Introduction: Datamining & Data Warehousing

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

School Of Computer Engineering

Datamining and Data warehousing (CS 2004)

Dr. Amiya Ranjan Panda
Assistant Professor [II]
School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT),
Deemed to be University,Odisha

3 Credit

Lecture Note 01

Acknoledgement

A Special

Thanks to

J. Han and M. Kamber.

&

Tan, Steinbach, Kumar

for their slides and books, which I have

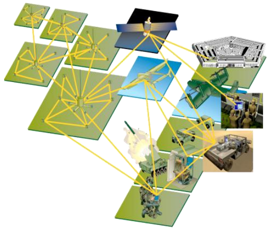
used for preparation of these slides.

Chapter Contents

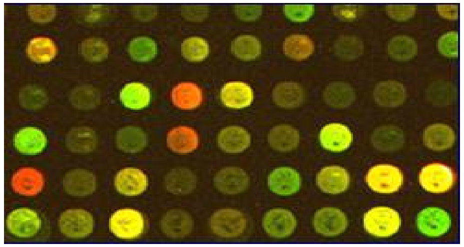
- ❑ Why Data Mining?
- ❑ What Is Data Mining?
- ❑ A Multi-Dimensional View of Data Mining
- ❑ What Kind of Data Can Be Mined?
- ❑ What Kinds of Patterns Can Be Mined?
- ❑ What Technology Are Used?
- ❑ What Kind of Applications Are Targeted?
- ❑ Major Issues in Data Mining
- ❑ A Brief History of Data Mining and Data Mining Society
- ❑ Summary

Why Data Mining?

- ❑ The Explosive Growth of Data: from terabytes to petabytes
  - ➢ Data collection and data availability
    - ✓ Automated data collection tools, database systems, Web, computerized society
  - ➢ Major sources of abundant data
    - ✓ *Business:* Web, e-commerce, transactions, stocks, …
    - ✓ *Science:* Remote sensing, bioinformatics, scientific simulation, …
    - ✓ *Society and everyone:* news, digital cameras, YouTube

