

Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

Chapter 1. Introduction

- 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
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets


Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
 - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
 - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
 - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
 - The flood of data from new scientific instruments and simulations
 - The ability to economically store and manage petabytes of data online
 - The Internet and computing Grid that makes all these archives universally accessible
 - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data mining** is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

- 1960s:
 - Data collection, database creation, IMS and network DBMS
- 1970s:
 - Relational data model, relational DBMS implementation
- 1980s:
 - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
 - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
 - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
 - Stream data management and mining
 - Data mining and its applications
 - Web technology (XML, data integration) and global information systems

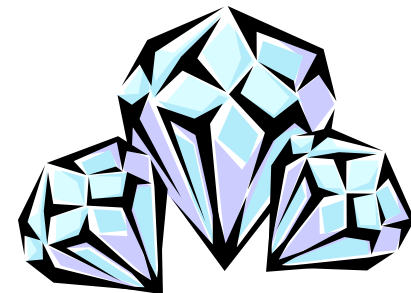
Chapter 1. Introduction

- 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

What Is Data Mining?

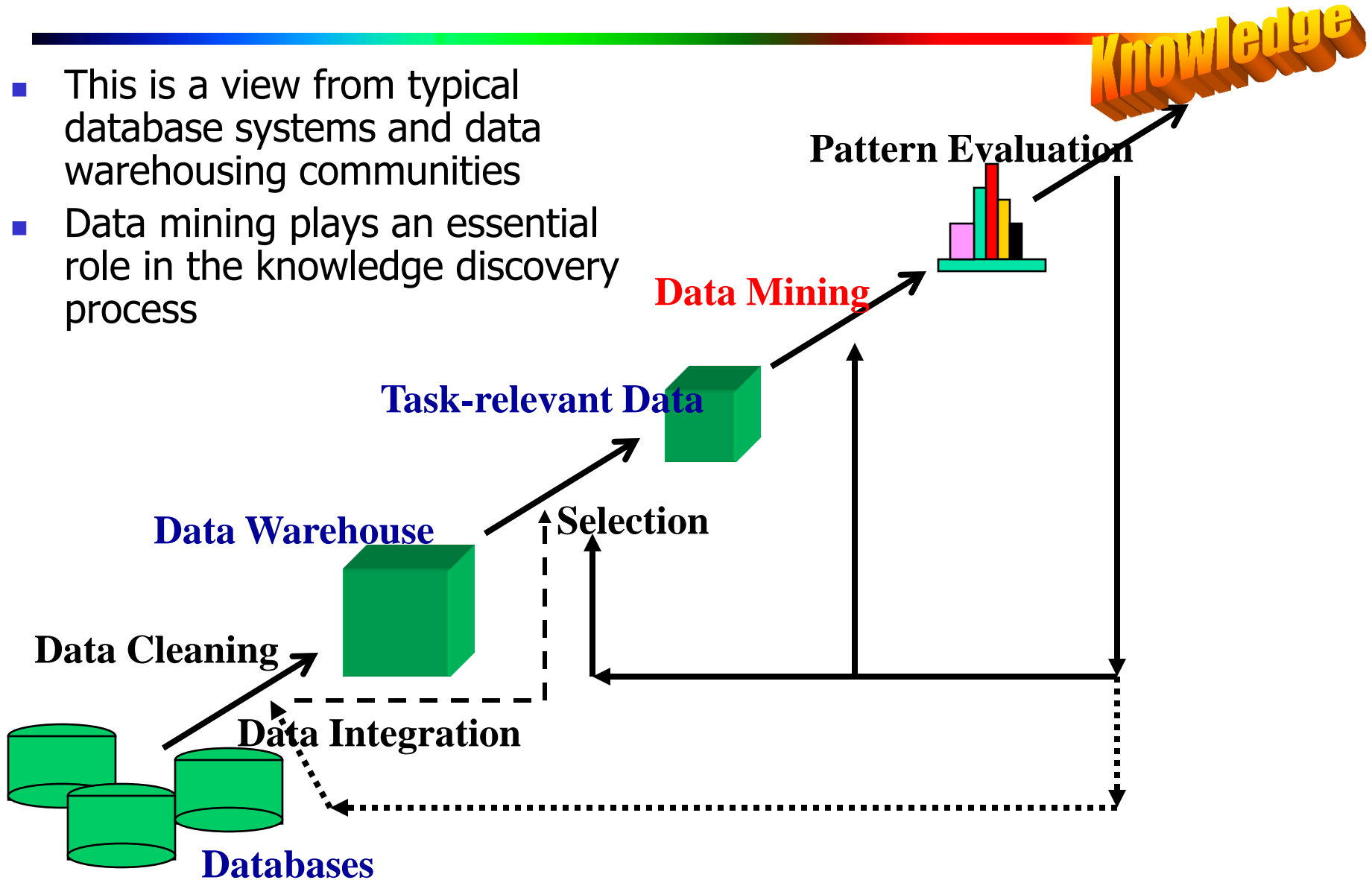


- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, 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



Knowledge Discovery (KDD) Process

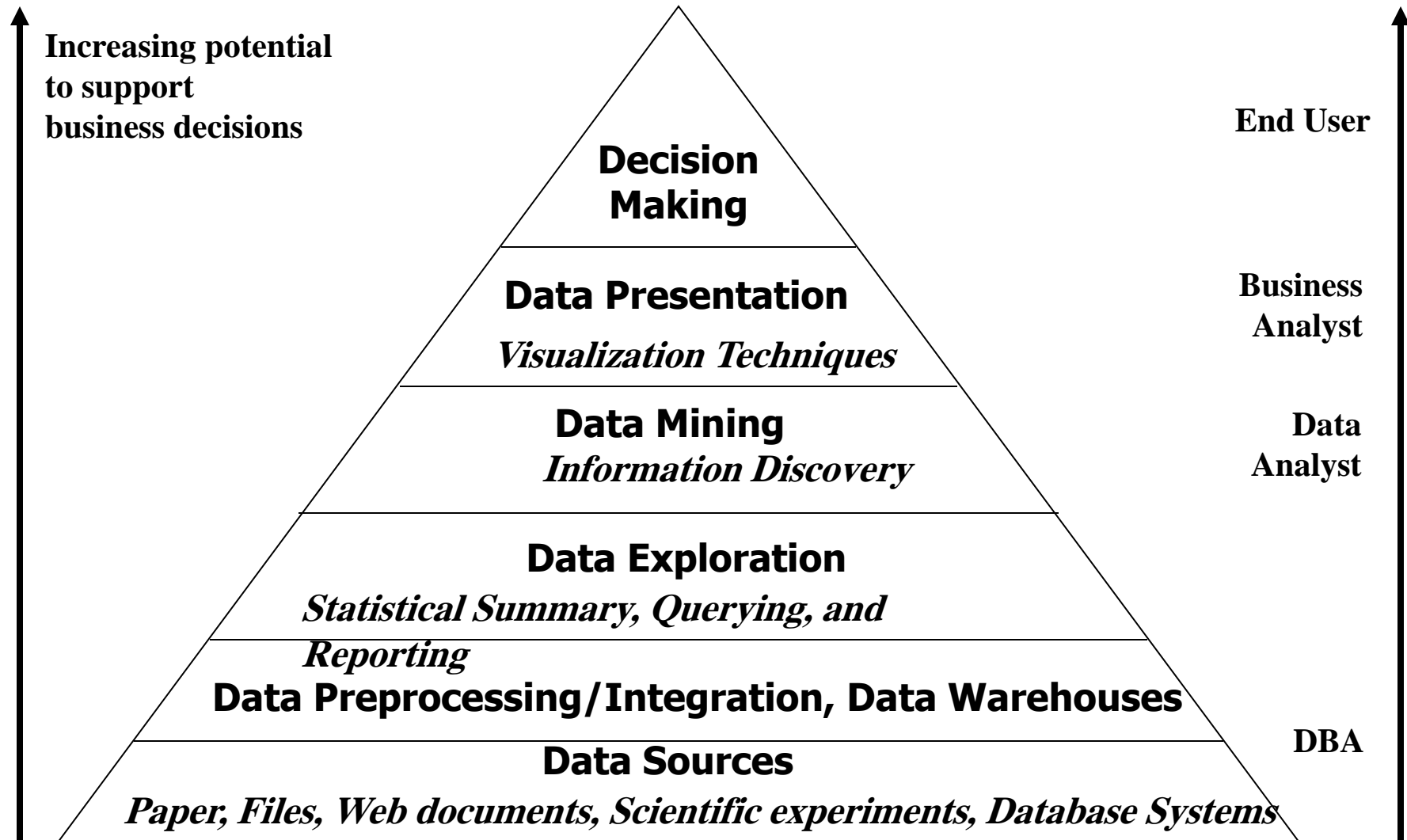
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Example: A Web Mining Framework

