Research Challenges for Data Mining in Science and Engineering

Jiawei Han

Department of Computer Science

University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

October 17, 2007

Why Data Mining in Science and Engineering?

- **Data is mounting up rapidly and is available!**
  - Giga-bytes → terabytes → peta-bytes in very fast pace
  - Data collection and dissemination tools: New generation of science and engineering equipments, sensors, Web
  - Gigantic data banks from genomics to astronomy: widely available!
- **Data mining: An imminent need in science and engineering**
  - Every discipline: "data poor" → "data rich" → "data richer" → …
  - Jim Gray: "We are in the era of **Data Science**"
  - Diversity of science/engineering data: Data streams, RFID, sensor networks, video/audio data, text and Web, computer systems & software, information networks, and biological/environmental data
  - Discoveries are often of high value

Data Mining Has Been Flourishing in S&E

- Statistical analysis, pattern recognition, machine learning, and data visualization have been popular tools in S & E
- Data mining research has brought in many new, fresh ideas and methods
  - Scalable mining methods
  - Pattern mining and analysis methods
  - Integrated information processing infrastructure: Integrated with database systems, data warehouses, and the Web
  - Invisible data mining: mining as a natural, hidden process
- There are still many open research problems

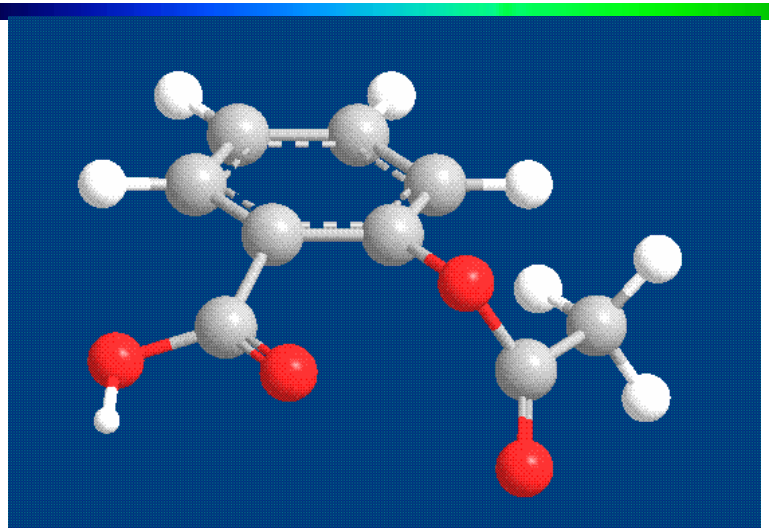
Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

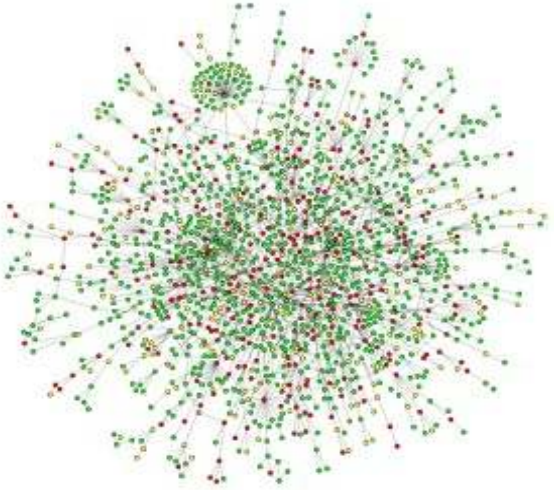
Exploring the Power of Links for Data Mining

- **Google: A shining example of the exploration of links**
  - PageRank and HITS: Identify authoritative Web pages and hubs
- **Science and engineering: Links are everywhere**
  - Beyond explicit homogeneous links: page links, computer network, …
  - Implictily, heterogeneou, multidimensional links: data relations and cross-relational “bridges”, gene/protein, wireless battlefield, pub, …
  - Knowledge and patterns: Hidden in massive links
- **The Power of links has been demonstrated in various tasks**
  - CrossMine: Classification of multi-relations by link analysis
  - CrossClus: Cross-relational clustering with user’s guidance
  - LinkClus: Efficient link-based clustering by exploring the power law distribution
  - Distinct: Distinguishing objects with identical names by link analysis
  - Veracity: Reliable facts with multiple conflicting information providers

Mining Graphs and Information Networks

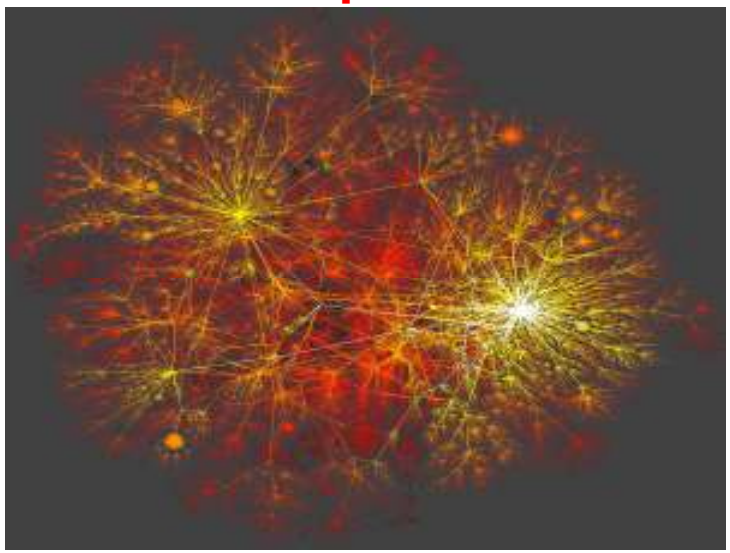


Aspirin

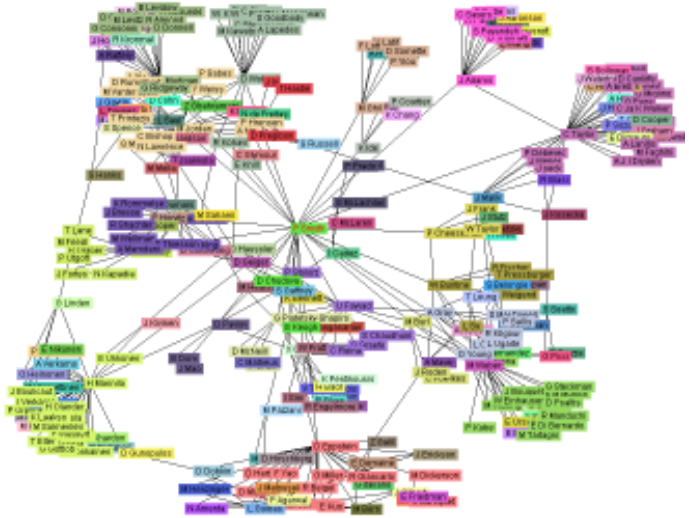


from H. Jeong et al Nature 411, 41 (2001)