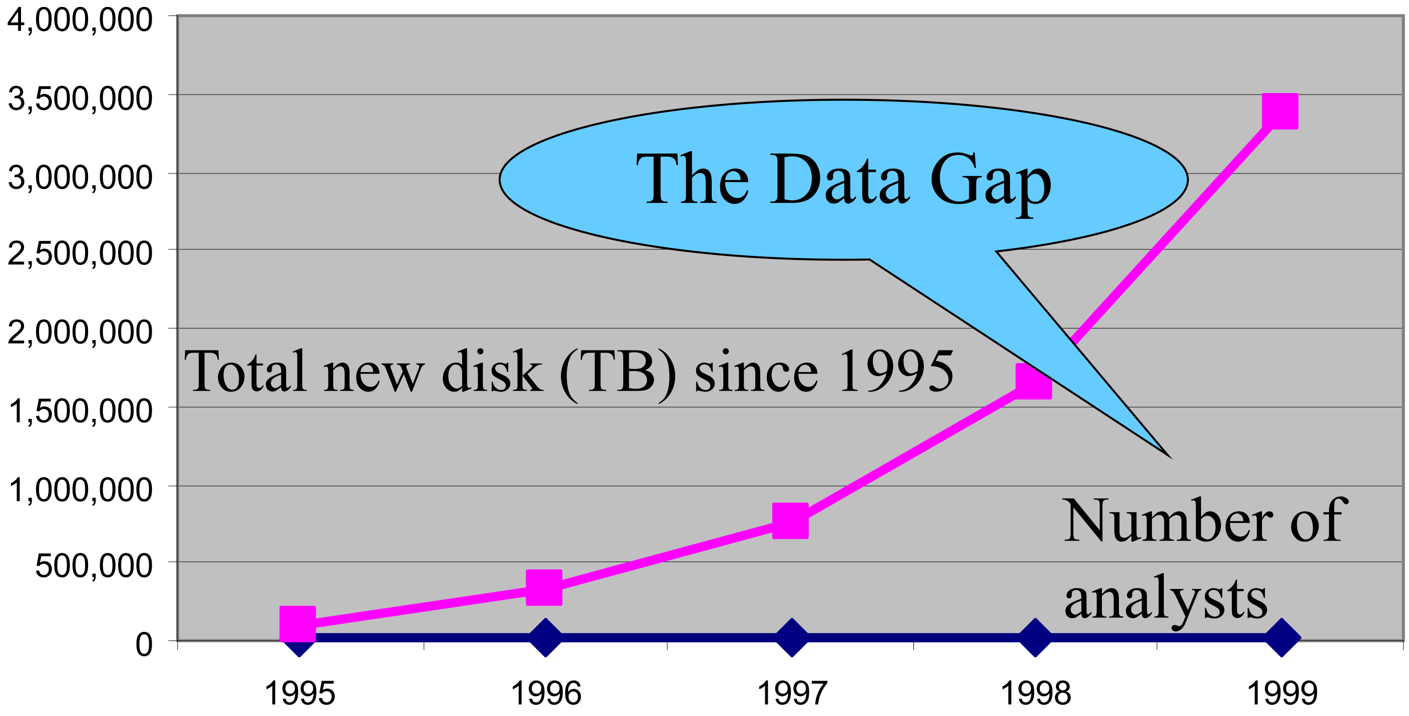


- ❑ **<u>We are drowning in data, but starving for knowledge !</u>**
- ❑ "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets



Why Data Mining?

- ❑ There is often information “hidden” in the data that is not readily evident
- ❑ Human analysts may take weeks to discover useful information
- ❑ Much of the data is never analyzed at all



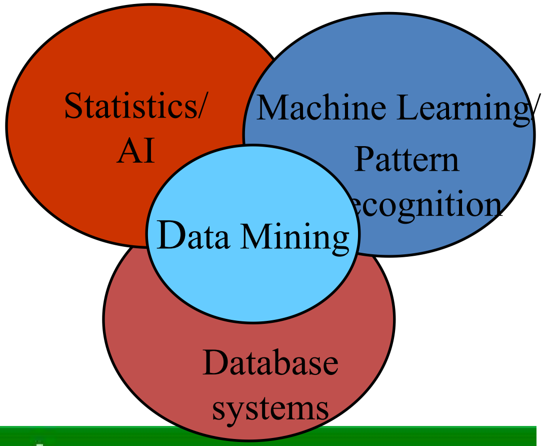
What Is Data Mining?

- ❑ Data mining (knowledge discovery from data)
  - ➢ Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - ➢ Data mining: a misnomer?
- ❑ Alternative names
  - ➢ Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- ❑ Watch out: Is everything "data mining"?
  - ➢ Simple search and query processing
  - ➢ (Deductive) expert systems

What Is Data Mining?