- Web mining usually involves
 - Data cleaning
 - Data integration from multiple sources
 - Warehousing the data
 - Data cube construction
 - Data selection for data mining
 - Data mining
 - Presentation of the mining results
 - Patterns and knowledge to be used or stored into knowledge-base

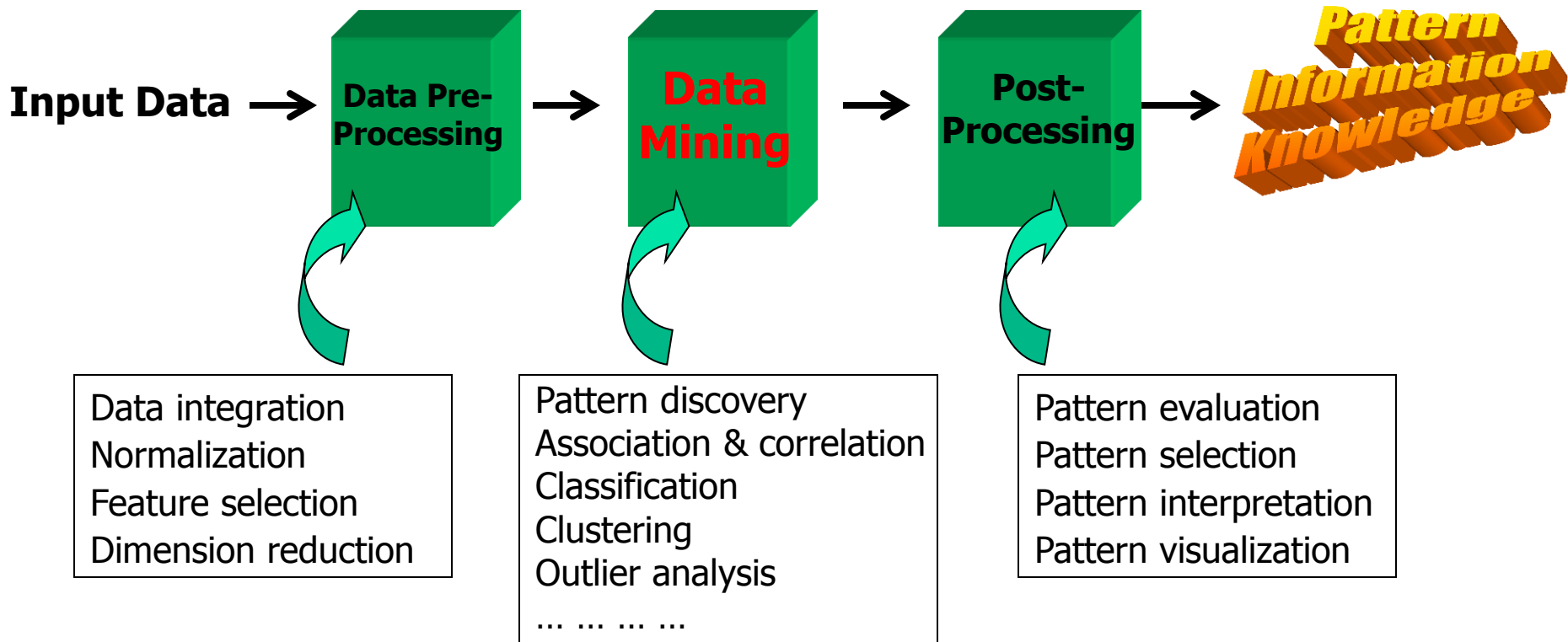
Data Mining in Business Intelligence



Example: Mining vs. Data Exploration

- Business intelligence view
 - Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

KDD Process: A Typical View from ML and Statistics




- This is a view from typical machine learning and statistics communities

Example: Medical Data Mining

- 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

Chapter 1. Introduction

- 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

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.


Chapter 1. Introduction

- 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

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

Chapter 1. Introduction

- 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

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

Data Mining Function: (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 \rightarrow 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?

Data Mining Function: (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, ...

Data Mining Function: (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

Data Mining Function: (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 fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, 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, ...

