

Introduction to Data Mining

Prepared by Dr. Amjad Qtaish

DS601
Data Mining

Lecture 01



Lecture: #01 Introduction to Data Mining



Dr. Amjad Qtaish

Faculty of Informatics
Midocean University



Data Mining (DS601)



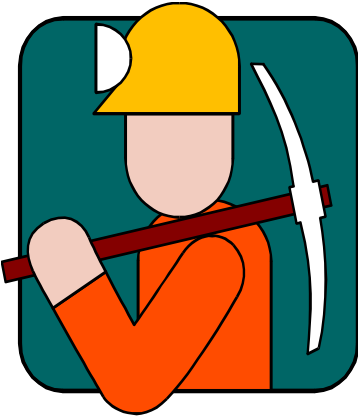
Introduction to Data Mining

- 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
- Summary





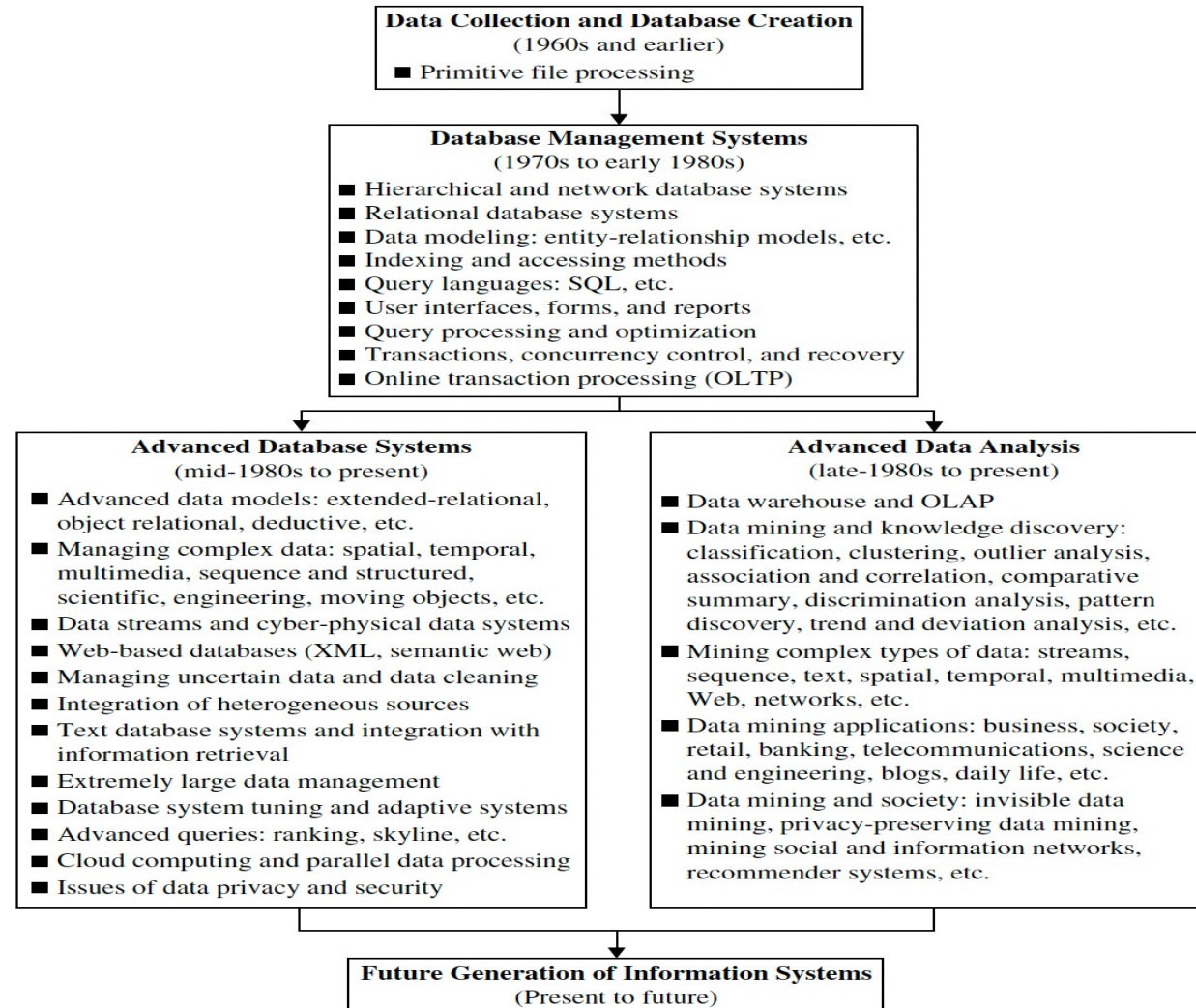
Why Data Mining?