Yeast protein interaction network



An Internet Web



Co-author network

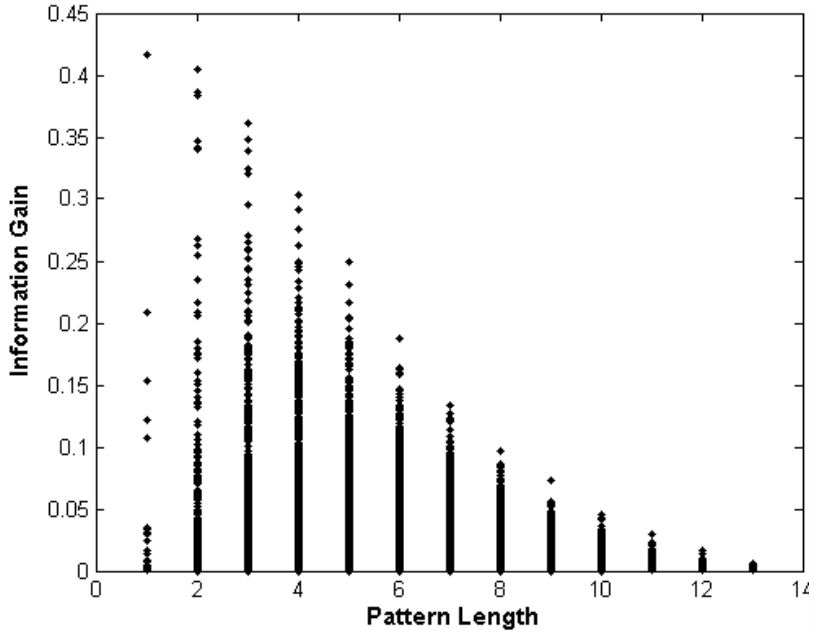
Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Pattern Mining, Pattern Usage and Pattern Understanding

- **Exploration of new, application-oriented pattern mining methods**
  - Applications, applications, applications!
  - Pattern exploration in software bug mining: frequent, sequential patterns
  - Mining *colossal* (i.e., rather large) patterns for bio-pattern analysis
  - Mining *approximate* substructures in large networks/graphs
- **Pattern usage**
  - Effective classification by frequent, discriminative pattern analysis
  - Indexing and substructure similarity search in structures
- **Pattern Understanding**
  - Too many patterns!? — redundancy-aware top-k patterns
  - Semantic annotation of frequent patterns

Discriminative Frequent Pattern Analysis for Effective Classification [ICDE'07]



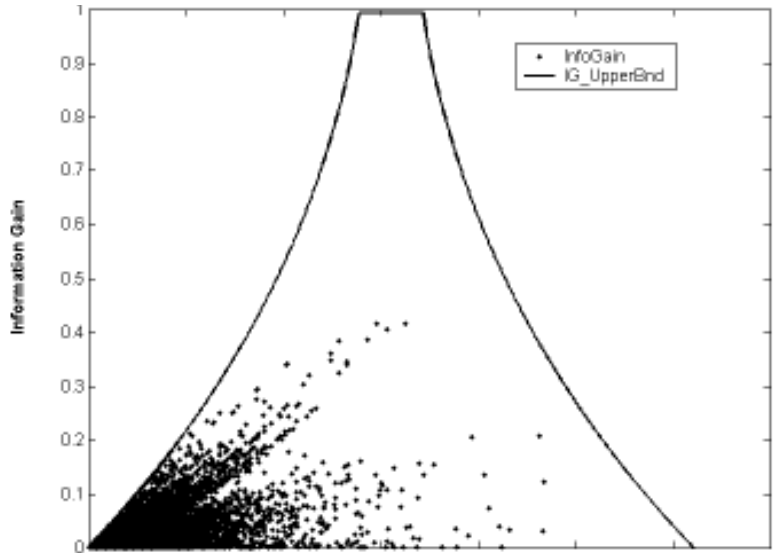


Table 1. Accuracy by SVM on Frequent Combined Features vs. Single Features

Data	Single Feature			Freq. Pattern	
	Item_All	*Item_FS*	*Item_RBF*	*Pat_All*	*Pat_FS*
anneal	**99.78**	**99.78**	99.11	99.33	99.67
austral	85.01	85.50	85.01	81.79	**91.14**
auto	83.25	84.21	78.80	74.97	**90.79**
breast	97.46	97.46	96.98	96.83	**97.78**
cleve	84.81	84.81	85.80	78.55	**95.04**
diabetes	74.41	74.41	74.55	77.73	**78.31**
glass	75.19	75.19	74.78	79.91	**81.32**
heart	84.81	84.81	84.07	82.22	**88.15**
hepatic	84.50	89.04	85.83	81.29	**96.83**
horse	83.70	84.79	82.36	82.35	**92.39**
iono	93.15	94.30	92.61	89.17	**95.44**
iris	94.00	**96.00**	94.00	95.33	**96.00**
labor	89.99	91.67	91.67	94.99	**95.00**
lymph	81.00	81.62	84.29	83.67	**96.67**
pima	74.56	74.56	76.15	76.43	**77.16**
sonar	82.71	86.55	82.71	84.60	**90.86**
vehicle	70.43	72.93	72.14	73.33	**76.34**
wine	98.33	99.44	98.33	98.30	**100**
zoo	97.09	97.09	95.09	94.18	**99.00**

Research Challenges

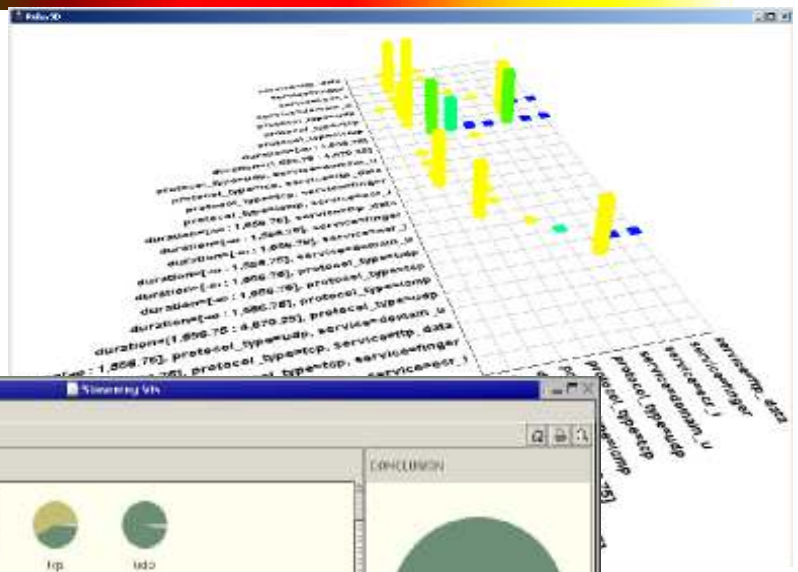
- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