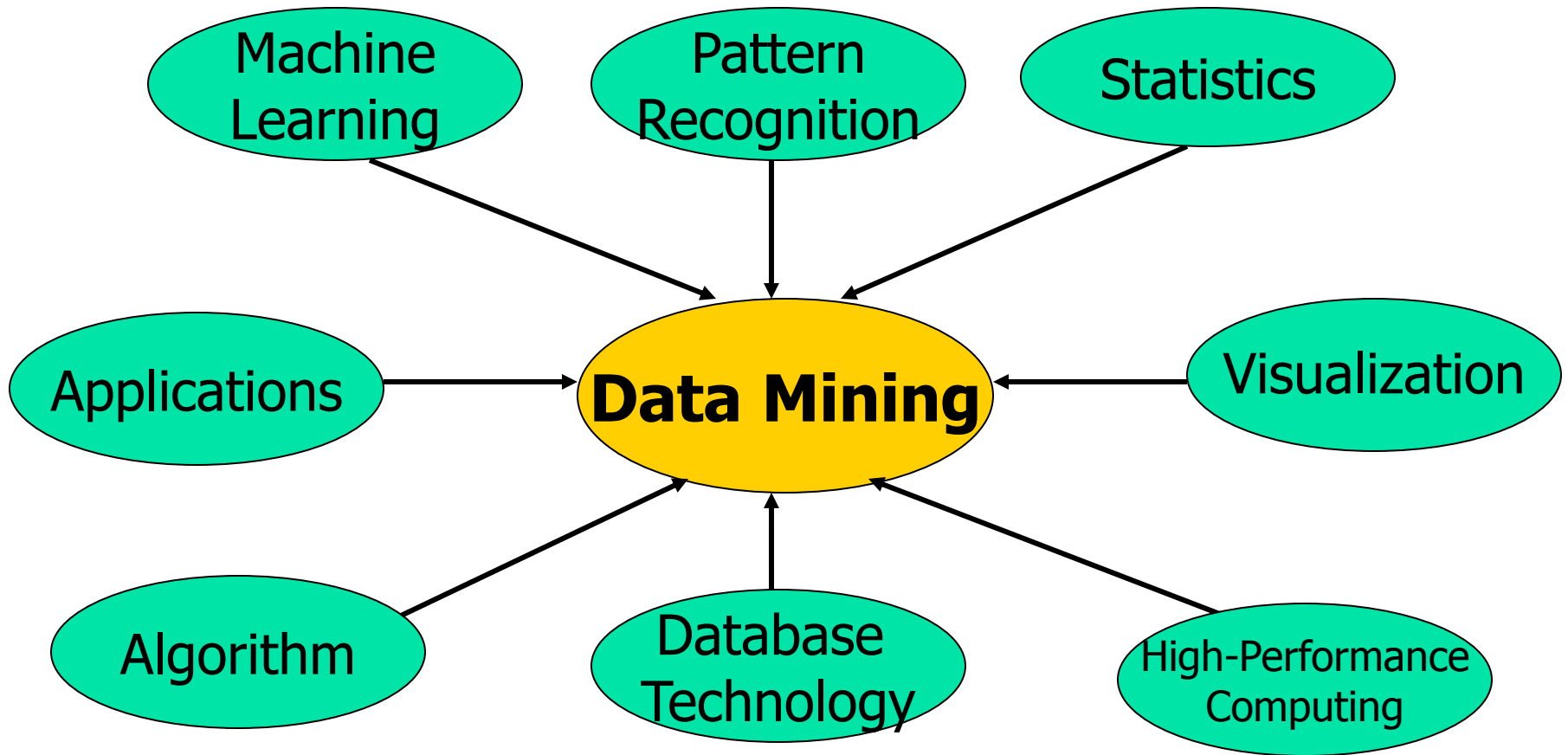
Evaluation of Knowledge

- 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
 - ...

Chapter 1. Introduction

- 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


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


Chapter 1. Introduction

- 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

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

Chapter 1. Introduction

- 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

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

Chapter 1. Introduction

- 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



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

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
 - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
 - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
 - Conferences: SIGIR, WWW, CIKM, etc.
 - Journals: WWW: Internet and Web Information Systems,
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

Chapter 1. Introduction

- 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 

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 Reference Books

- **S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data. Morgan Kaufmann, 2002**
- **R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000**
- **T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003**
- **U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996**
- **U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001**
- **J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011**
- **D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001**
- **T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009**
- **B. Liu, Web Data Mining, Springer 2006.**
- **T. M. Mitchell, Machine Learning, McGraw Hill, 1997**
- **G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991**
- **P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005**
- **S. M. Weiss and N. Indurkha, Predictive Data Mining, Morgan Kaufmann, 1998**
- **I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005**



Reference

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

Why Data Mining

⌘ Credit ratings/targeted marketing:

- ☒ Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
- ☒ Identify likely responders to sales promotions

⌘ Fraud detection

- ☒ Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?

⌘ Customer relationship management:

- ☒ Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? :

Data Mining helps extract such information

Data mining



- ⌘ Process of semi-automatically analyzing large databases to find patterns that are:
 - ☑ valid: hold on new data with some certainty
 - ☑ novel: non-obvious to the system
 - ☑ useful: should be possible to act on the item
 - ☑ understandable: humans should be able to interpret the pattern
- ⌘ Also known as Knowledge Discovery in Databases (KDD)

Applications



- ⌘ Banking: loan/credit card approval
 - ☑ predict good customers based on old customers
- ⌘ Customer relationship management:
 - ☑ identify those who are likely to leave for a competitor.
- ⌘ Targeted marketing:
 - ☑ identify likely responders to promotions
- ⌘ Fraud detection: telecommunications, financial transactions
 - ☑ from an online stream of event identify fraudulent events
- ⌘ Manufacturing and production:
 - ☑ automatically adjust knobs when process parameter changes

Applications (continued)



⌘ Medicine: disease outcome, effectiveness of treatments

☑ analyze patient disease history: find relationship between diseases

⌘ Molecular/Pharmaceutical: identify new drugs

⌘ Scientific data analysis:

☑ identify new galaxies by searching for sub clusters

⌘ Web site/store design and promotion:

☑ find affinity of visitor to pages and modify layout

The KDD process

⌘ Problem formulation

⌘ Data collection

☒ subset data: sampling might hurt if highly skewed data

☒ feature selection: principal component analysis, heuristic search

⌘ Pre-processing: cleaning

☒ name/address cleaning, different meanings (annual, yearly), duplicate removal, supplying missing values

⌘ Transformation:

☒ map complex objects e.g. time series data to features e.g. frequency

⌘ Choosing mining task and mining method:

⌘ Result evaluation and Visualization:

Knowledge discovery is an iterative process

Relationship with other fields



- ⌘ Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
 - ☑ scalability of number of features and instances
 - ☑ stress on algorithms and architectures whereas foundations of methods and formulations provided by statistics and machine learning.
 - ☑ automation for handling large, heterogeneous data

Some basic operations



⌘ Predictive:

- ☑ Regression
- ☑ Classification
- ☑ Collaborative Filtering

⌘ Descriptive:

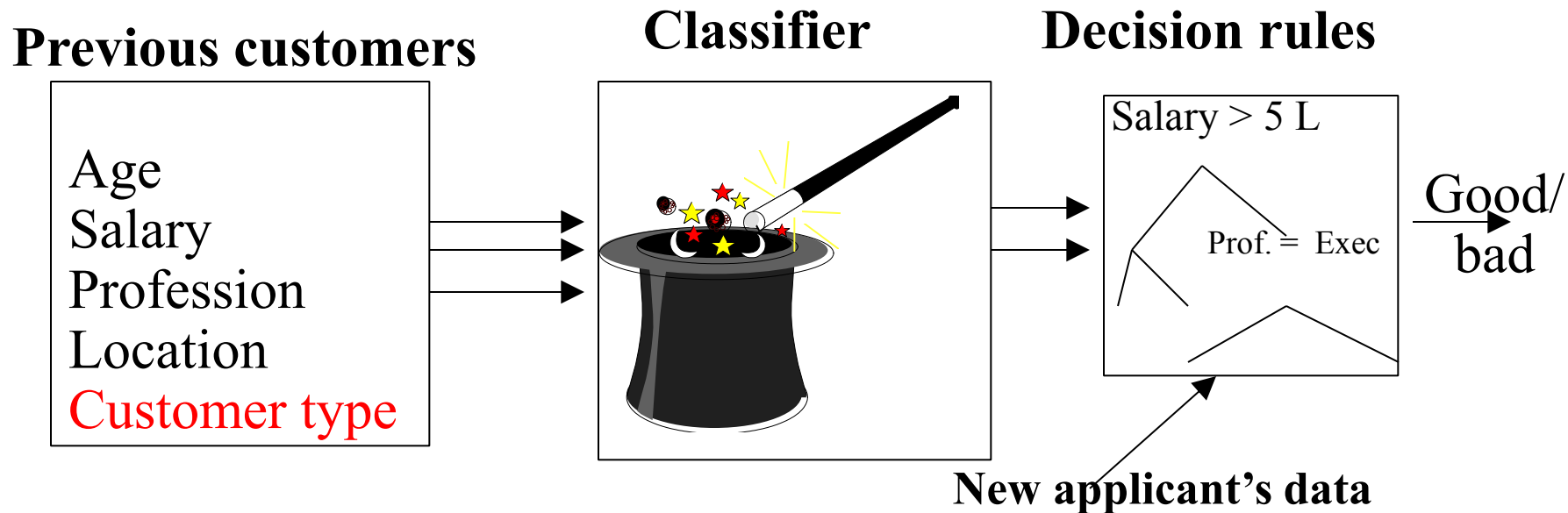
- ☑ Clustering / similarity matching
- ☑ Association rules and variants
- ☑ Deviation detection



Classification (Supervised learning)

Classification

⌘ Given old data about customers and payments, predict new applicant's loan eligibility.