- **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

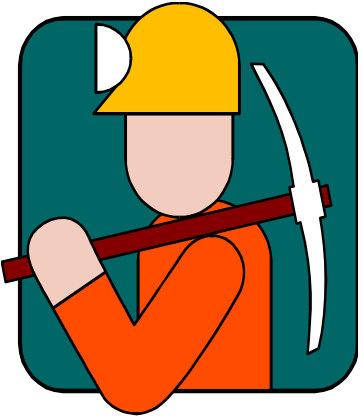


Why Data Mining?





Why Data Mining?

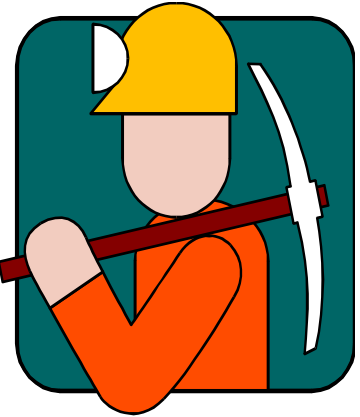




Introduction to Data Mining

- 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
- Summary

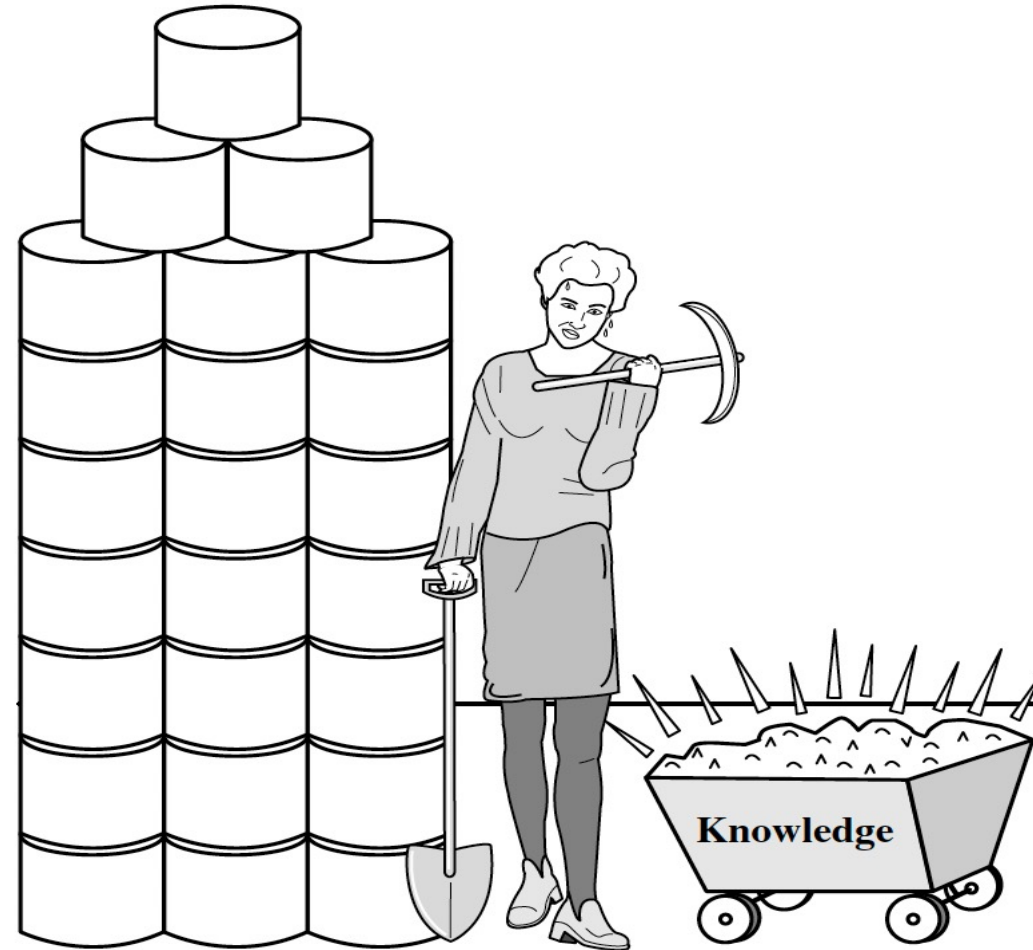
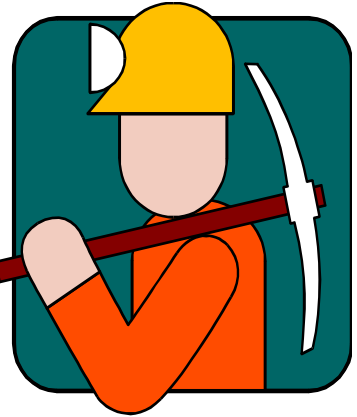