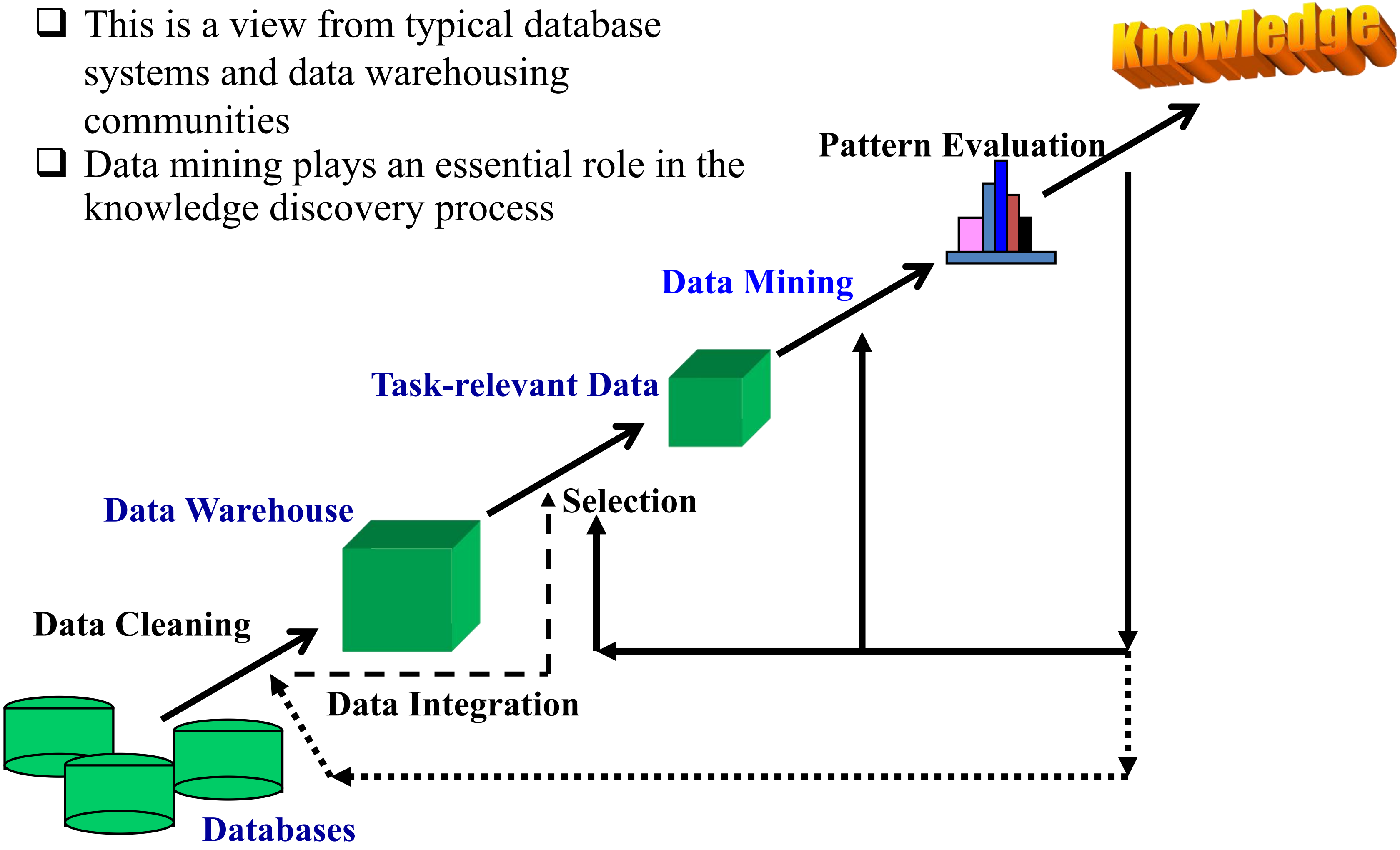
- ❑ **What is not Data Mining?**
  - ➢ Look up phone number in phone directory
  - ➢ Query a Web search engine for information about "Amazon"
- ❑ Origin of Data Mining
  - ➢ Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
  - ➢ Traditional Techniques may be unsuitable due to
    - ➢ Enormity of data
    - ➢ High dimensionality of data
    - ➢ Heterogeneous, distributed nature of data

- ❑ **What is Data Mining?**
  - ➢ Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - ➢ Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