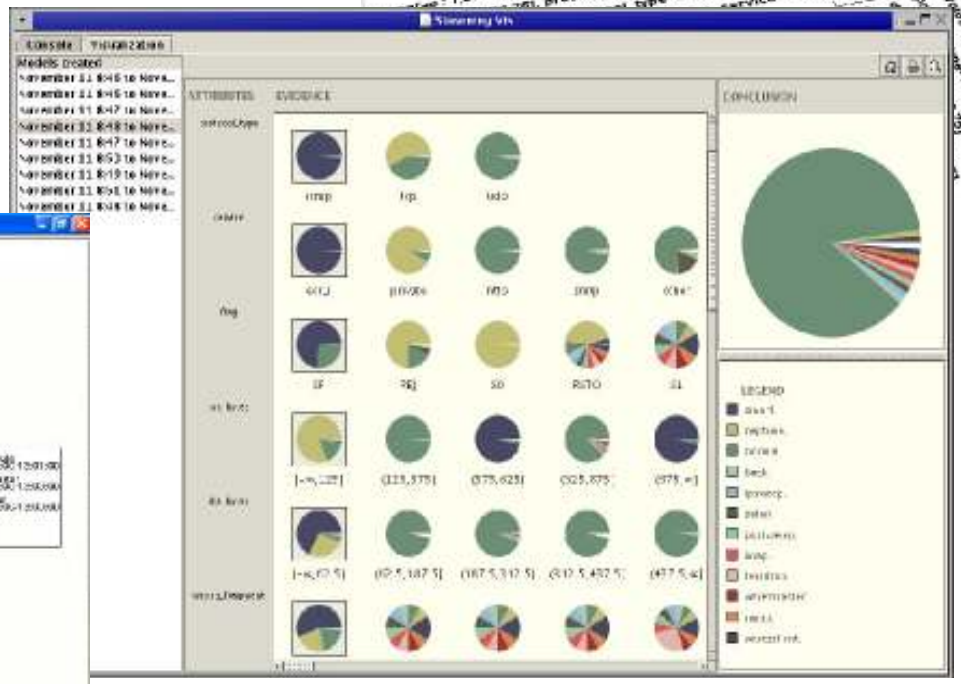
Stream Data Mining

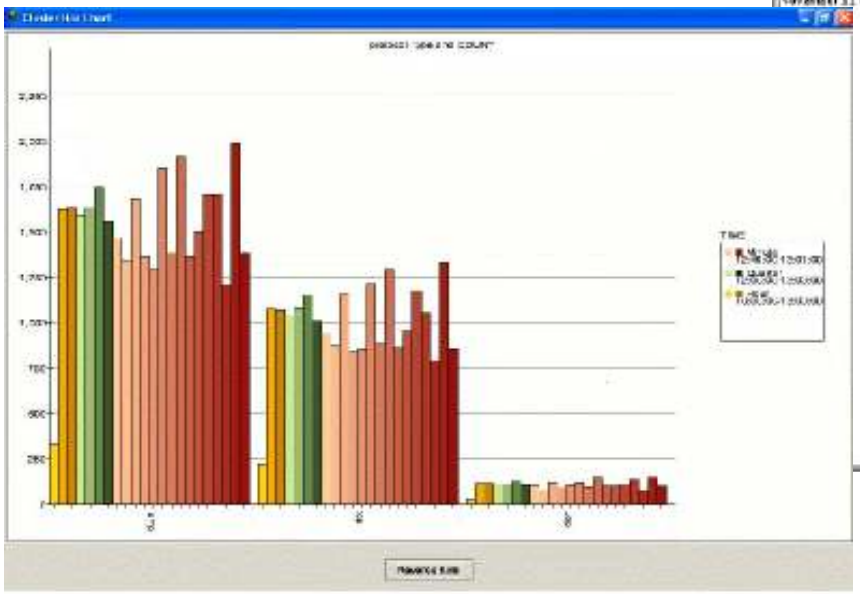
- **Streams are common in science, engineering, and industry**
  - Scientific experiments/observations, production processes
  - Data could be too huge to be scanned multiple times
- **Not just DSMS (data stream management systems)**
- **More on Stream OLAP and stream data mining**
  - Stream data cube: fading model, evolution, summary, and multidimensional drilling
  - Stream sample counting and frequent pattern analysis
  - Classification of data streams: rare events, unexpected distributions
  - Clustering evolving data streams: microclusters vs. macroclusters

MAIDS: Mining Alarming Incidents from Data Streams [SIGMOD'04 demo]

- Stream cubing: Tilted time frame, H-tree structure, online partial aggregation
- Mining evolving clusters in stream data
- Stream classification
- Stream FP analysis
- Mining network intrusion data







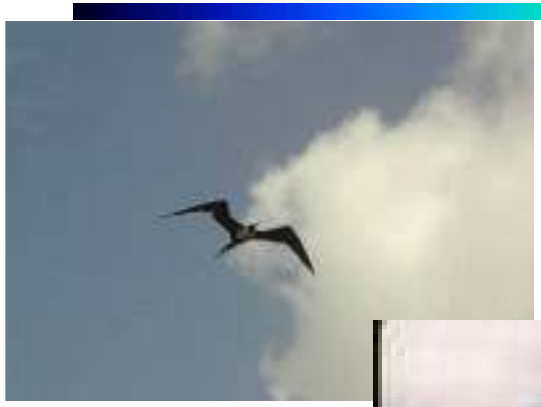
Research Challenges

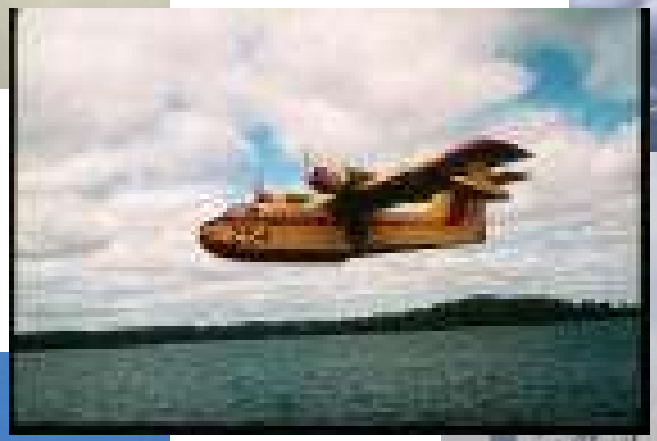
- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Mining Moving Object Data, RFID Data, and Data from Sensor Networks