- **Data mining (knowledge discovery from data)**
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- **Alternative names**
 - Knowledge discovery (mining) in databases (KDD), knowledge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

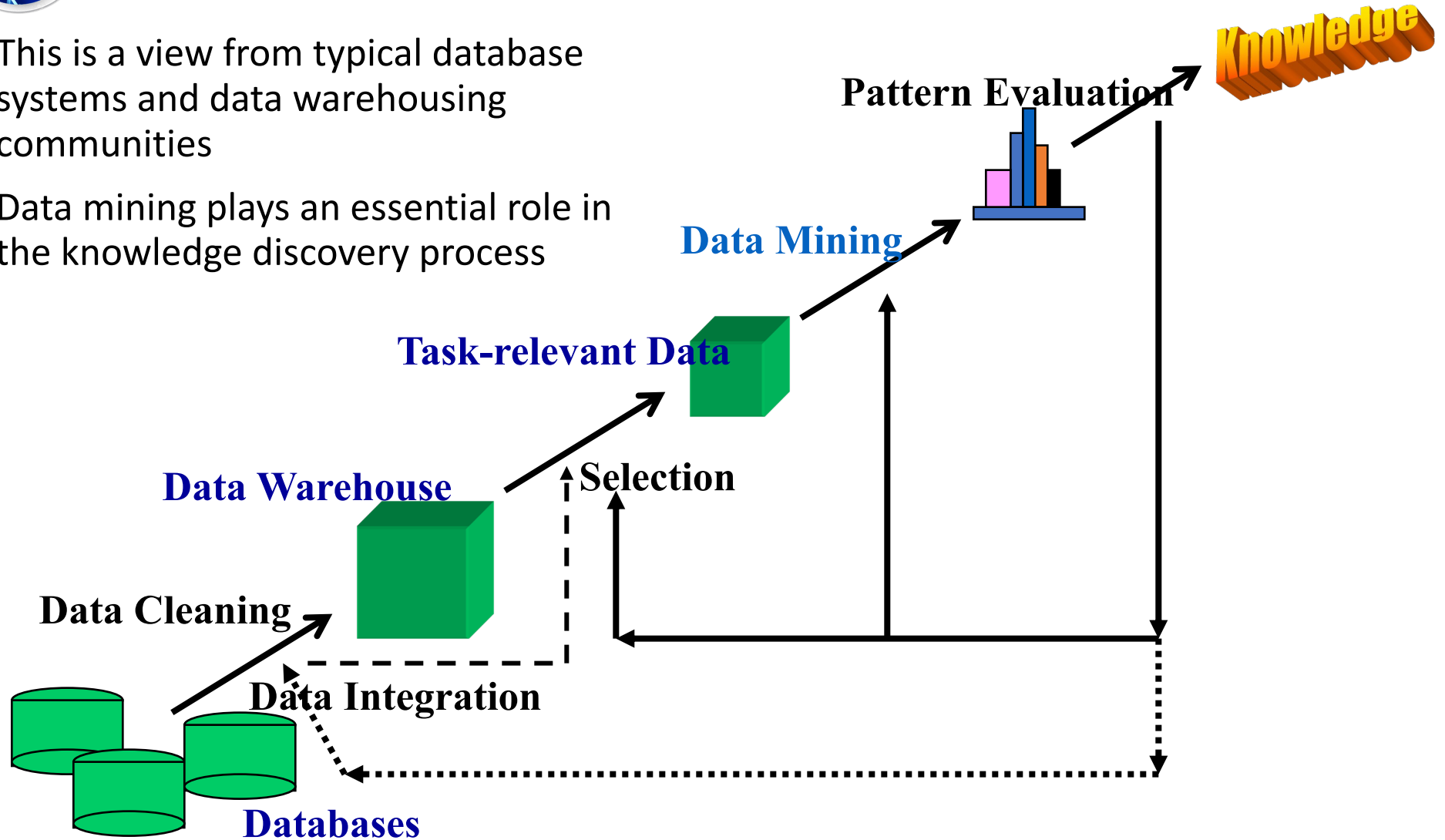
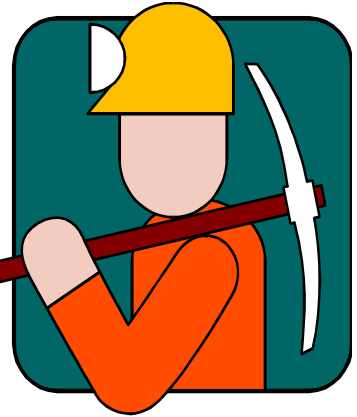