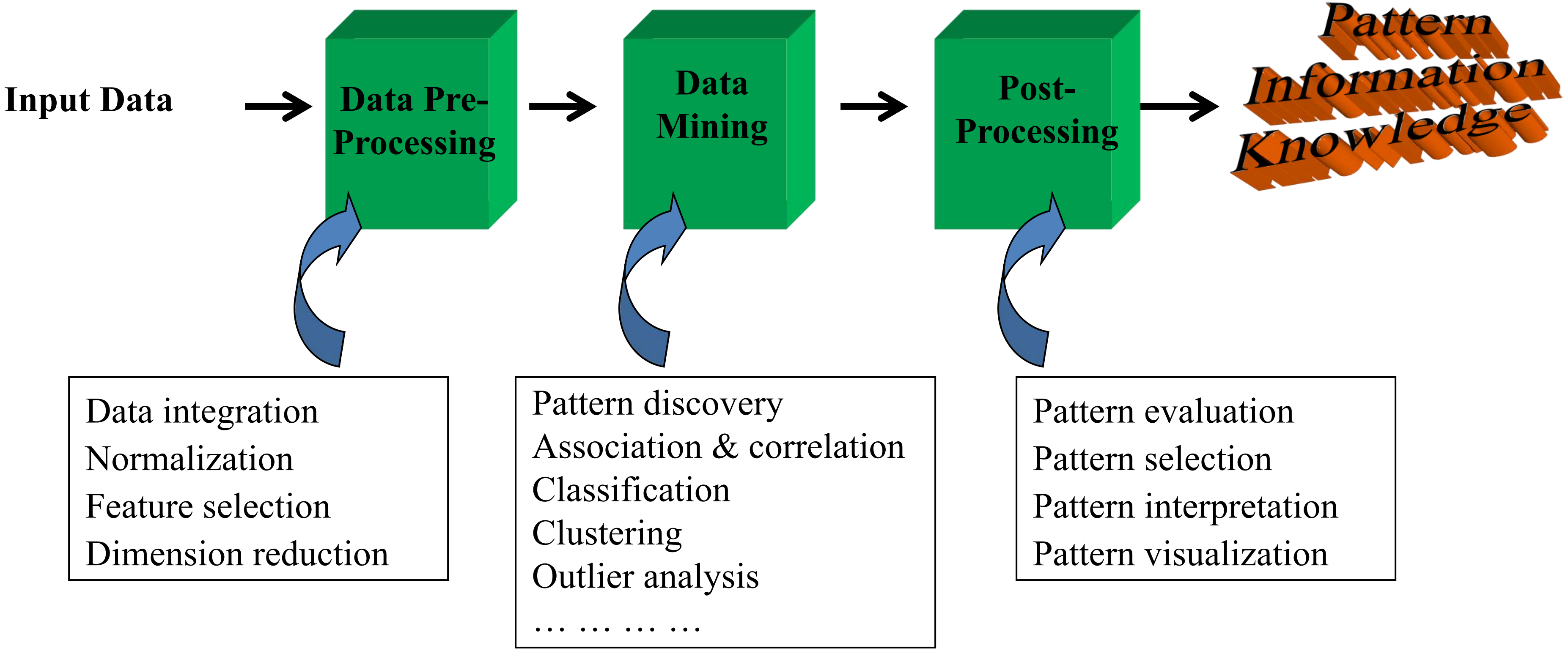


Knowledge Discovery (KDD) Process

- ❑ This is a view from typical database systems and data warehousing communities
- ❑ Data mining plays an essential role in the knowledge discovery process



KDD Process: A Typical View from ML and Statistics



Input Data

Data Pre-Processing

Data Mining

Post-Processing

Pattern Information Knowledge

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

- **Example:** Health care & medical data mining – often adopted such a view in statistics and machine learning
  - Preprocessing of the data (including feature extraction and dimension reduction)
  - Classification or/and clustering processes
  - Post-processing for presentation

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On What Kinds of Data?

- ❑ Database-oriented data sets and applications
  - ➢ Relational database, data warehouse, transactional database
- ❑ Advanced data sets and advanced applications
  - ➢ Data streams and sensor data
  - ➢ Time-series data, temporal data, sequence data (incl. bio-sequences)
  - ➢ Structure data, graphs, social networks and multi-linked data
  - ➢ Object-relational databases
  - ➢ Heterogeneous databases and legacy databases
  - ➢ Spatial data and spatiotemporal data
  - ➢ Multimedia database
  - ➢ Text databases
  - ➢ The World-Wide Web