- **RFID/sensor data warehousing and mining**
  - Deep compression leads to lean and powerful warehouse
  - Paths are important: FlowCube: Multi-Dimensional Analysis of Commodity Flows
  - Mining should explore motifs, paths, traffic, and layers
- **Mining moving object data and trajectories**
  - Classification & outlier detection in moving objects
  - Clustering and classification of trajectory data
- **Mining integrated networks, e.g., sensor network + information network**
  - Medical sensor monitoring and information networks

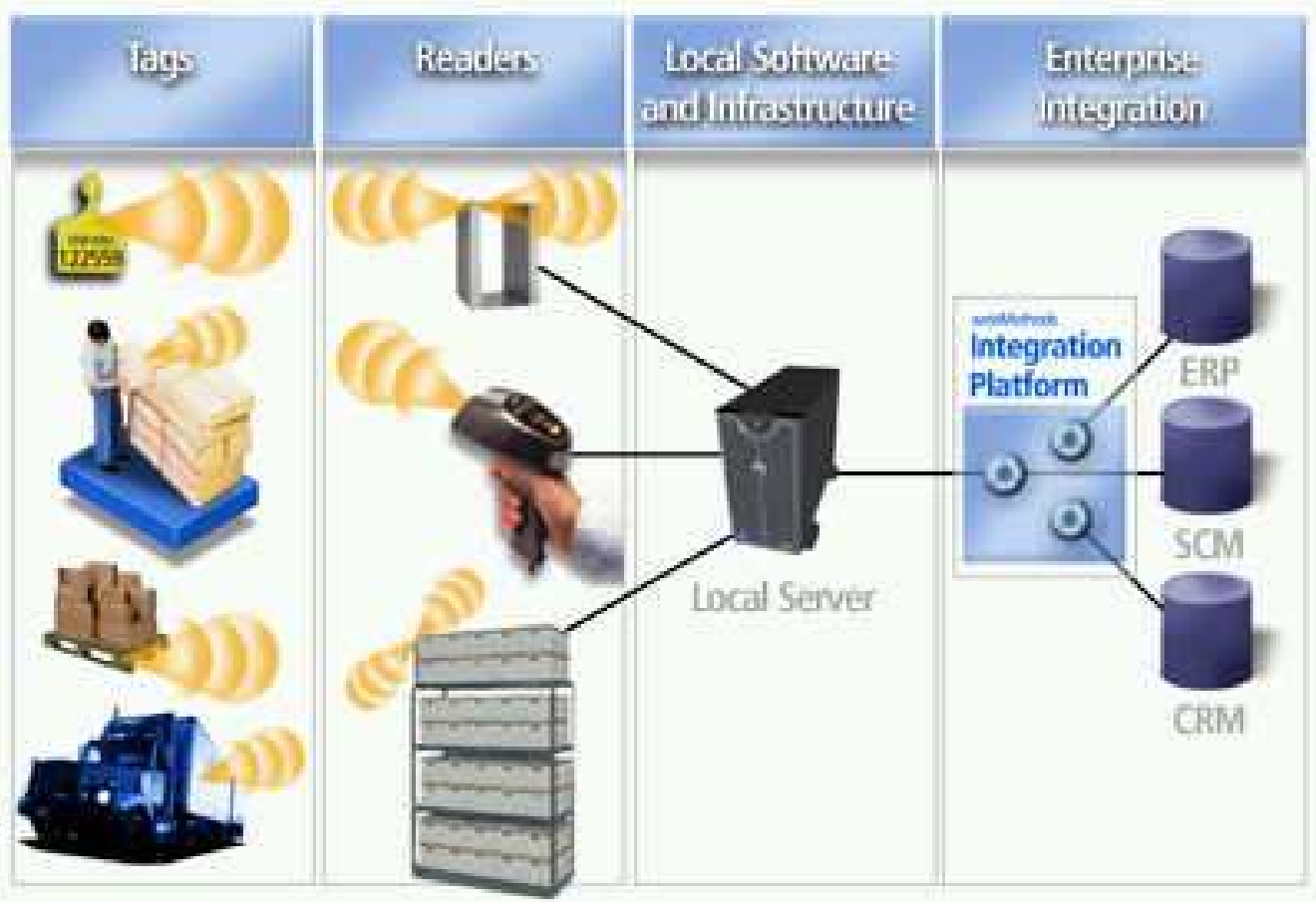
Typical Unrestricted Moving Paths





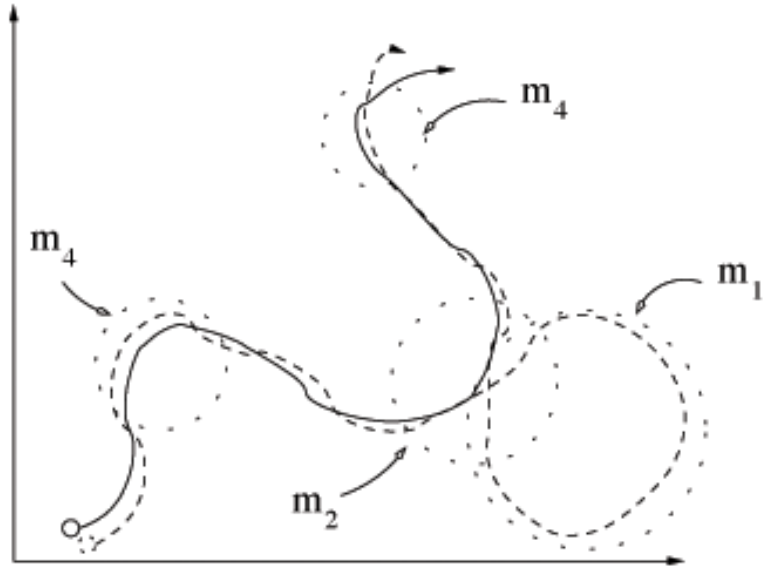
Typical Restricted Moving Paths

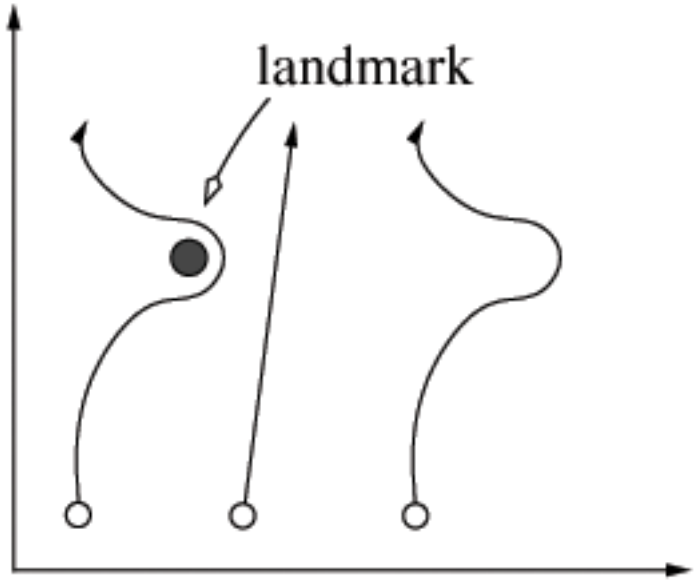
Typical Scattered Movement Traces



Motion-Alert: Automatic Detection of Anomalies in Massive Moving Objects [SDM'07]