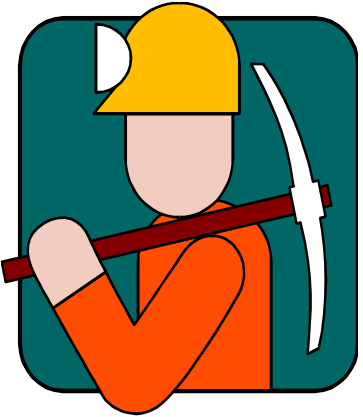
What Is Data Mining?





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

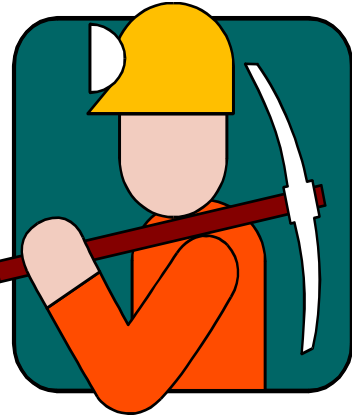




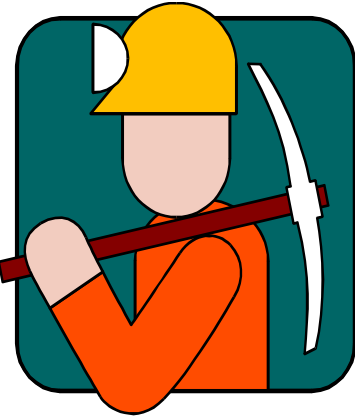
- **Data cleaning** (to remove noise and inconsistent data)
- **Data integration** (where multiple data sources may be combined)
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)