Data Mining Tasks

- □ Prediction Methods
  - ➢ Use some variables to predict unknown or future values of other variables.
    - ➢ Classification
    - ➢ Regression
    - ➢ Deviation Detection
- □ Description Methods
  - ➢ Find human-interpretable patterns that describe the data.
    - ➢ Clustering
    - ➢ Association Rule Discovery
    - ➢ Sequential Pattern Discovery

Data Mining Function: (1) Generalization

- ❑ Information integration and data warehouse construction
  - ➢ Data cleaning, transformation, integration, and multidimensional data model
- ❑ Data cube technology
  - ➢ Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - ➢ OLAP (online analytical processing)
- ❑ Multidimensional concept description: Characterization and discrimination
  - ➢ Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

(2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - ✓ Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

TID	***Items***
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

(3) Classification

- ❑ Classification and label prediction
  - ➢ Construct models (functions) based on some training examples
  - ➢ Describe and distinguish classes or concepts for future prediction
    - ✓ E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - ➢ Predict some unknown class labels
- ❑ Typical methods
  - ➢ Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- ❑ Typical applications:
  - ➢ Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

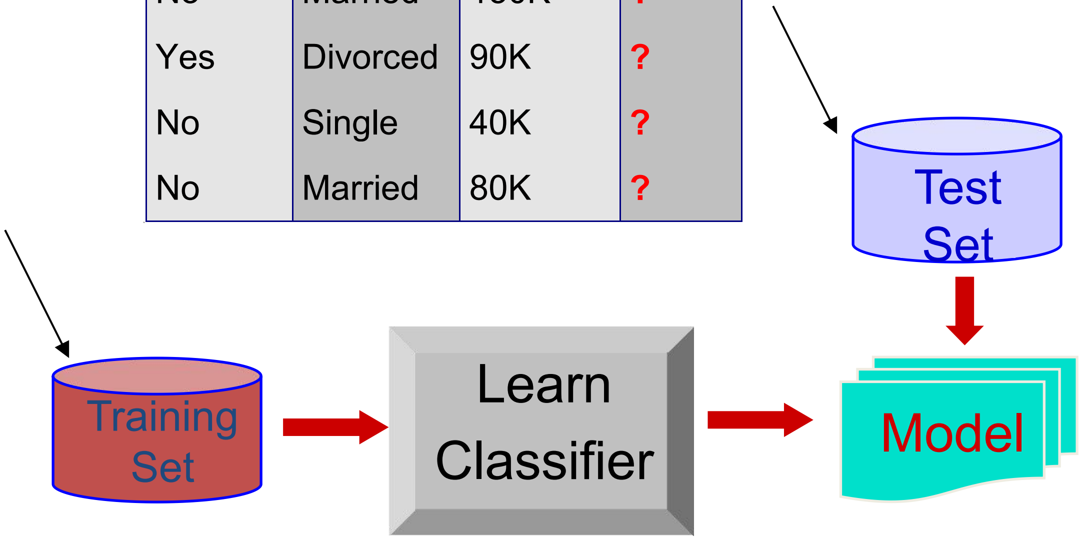
(3) Classification - Regression

- ❑ Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- ❑ Greatly studied in statistics, neural network fields.
- ❑ Examples:
  - ➢ Predicting sales amounts of new product based on advetising expenditure.
  - ➢ Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - ➢ Time series prediction of stock market indices.