- **Extraction of movement fragments as motifs**
  - Trajectories are converted to sets of motif expressions
- **Combination of motif expressions with other multidimensional features**
- **Motif-based feature extraction and clustering**
- **Micro-cluster-based classification to find anomalies**





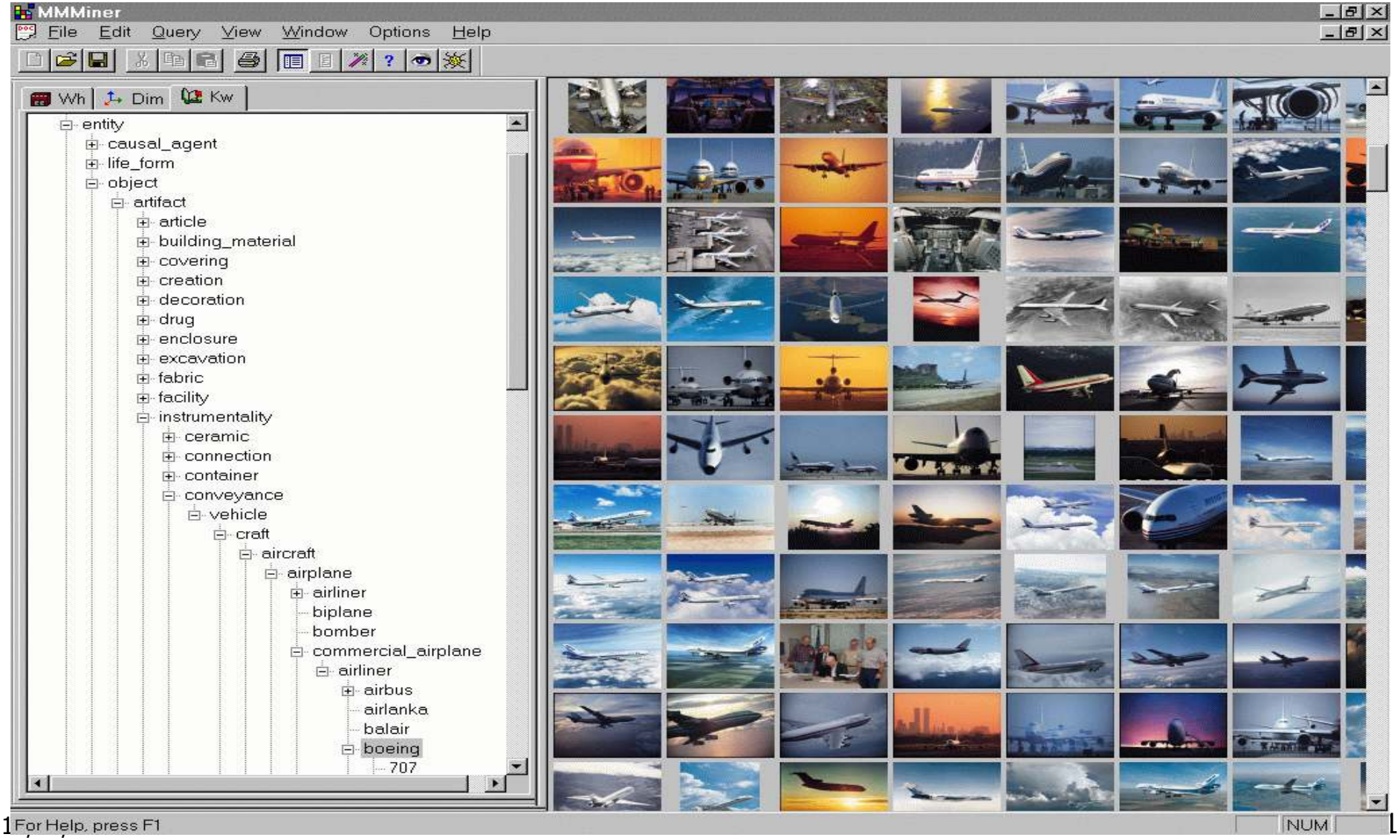
Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

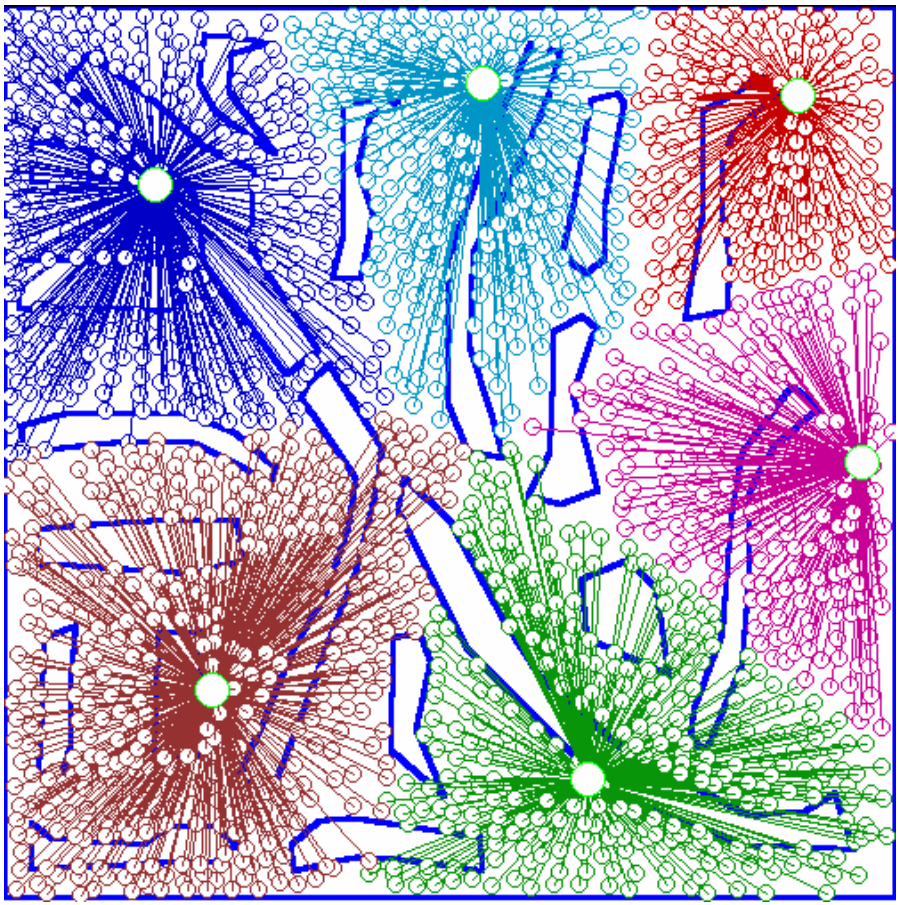
Spatiotemporal and Multimedia Data Mining