Classification methods

- ⌘ **Goal:** Predict class $C_i = f(x_1, x_2, \dots, X_n)$
- ⌘ Regression: (linear or any other polynomial)
 - ⊞ $a*x_1 + b*x_2 + c = C_i.$
- ⌘ Nearest neighbour
- ⌘ Decision tree classifier: divide decision space into piecewise constant regions.
- ⌘ Probabilistic/generative models
- ⌘ Neural networks: partition by non-linear boundaries

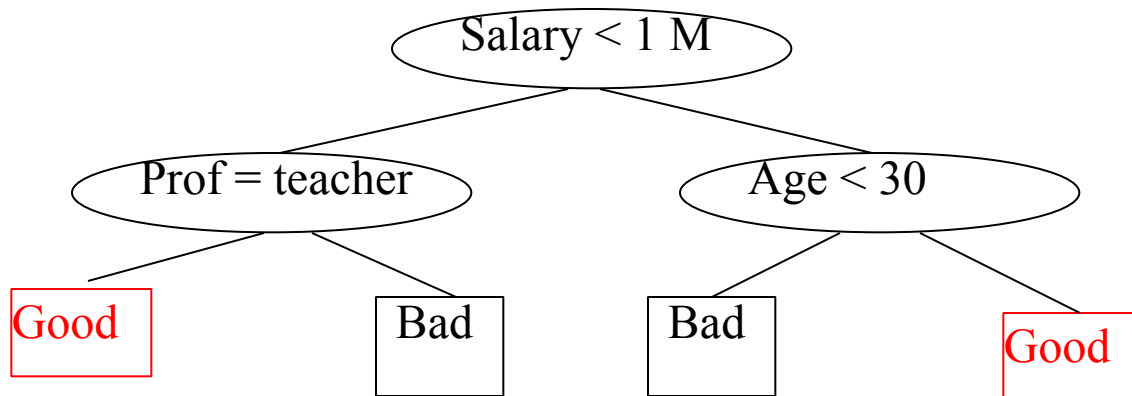
Nearest neighbor



- ⌘ Define proximity between instances, find neighbors of new instance and assign majority class
- ⌘ Case based reasoning: when attributes are more complicated than real-valued.
 - Pros
 - + Fast training
 - Cons
 - Slow during application.
 - No feature selection.
 - Notion of proximity vague

Decision trees

- Tree where internal nodes are simple decision rules on one or more attributes and leaf nodes are predicted class labels.



Decision tree classifiers

- ⌘ Widely used learning method
- ⌘ Easy to interpret: can be re-represented as if-then-else rules
- ⌘ Approximates function by piece wise constant regions
- ⌘ Does not require any prior knowledge of data distribution, works well on noisy data.
- ⌘ Has been applied to:
 - ⌘ classify medical patients based on the disease,
 - ⌘ equipment malfunction by cause,
 - ⌘ loan applicant by likelihood of payment.

Pros and Cons of decision trees



· Pros

- + Reasonable training time
- + Fast application
- + Easy to interpret
- + Easy to implement
- + Can handle large number of features

More information:

<http://www.stat.wisc.edu/~limt/treeprogs.html>

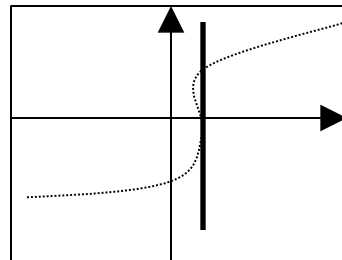
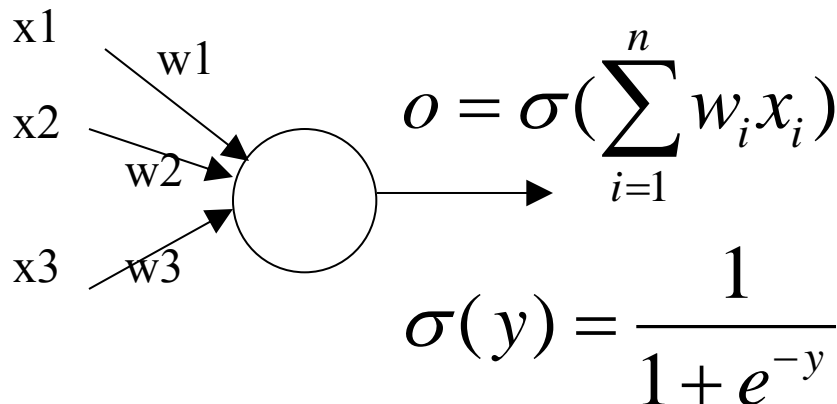
· Cons

- Cannot handle complicated relationship between features
- simple decision boundaries
- problems with lots of missing data

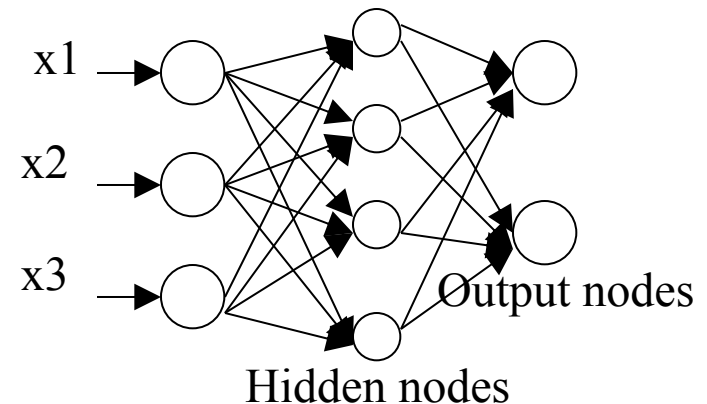
Neural network

⌘ Set of nodes connected by directed weighted edges

Basic NN unit



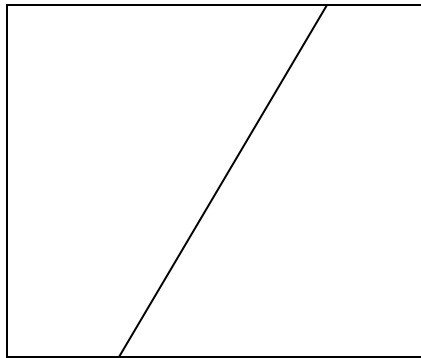
A more typical NN



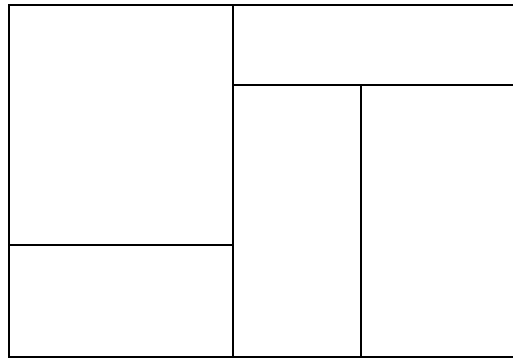
Neural networks

⌘ Useful for learning complex data like handwriting, speech and image recognition

