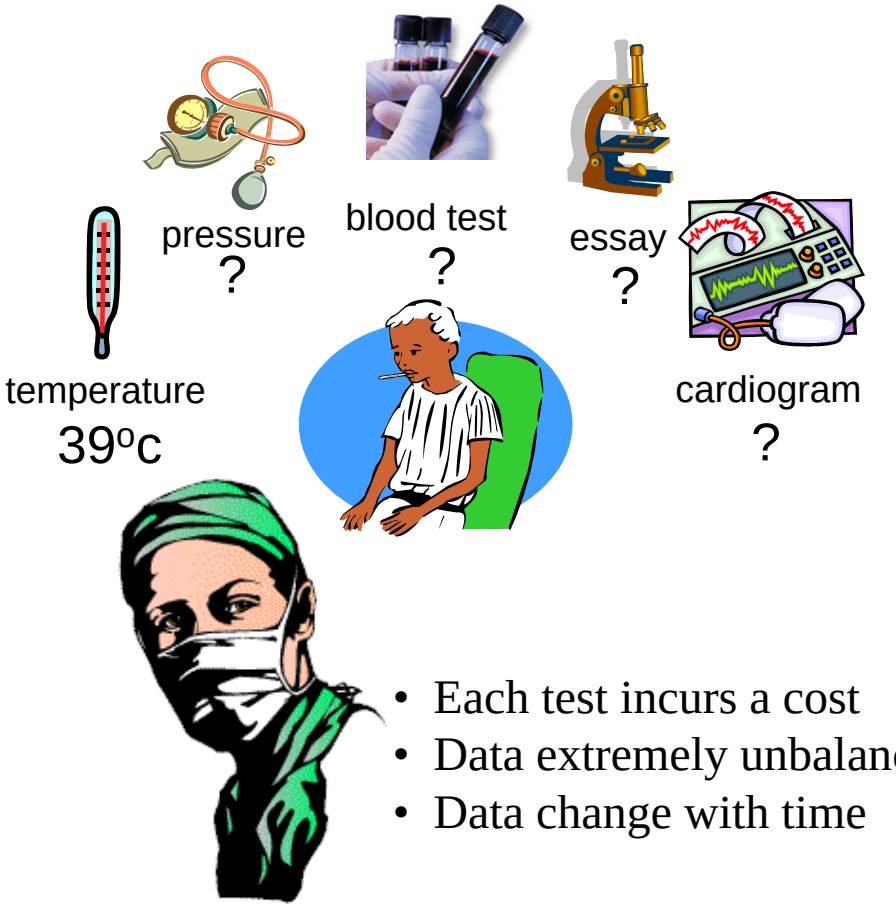
- How to ensure the users privacy while their data are being mined?
- How to do data mining for protection of security and privacy?
- Knowledge integrity assessment.
  - Data are intentionally modified from their original version, in order to misinform the recipients or for privacy and security
  - Development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

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

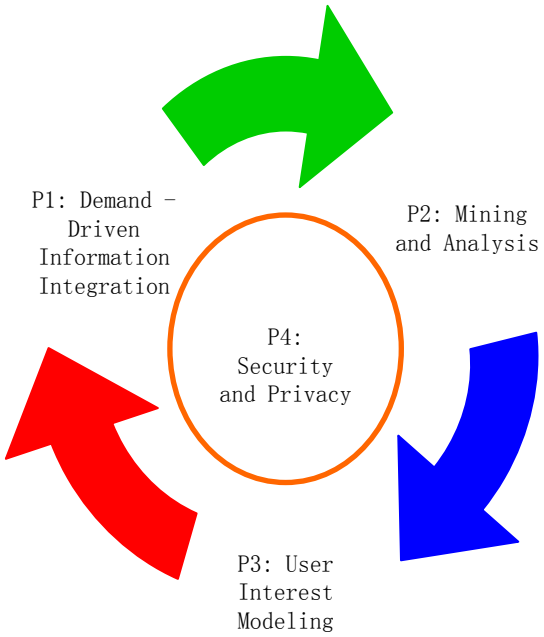
Headlines (Nov 21 2005)
Senate Panel Approves Data Security Bill - The Senate Judiciary Committee on Thursday passed legislation designed to protect consumers against data security failures by, among other things, requiring companies to notify consumers when their personal information has been compromised. 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.



Beyond 10 Challenging Problems: Active Information Fusion



A positive cycle

with

P1: Demand-driven integration of information sources

P2: Mining and analysis

P3: User interest modeling

P4: Security and privacy.

10 Challenging Problems: Recap

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 Graph Structured Data
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

Open Questions

- **Will the challenging problems become more challenging?**
  - **Yes, more research will find new problems.**
- **Will the top-10 algorithms change in the future?**
  - **Sure, let's design better ones to replace them in the next 10 years!**