- **Spatiotemporal and multimedia data**
  - Digital photos, videos, YouTube, GoogleMaps, weather services, satellite images, Digital Earth, traffic monitor, …
- **Abundant applications**
  - Detect forest fire, predict hurricane path, weather patterns, global warming, traffic routing, …
- **Warehouse spatiotemporal and multimedia data**
  - More dimensions: Spatial, temporal, color, shape, relationships, …
  - Multidimensional analysis and OLAP
- **Confluence of multi-disciplines: CVPR, GIS, stat., HPC**
  - Frequent pattern and correlation analysis
  - Classification, clustering and outlier analysis

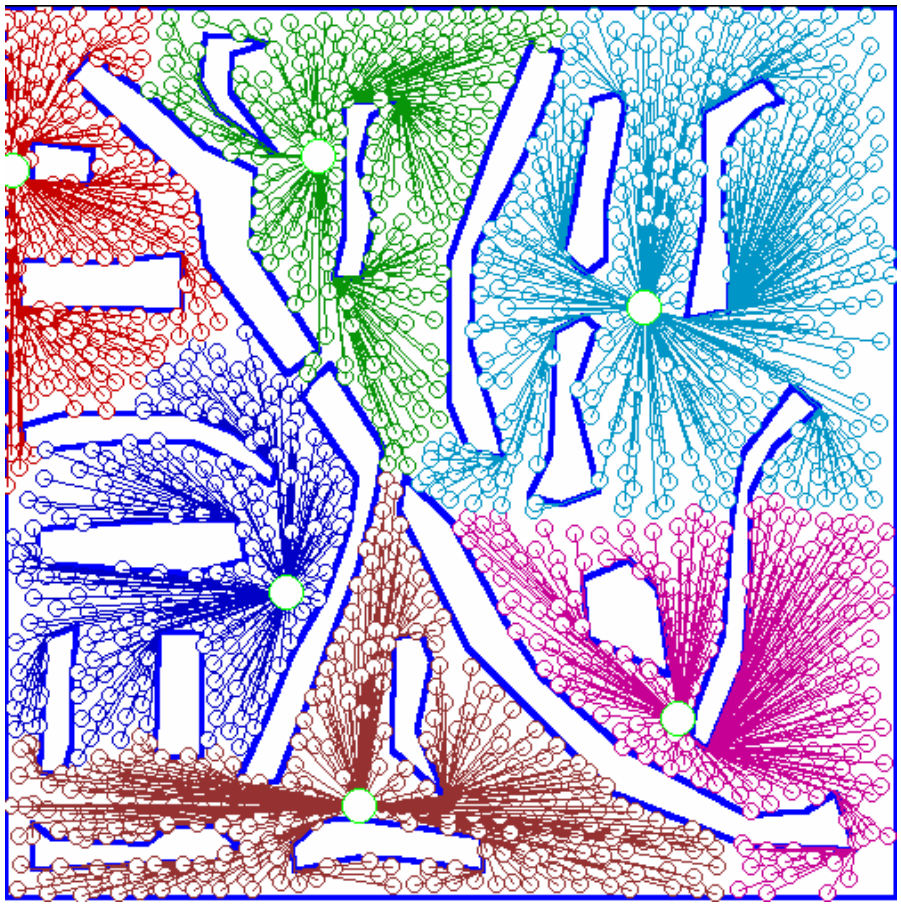
Mining Multimedia Databases in MultiMediaMiner



Clustering with Obstacle Objects



Not Taking obstacles into account



Taking obstacles into account

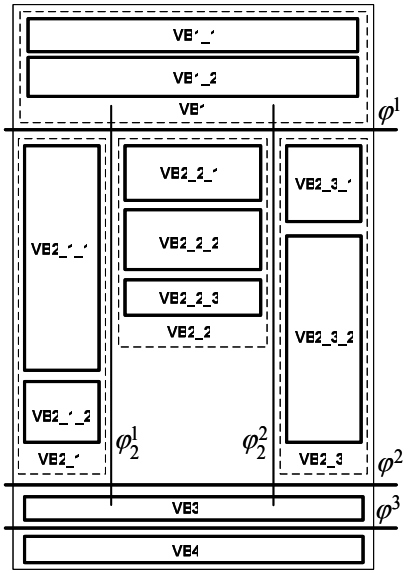
Research Challenges

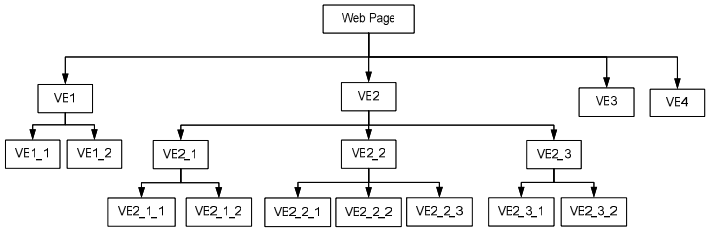
- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Text and Web Mining

- **Web: The ultimate information access and processing platform**
  - The common place for S & E to pub., share and exchange data and ideas, e.g., Bio: GenBank, ProtienBank, GO, Pubmed, …
- **Web modeling:** an evolving, collaborative, social network
- **Web, text, email and blog mining**
  - Mining digital libraries and research literature databases
  - Document clustering, classification: Exploiting text, links and usages for multi-level, many-class clustering and classification
  - Mining unstructured, textual information: emails, blogs
- **Semantic Web and information repository construction**
  - Information extraction, exploiting markup structure to extract structured data from pages meant for human consumption
- **Web usage mining and adaptive Web sites**

Mining Web Page Layout Structure for Advanced Search





- VIPS: VIsion based Page Segmentation
- Block-based Web Search
- Block-level Link analysis
- Web image search and organization

Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Data Cube-Oriented Multidimensional Online Analytical Mining