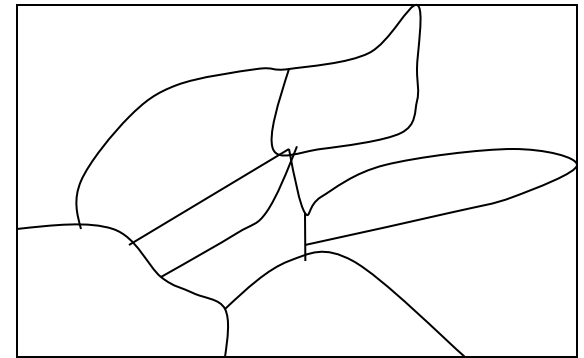
Decision boundaries:



Linear regression



Classification tree



Neural network

Pros and Cons of Neural Network



· Pros

- + Can learn more complicated class boundaries
- + Fast application
- + Can handle large number of features

· Cons

- Slow training time
- Hard to interpret
- Hard to implement: trial and error for choosing number of nodes

Conclusion: Use neural nets only if decision-trees/NN fail.

Bayesian learning

- ⌘ Assume a probability model on generation of data.
- ⌘ predicted class : $c = \max_{c_j} p(c_j | d) = \max_{c_j} \frac{p(d | c_j) p(c_j)}{p(d)}$
- ⌘ Apply bayes theorem to find most likely class as:

$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^n p(a_i | c_j)$$

- ⌘ Naïve bayes: Assume attributes conditionally independent given class value
- ⌘ Easy to learn probabilities by counting,
- ⌘ Useful in some domains e.g. text



Clustering or Unsupervised Learning

Clustering



- ⌘ Unsupervised learning when old data with class labels not available e.g. when introducing a new product.
- ⌘ Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
- ⌘ Key requirement: Need a good measure of similarity between instances.
- ⌘ Identify micro-markets and develop policies for each

Applications



- ⌘ Customer segmentation e.g. for targeted marketing
 - ☑ Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
 - ☑ Identify micro-markets and develop policies for each
- ⌘ Collaborative filtering:
 - ☑ group based on common items purchased
- ⌘ Text clustering
- ⌘ Compression

Distance functions



- ⌘ Numeric data: euclidean, manhattan distances
- ⌘ Categorical data: 0/1 to indicate presence/absence followed by
 - ☑ Hamming distance (# dissimilarity)
 - ☑ Jaccard coefficients: $\frac{\# \text{similarity in 1s}}{\# \text{ of 1s}}$
 - ☑ data dependent measures: similarity of A and B depends on co-occurrence with C.
- ⌘ Combined numeric and categorical data:
 - ☑ weighted normalized distance:

Clustering methods



⌘ Hierarchical clustering

- ☒ agglomerative Vs divisive
- ☒ single link Vs complete link

⌘ Partitional clustering

- ☒ distance-based: K-means
- ☒ model-based: EM
- ☒ density-based:

Partitional methods: K-means

⌘ Criteria: minimize sum of square of distance

- ⊗ Between each point and centroid of the cluster.

- ⊗ Between each pair of points in the cluster

⌘ Algorithm:

- ⊗ Select initial partition with K clusters: random, first K, K separated points

- ⊗ Repeat until stabilization:

- ⊗ Assign each point to closest cluster center

- ⊗ Generate new cluster centers

- ⊗ Adjust clusters by merging/splitting

Collaborative Filtering

- ⌘ Given database of user preferences, predict preference of new user
- ⌘ Example: predict what new movies you will like based on
 - ☑ your past preferences
 - ☑ others with similar past preferences
 - ☑ their preferences for the new movies
- ⌘ Example: predict what books/CDs a person may want to buy
 - ☑ (and suggest it, or give discounts to tempt customer)

Collaborative recommendation

- Possible approaches:

- Average vote along columns [Same prediction for all]
- Weight vote based on similarity of likings [GroupLens]

	Rangeela	QSQT	100 day	Anand	Sholay	Deewar	Vertigo
Smita							
Vijay							
Mohan							
Rajesh							
Nina							
Nitin	?	?		?	?	?	?

Cluster-based approaches

- ⌘ External attributes of people and movies to cluster

- ⊞ age, gender of people

- ⊞ actors and directors of movies.

- ⊞ [May not be available]

- ⌘ Cluster people based on movie preferences

- ⊞ misses information about similarity of movies

- ⌘ Repeated clustering:

- ⊞ cluster movies based on people, then people based on movies, and repeat

- ⊞ ad hoc, might smear out groups

Example of clustering

	Anand QSQT		Rangeela	100 days	Vertigo	Deewar	Sholay
Vijay							
Rajesh							
Mohan							
Nina							
Smita							
Nitin	?	?	?		?	?	?

Model-based approach

- ⌘ People and movies belong to unknown classes
- ⌘ P_k = probability a random person is in class k
- ⌘ P_l = probability a random movie is in class l
- ⌘ P_{kl} = probability of a class- k person liking a class- l movie
- ⌘ Gibbs sampling: iterate
 - ⊞ Pick a person or movie at random and assign to a class with probability proportional to P_k or P_l
 - ⊞ Estimate new parameters
 - ⊞ Need statistics background to understand details



Association Rules

Association rules

T

- ⌘ Given set T of groups of items
- ⌘ Example: set of item sets purchased
- ⌘ Goal: find all rules on itemsets of the form $a \rightarrow b$ such that
 - ⌘ support of a and b $>$ user threshold s
 - ⌘ conditional probability (confidence) of b given a $>$ user threshold c
- ⌘ Example: Milk \rightarrow bread
- ⌘ Purchase of product A \rightarrow service B

Milk, cereal
Tea, milk
Tea, rice, bread
cereal

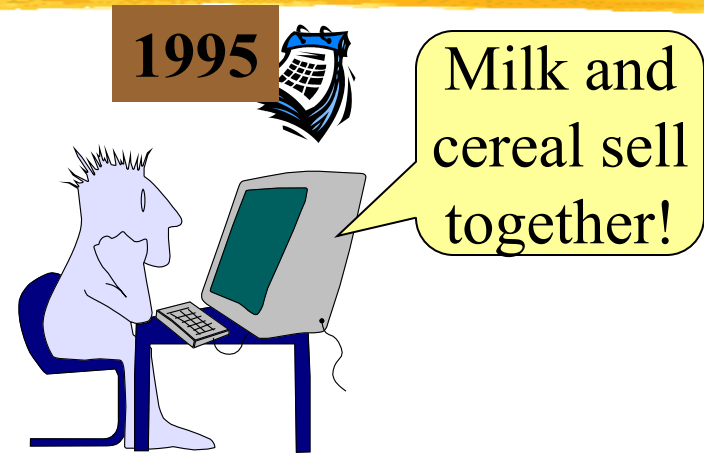
Variants



- ⌘ High confidence may not imply high correlation
- ⌘ Use correlations. Find expected support and large departures from that interesting..
 - ⊠ see statistical literature on contingency tables.
- ⌘ Still too many rules, need to prune...