(3) Classification Example

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



(4) Cluster Analysis

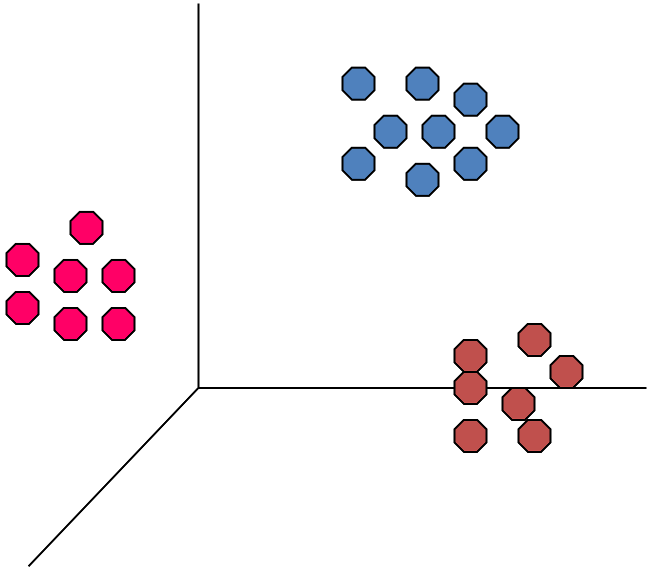
- ❑ Unsupervised learning (i.e., Class label is unknown)
- ❑ Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- ❑ Principle: Maximizing intra-class similarity & minimizing interclass similarity
- ❑ Many methods and applications

(4) Cluster Analysis Illustration

❑ Euclidean Distance Based Clustering in 3-D space.

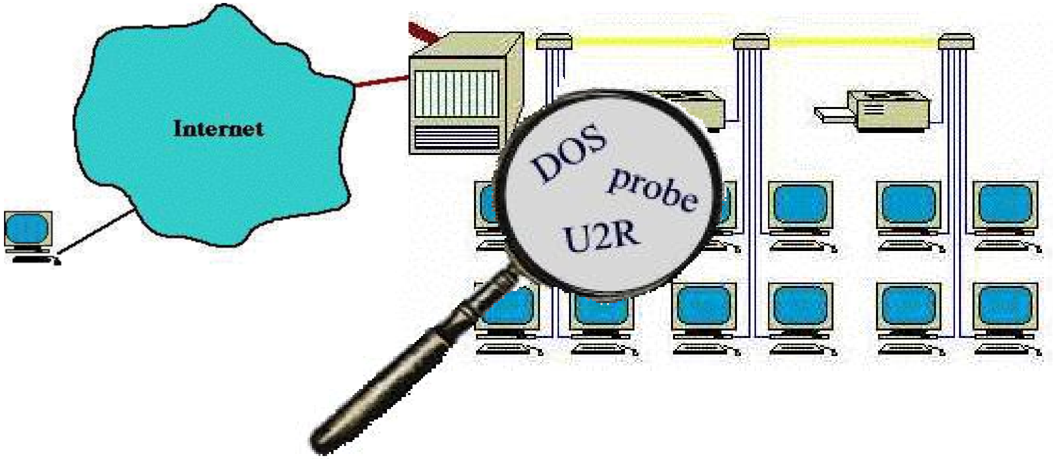
Intracluster distances are minimized

Intercluster distances are maximized



(5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in network intrusion detection, credit card fraud detection, rare events analysis



Time and Ordering:

- ❑ **Sequential Pattern, Trend and Evolution Analysis**
  - ➢ Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - ➢ Sequential pattern mining
    - ✓ e.g., first buy digital camera, then buy large SD memory cards
  - ➢ Periodicity analysis
  - ➢ Motifs and biological sequence analysis
    - ✓ Approximate and consecutive motifs
  - ➢ Similarity-based analysis
- ❑ Mining data streams
  - ➢ Ordered, time-varying, potentially infinite, data streams