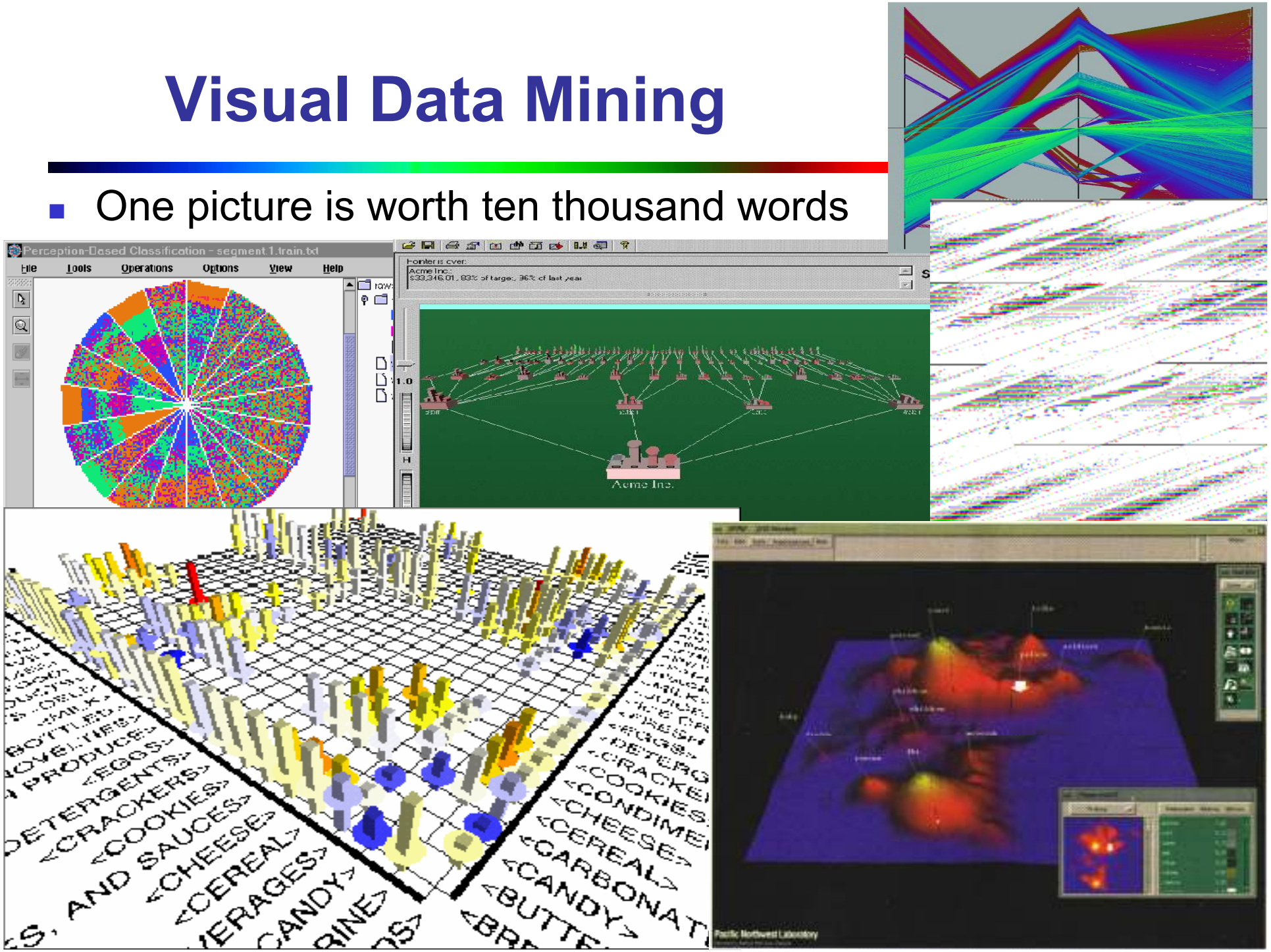
- Science and eng. data sets are often high-D in mature
- Viewing and mining data in multidimensional space
- Data cube and OLAP technology
- Beyond conventional data cubes
  - Regression cubes, prediction cubes
  - Integration cube and ranking query processing: *The Ranking Cube Approach*" [VLDB'06, SIGMOD'07]
  - High-dimensional OLAP: *A Minimal Cubing Approach* [VLDB'04]
- OLAP mining may substantially enhance the power and flexibility of data analysis
  - Exploratory-based science and eng. data mining

Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Visual Data Mining

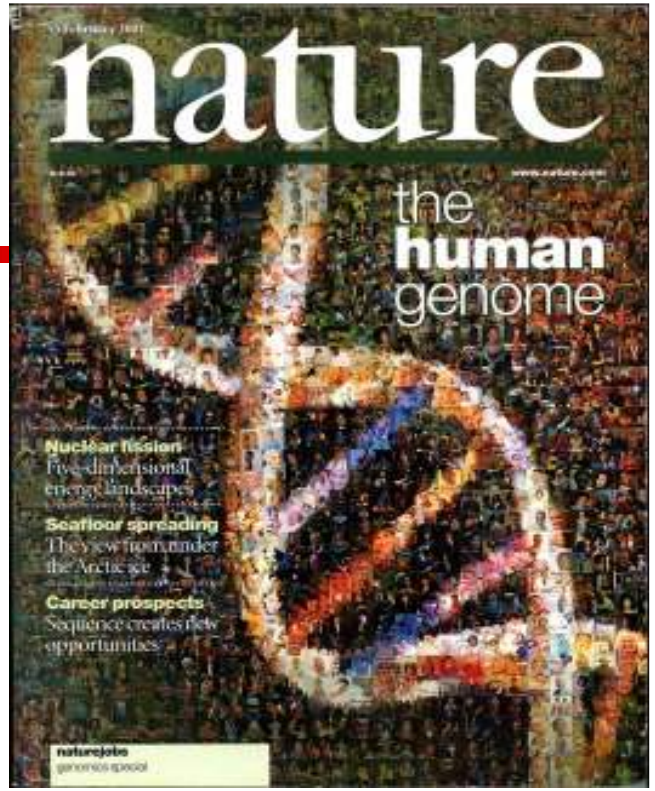
- One picture is worth ten thousand words



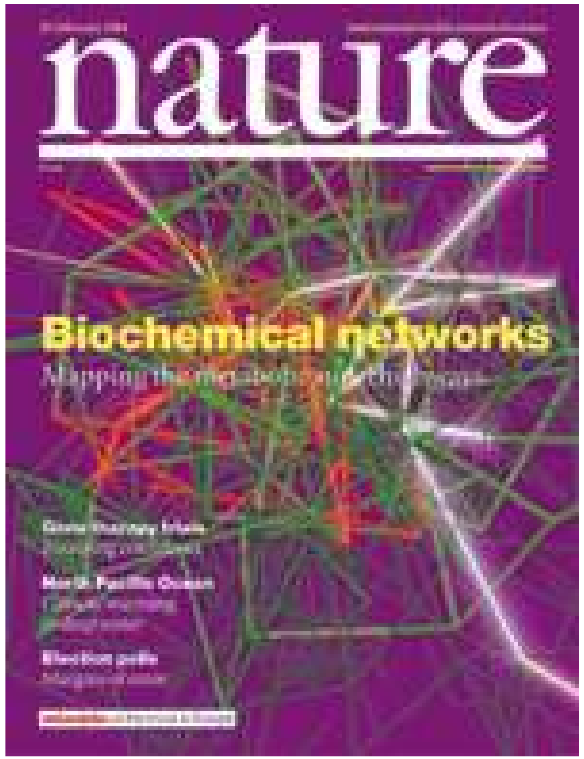
Research Challenges

- Information network analysis
- Discovery, understanding, and usage of patterns
- Stream data mining
- Mining moving object data, RFID data, and data from sensor networks
- Spatiotemporal and multimedia data mining
- Mining text, Web, and other unstructured data
- Data cube-oriented multidimensional online analytical mining
- Visual data mining
- Domain-specific data mining: Work in each scientific and engineering domain