Prevalent \neq Interesting

- ⌘ Analysts already know about prevalent rules
- ⌘ Interesting rules are those that *deviate* from prior expectation
- ⌘ Mining's payoff is in finding *surprising* phenomena



What makes a rule surprising?

⌘ Does not match prior expectation

☑ Correlation between milk and cereal remains roughly constant over time

⌘ Cannot be trivially derived from simpler rules

☑ Milk 10%, cereal 10%

☑ Milk and cereal 10% ... surprising

☑ Eggs 10%

☑ Milk, cereal and eggs 0.1% ... surprising!

☑ Expected 1%

Applications of fast itemset counting



Find correlated events:

- ⌘ Applications in medicine: find redundant tests
- ⌘ Cross selling in retail, banking
- ⌘ Improve predictive capability of classifiers that assume attribute independence
- ⌘ New similarity measures of categorical attributes [**Mannila et al, KDD 98**]



Data Mining in Practice

Application Areas



Industry

Finance

Insurance

Telecommunication

Transport

Consumer goods

Data Service providers

Utilities

Application

Credit Card Analysis

Claims, Fraud Analysis

Call record analysis

Logistics management

promotion analysis

Value added data

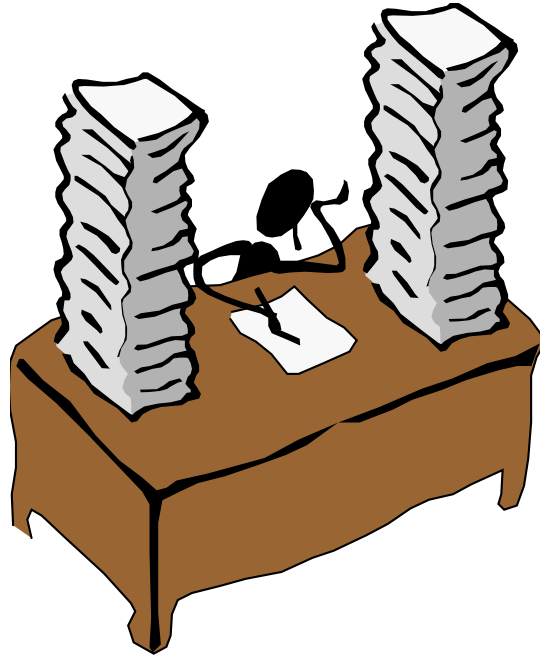
Power usage analysis

Why Now?



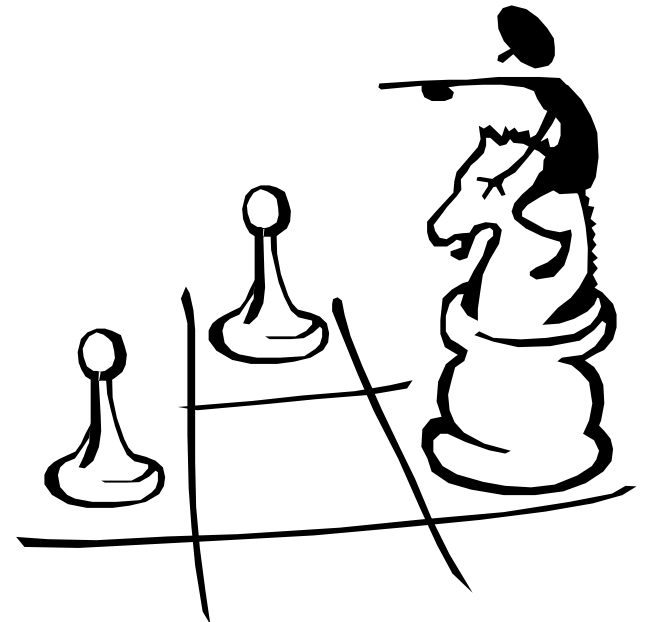
- ⌘ Data is being produced
- ⌘ Data is being warehoused
- ⌘ The computing power is available
- ⌘ The computing power is affordable
- ⌘ The competitive pressures are strong
- ⌘ Commercial products are available

Data Mining works with Warehouse Data



⌘ Data Warehousing provides the Enterprise with a memory

Ñ Data Mining provides the Enterprise with intelligence



Usage scenarios



⌘ Data warehouse mining:

- ☑ assimilate data from operational sources

- ☑ mine static data

⌘ Mining log data

⌘ Continuous mining: example in process control

⌘ Stages in mining:

- ☑ data selection → pre-processing: cleaning

- transformation → mining → result

- evaluation → visualization

Mining market



⌘ Around 20 to 30 mining tool vendors

⌘ Major tool players:

- ☒ Clementine,
- ☒ IBM's Intelligent Miner,
- ☒ SGI's MineSet,
- ☒ SAS's Enterprise Miner.

⌘ All pretty much the same set of tools

⌘ Many embedded products:

- ☒ fraud detection:
- ☒ electronic commerce applications,
- ☒ health care,
- ☒ customer relationship management: Epiphany

Vertical integration:

Mining on the web



⌘ Web log analysis for site design:

☑ what are popular pages,

☑ what links are hard to find.

⌘ Electronic stores sales enhancements:

☑ recommendations, advertisement:

☑ **Collaborative filtering**: Net perception, Wisewire

☑ Inventory control: what was a shopper looking for and could not find..

OLAP Mining integration



⌘ OLAP (On Line Analytical Processing)

- ☑ Fast interactive exploration of multidim. aggregates.
 - ☑ Heavy reliance on manual operations for analysis:
 - ☑ Tedious and error-prone on large multidimensional data
- ⌘ Ideal platform for vertical integration of mining but needs to be interactive instead of batch.

State of art in mining OLAP integration

- ⌘ Decision trees [**Information discovery**, Cognos]
 - ☒ find factors influencing high profits
- ⌘ Clustering [Pilot software]
 - ☒ segment customers to define hierarchy on that dimension
- ⌘ Time series analysis: [Seagate's HoloS]
 - ☒ Query for various shapes along time: eg. spikes, outliers
- ⌘ Multi-level Associations [Han et al.]
 - ☒ find association between members of dimensions
- ⌘ Sarawagi [VLDB2000]

Data Mining in Use



- ⌘ The US Government uses Data Mining to track fraud
- ⌘ A Supermarket becomes an information broker
- ⌘ Basketball teams use it to track game strategy
- ⌘ Cross Selling
- ⌘ Target Marketing
- ⌘ Holding on to Good Customers
- ⌘ Weeding out Bad Customers