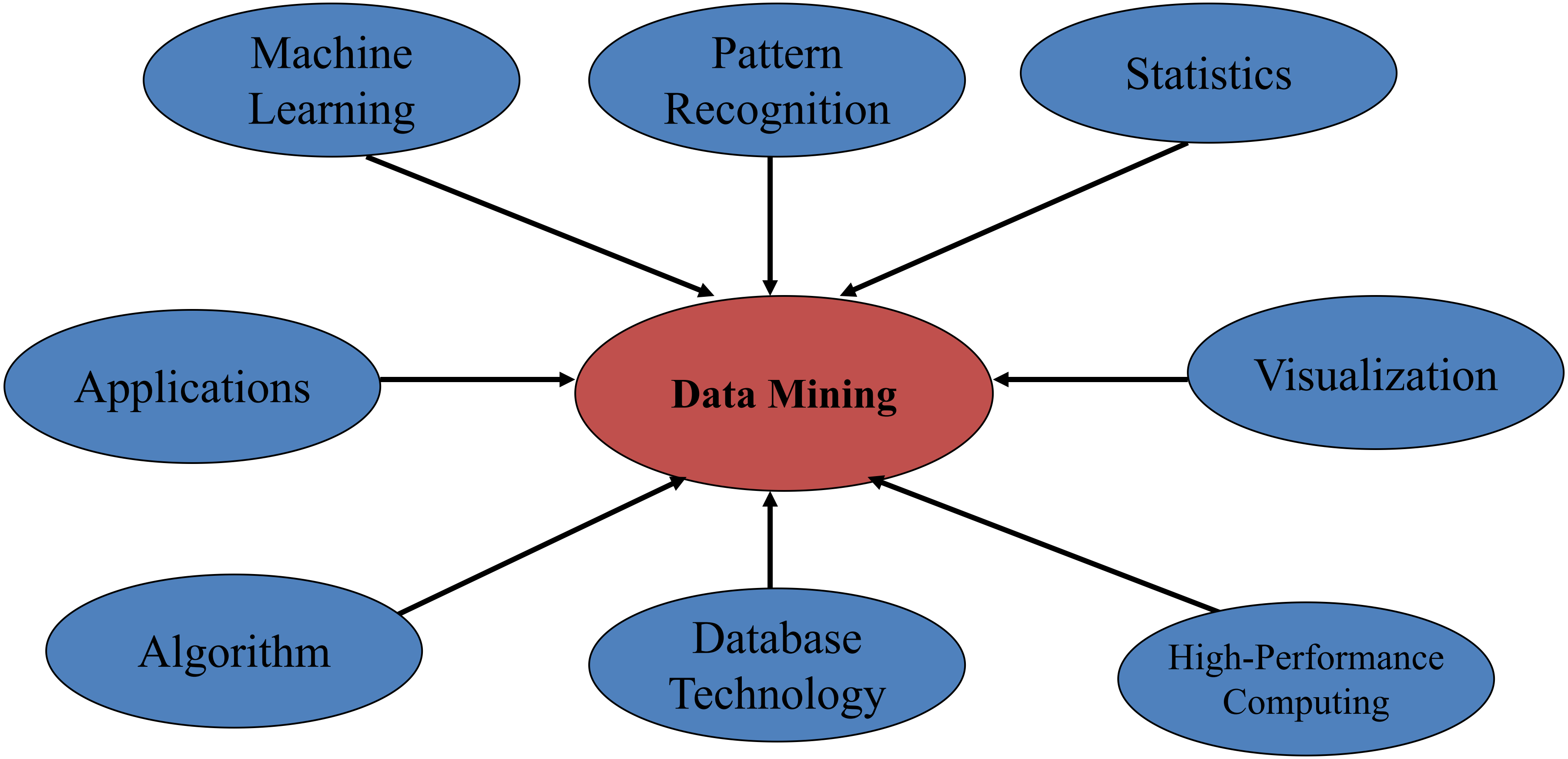
Structure and Network Analysis

- ❑ Graph mining
  - ➢ Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- ❑ Information network analysis
  - ➢ Social networks: actors (objects, nodes) and relationships (edges)
    - ✓ e.g., author networks in CS, terrorist networks
  - ➢ Multiple heterogeneous networks
    - ✓ A person could be multiple information networks: friends, family, classmates, …
  - ➢ Links carry a lot of semantic information: Link mining
- ❑ Web mining
  - ➢ Web is a big information network: from PageRank to Google
  - ➢ Analysis of Web information networks
    - ✓ Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- ❑ Are all mined knowledge interesting?
    - ➢ One can mine tremendous amount of “patterns” and knowledge
    - ➢ Some may fit only certain dimension space (time, location, …)
    - ➢ Some may not be representative, may be transient, …
- ❑ Evaluation of mined knowledge → directly mine only interesting knowledge?
    - ➢ Descriptive vs. predictive
    - ➢ Coverage
    - ➢ Typicality vs. novelty
    - ➢ Accuracy
    - ➢ Timeliness
    - ➢ …