Biological Data Mining



- **Bioinformatics: A driving force in data mining**
  - Huge, complex and valuable data sources for data mining
  - Gene or protein chips, molecular structures, bio-medical data, biomedical literature databases, mass spectrometry, spatial/image data, …

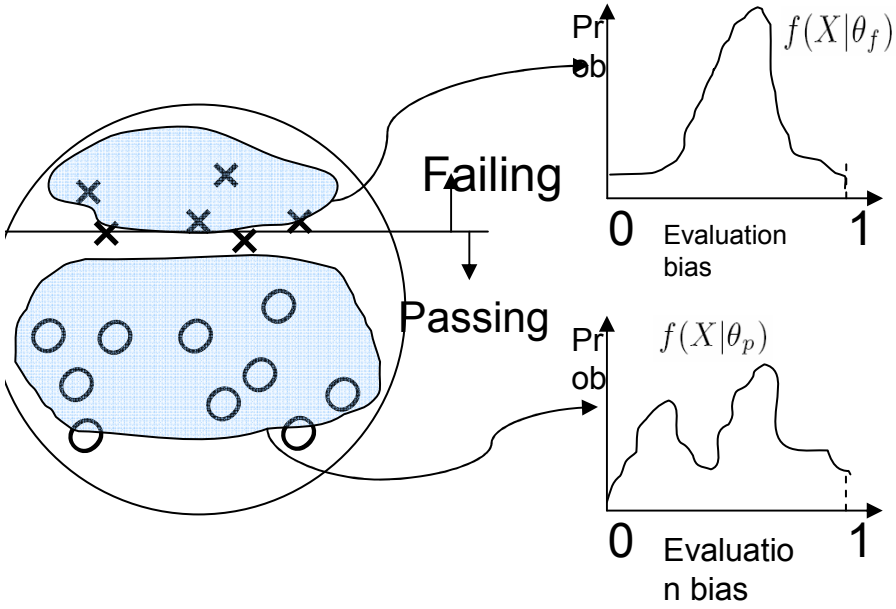


- **Data mining for genomics and proteomics**
  - Mining motif patterns, phylogenetic and functional prediction, biological networks, system biology, bioliterature
  - Data mining is at very primitive stage in bioinformatics
  - A very rewarding frontier for data mining

System/Software Engineering Data Mining

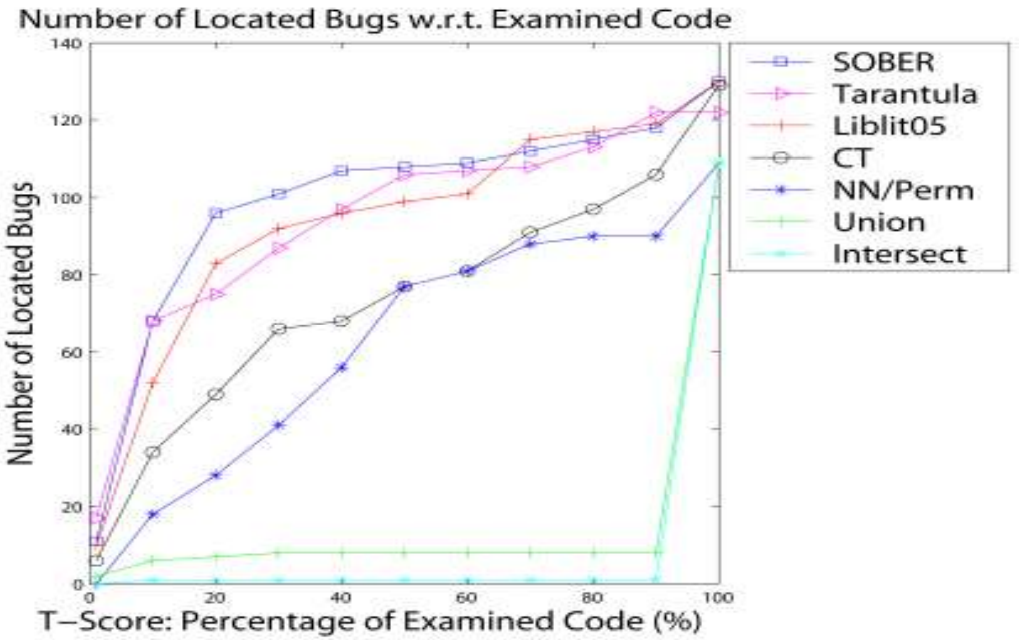
- **Computer and software system generate a huge amount of data**
- **Data mining will improve their performance, reliability & security**
- **Autonomic computing: Automate the construction, maintenance and evolution of sophisticated systems (esp. software systems)**
- **Frontiers in computer/software system data mining**
  - Software bug isolation and analysis
  - Computer network intrusion detection
  - System performance improvement, diagnosis, and maintenance
  - Protection of intellectual property (e.g., plagiarism detection)

SOBER: Bug Localization based on Classification of Statistical Distribution of Statement Execution



- Based on statistical differences between passing and failing runs, identify which lines or portions of code are most bug relevant

- Performance study [FSE'05] shows it can detect buggy lines by examining a smaller portions of code



Other Research Themes

- **Privacy-preserving data mining**
  - Encryption of data while preserving mining statistics
  - Data perturbation: adding noise and randomization
  - Data transformation: Projection in different angles while preserving mining results
  - Privacy-preserving mining in distributed environments
- **Invisible data mining**
  - Embedded functions: Web search engine (link analysis, authoritative pages, user profiles), Google News, adaptive web sites, query optimization via data mining
- **Towards integrated information systems**
  - Integration of database, data warehouse, Web, and data mining: Infrastructure for integrated data/information systems
  - Languages, optimization, automated system tuning & adaptation

Conclusions

- Science and engineering are fertile lands for data mining
- Gigantic amounts of S & E data are constantly being generated and collected
- Data mining: An essential scientific discovery process
- Lots of research themes will be flourishing
  - Information network analysis
  - Stream data mining
  - Mining moving object data, RFID data, and data from sensor networks
  - Spatiotemporal and multimedia data mining
  - Text and Web mining, visual data mining, …
  - Working on every S&E domain!
- NSF CDI: A long-waited, exciting call!

Thanks and Questions