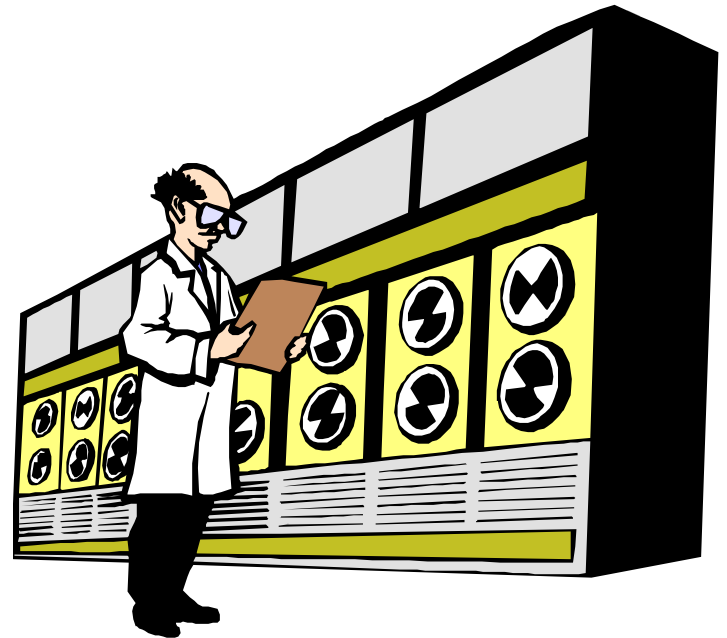
Some success stories



- ⌘ Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
 - ⊞ Won over (manual) knowledge engineering approach
 - ⊞ <http://www.cs.columbia.edu/~sal/JAM/PROJECT/> provides good detailed description of the entire process
- ⌘ Major US bank: customer attrition prediction
 - ⊞ First segment customers based on financial behavior: found 3 segments
 - ⊞ Build attrition models for each of the 3 segments
 - ⊞ 40-50% of attritions were predicted == factor of 18 increase
- ⌘ Targeted credit marketing: major US banks
 - ⊞ find customer segments based on 13 months credit balances
 - ⊞ build another response model based on surveys
 - ⊞ increased response 4 times -- 2%

Prof. S. Sudarshan
CSE Dept, IIT Bombay

Most slides courtesy:
Prof. Sunita Sarawagi
School of IT, IIT Bombay



Data Mining

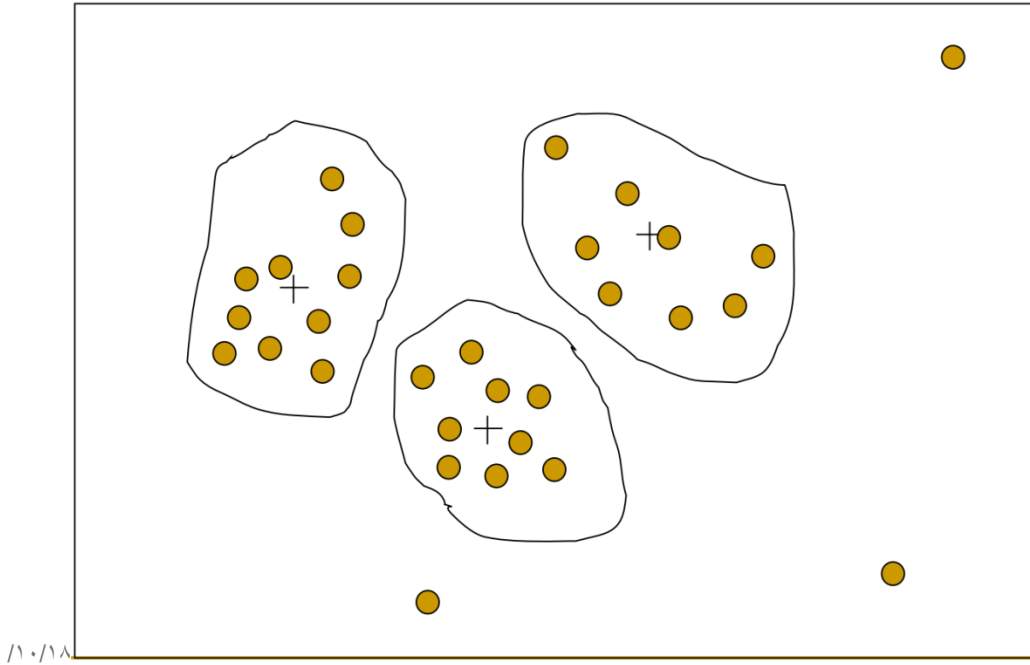
Data Mining Steps

- Data Cleaning.
- Data Integration.
- Data Reduction.
- Data Transformation.
- Data Mining.
- Pattern Evaluation.
- Knowledge Representation.

Data Cleaning

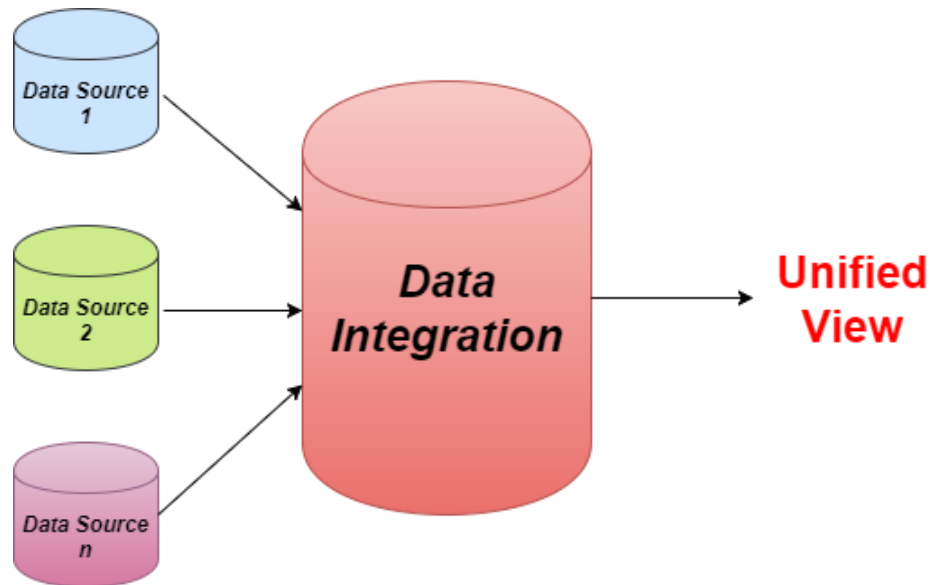
Detecting and removing corrupt or inaccurate records from a record set, table or database. Some data cleaning methods :- 1 You can ignore the tuple. This is done when class label is missing

- Remove irrelevant data.
- Standardize capitalization.
- Convert data type.
- Clear formatting.
- Fix errors.
- Language translation.
- Handle missing values.