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- ❑ Tremendous amount of data
  - ➢ Algorithms must be highly scalable to handle such as tera-bytes of data
- ❑ High-dimensionality of data
  - ➢ Micro-array may have tens of thousands of dimensions
- ❑ High complexity of data
  - ➢ Data streams and sensor data
  - ➢ Time-series data, temporal data, sequence data
  - ➢ Structure data, graphs, social networks and multi-linked data
  - ➢ Heterogeneous databases and legacy databases
  - ➢ Spatial, spatiotemporal, multimedia, text and Web data
  - ➢ Software programs, scientific simulations
- ❑ New and sophisticated applications

Applications of Data Mining

- ❑ Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- ❑ Collaborative analysis & recommender systems
- ❑ Basket data analysis to targeted marketing
- ❑ Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- ❑ Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- ❑ From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Major Issues in Data Mining (1)

- ❑ Mining Methodology
  - ➢ Mining various and new kinds of knowledge
  - ➢ Mining knowledge in multi-dimensional space
  - ➢ Data mining: An interdisciplinary effort
  - ➢ Boosting the power of discovery in a networked environment
  - ➢ Handling noise, uncertainty, and incompleteness of data
  - ➢ Pattern evaluation and pattern- or constraint-guided mining
- ❑ User Interaction
  - ➢ Interactive mining
  - ➢ Incorporation of background knowledge
  - ➢ Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- ❑ Efficiency and Scalability
  - ➢ Efficiency and scalability of data mining algorithms
  - ➢ Parallel, distributed, stream, and incremental mining methods
- ❑ Diversity of data types
  - ➢ Handling complex types of data
  - ➢ Mining dynamic, networked, and global data repositories
- ❑ Data mining and society
  - ➢ Social impacts of data mining
  - ➢ Privacy-preserving data mining
  - ➢ Invisible data mining

A Brief History of Data Mining Society

- ❑ 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - ➢ Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- ❑ 1991-1994 Workshops on Knowledge Discovery in Databases
  - ➢ Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- ❑ 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - ➢ Journal of Data Mining and Knowledge Discovery (1997)
- ❑ ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- ❑ More conferences on data mining
  - ➢ PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ❑ ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- ❑ KDD Conferences
  - ➢ ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - ➢ SIAM Data Mining Conf. (SDM)
  - ➢ (IEEE) Int. Conf. on Data Mining (ICDM)
  - ➢ European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - ➢ Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - ➢ Int. Conf. on Web Search and Data Mining (WSDM)

- ❑ Other related conferences
  - ➢ DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - ➢ Web and IR conferences: WWW, SIGIR, WSDM
  - ➢ ML conferences: ICML, NIPS
  - ➢ PR conferences: CVPR,
- ❑ Journals
  - ➢ Data Mining and Knowledge Discovery (DAMI or DMKD)
  - ➢ IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - ➢ KDD Explorations
  - ➢ ACM Trans. on KDD

Summary

- ❑ Data mining: Discovering interesting patterns and knowledge from massive amount of data
- ❑ A natural evolution of database technology, in great demand, with wide applications
- ❑ A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- ❑ Mining can be performed in a variety of data
- ❑ Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- ❑ Data mining technologies and applications
- ❑ Major issues in data mining

Recommended Text and Reference Books

- ❑ **Text Book:**
  - ➢ J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- ❑ **Reference Books:**
  - ➢ H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2006.
  - ➢ I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2000.
  - ➢ D. Hand, H. Mannila and P. Smyth. Principles of Data Mining.Prentice-Hall. 2001.

THANK YOU!