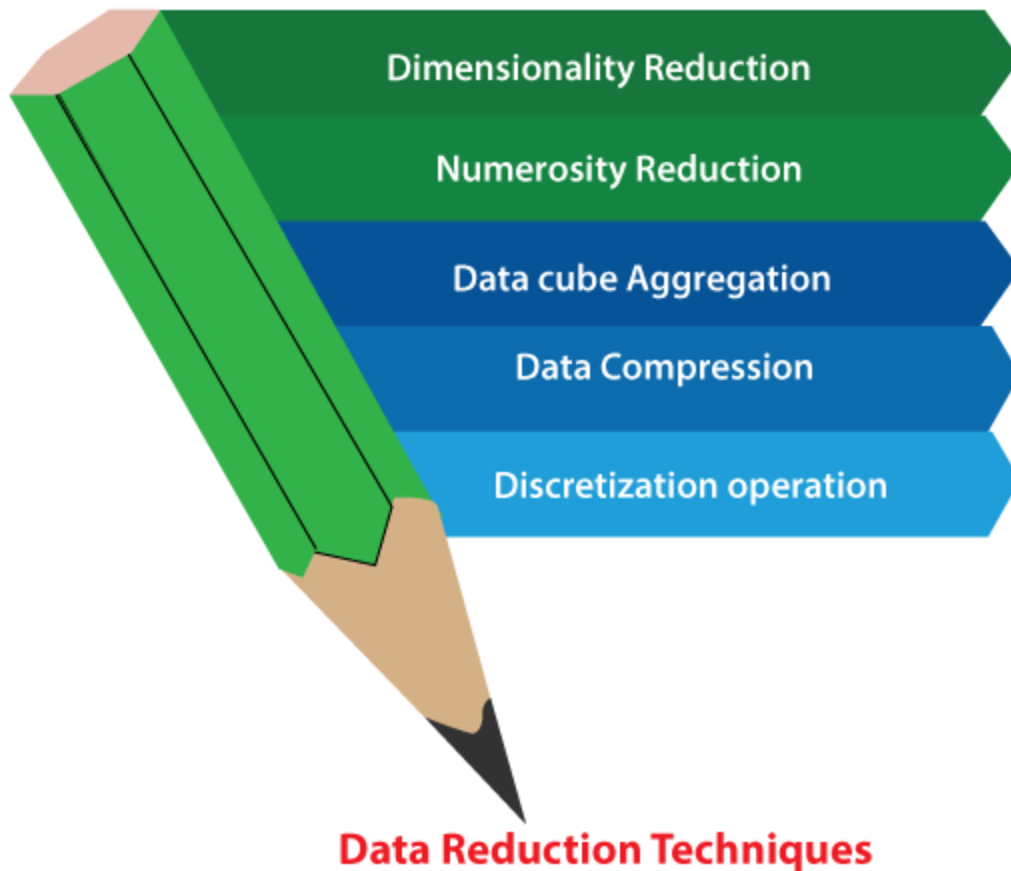
Data Integration

Disparate sources together to provide users with a unified view. The premise of data integration is to make data more freely available and easier to consume and process by systems and users



Data Reduction.

Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data



Data Transformation

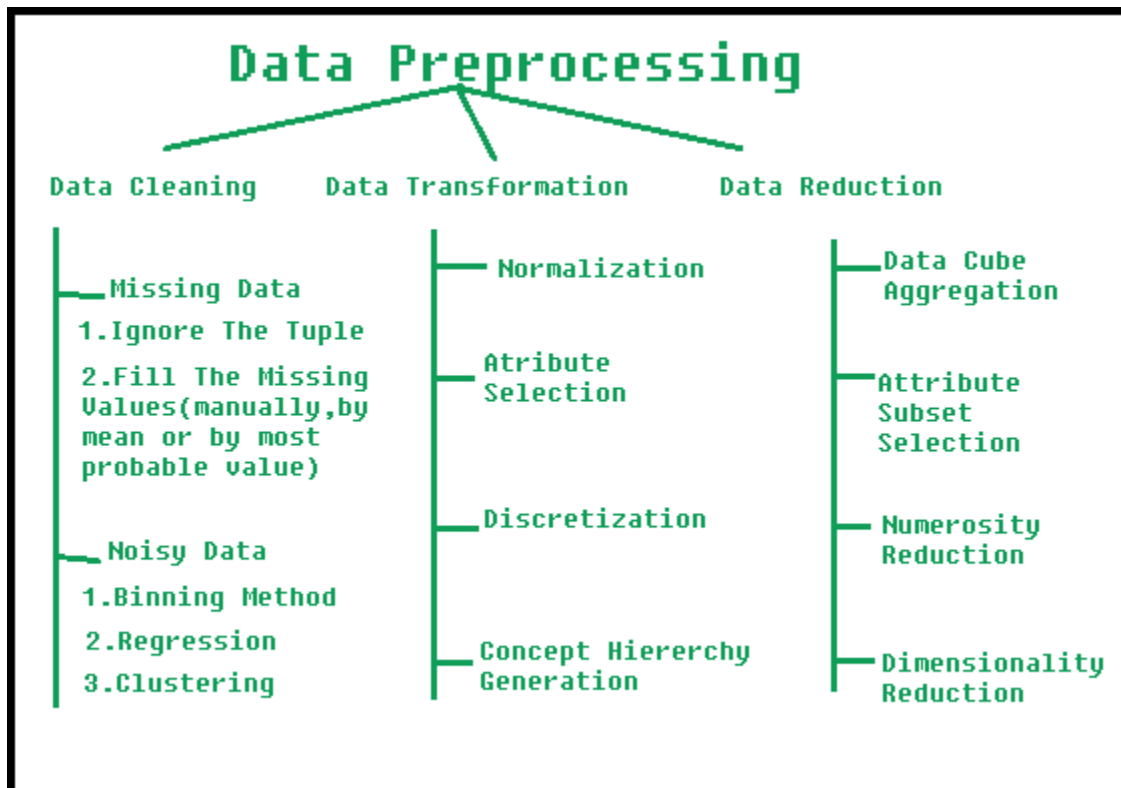
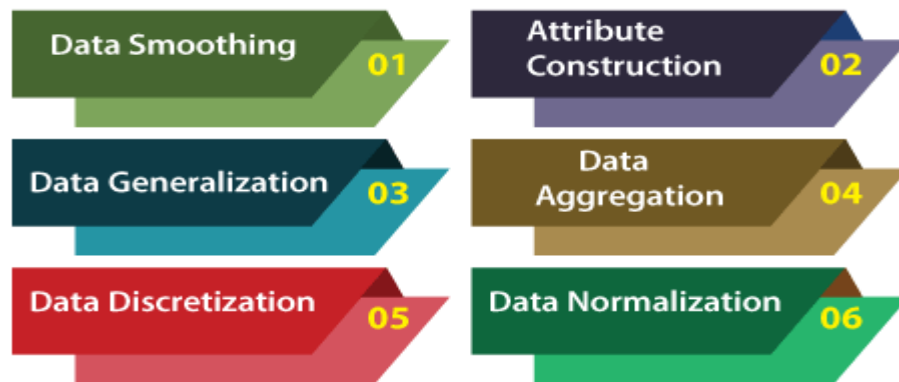
Data transformation is a **technique used to convert the raw data into a suitable format that efficiently eases data mining and retrieves strategic information.** Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

Types of Data Transformations

- Bucketing/Binning.
- Data Aggregation.
- Data Cleansing.
- Data Deduplication.

- Data Derivation.
- Data Filtering.
- Data Integration.
- Data Joining.

Data Transformation Techniques



Data Mining

Data Mining is a process to identify interesting patterns and knowledge from a large amount of data. In these steps, **intelligent patterns are applied to extract the data patterns**. The data is represented in the form of patterns and models are structured using classification and clustering techniques.

Organizations use data mining applications to extract useful trends and optimize knowledge discovery to generate business intelligence. This is only possible if a company takes full advantage of big data and collects the correct type of information.

Engineers apply intelligent patterns to the available data before they extract it. They then represent all information as models. Specialists use clustering, classification, or other modeling techniques to ensure accuracy.

- Classification analysis. This analysis is used to retrieve important and relevant information about data, and metadata. ...
- Association rule learning. ...
- Anomaly or outlier detection. ...
- Clustering analysis. ...
- Regression analysis.

Pattern Evaluation

This is the stage where engineers stop working behind the scenes and bring insights into the real world. Specialists will pinpoint any useful patterns that can generate business knowledge.

They will use their models, historical data, and real-time information to find out more about customers, employees, and sales. Teams will also summarize information data or use visualization data mining techniques to make it easier to understand.

Representing Knowledge in Data Mining

Knowledge representation is **the presentation of knowledge to the user for visualization in terms of trees, tables, rules graphs, charts, matrices, etc.**