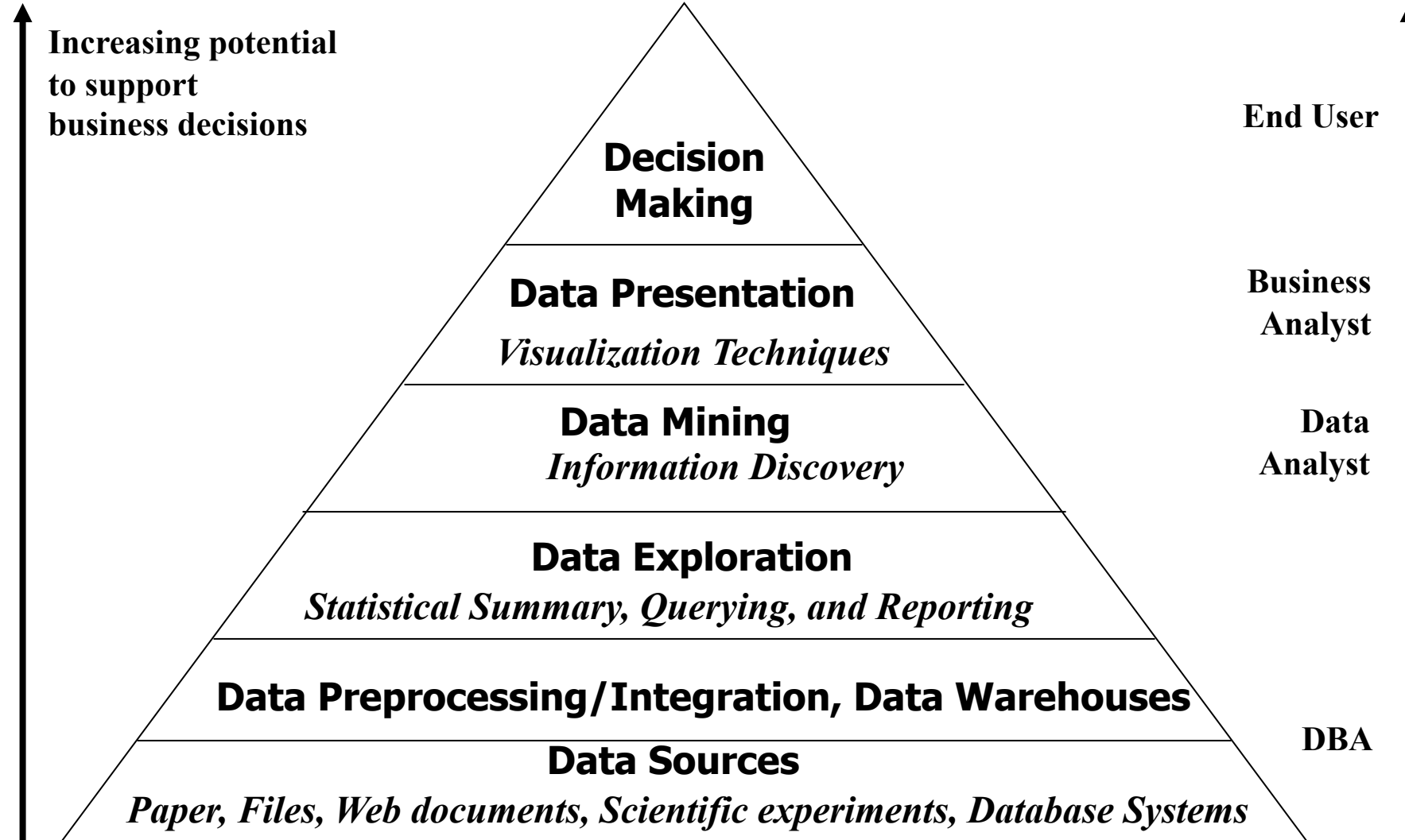
- **Web mining usually involves**

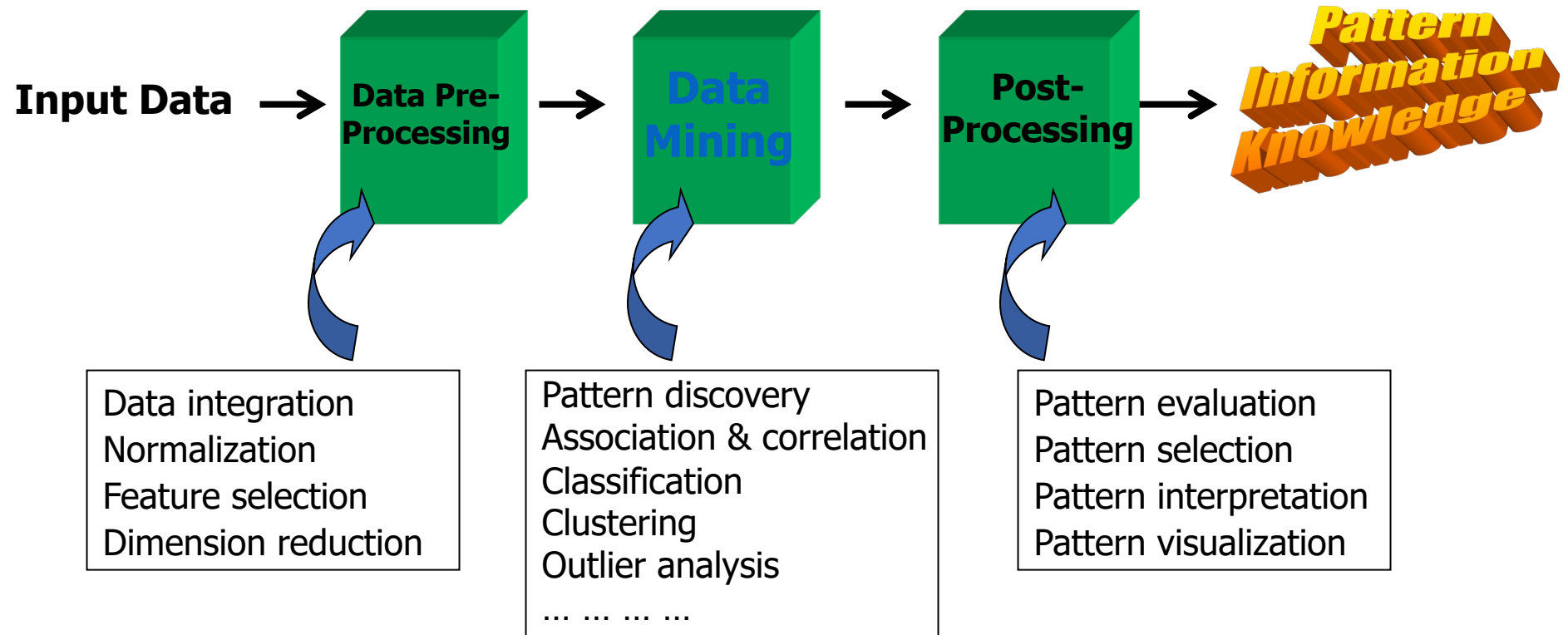
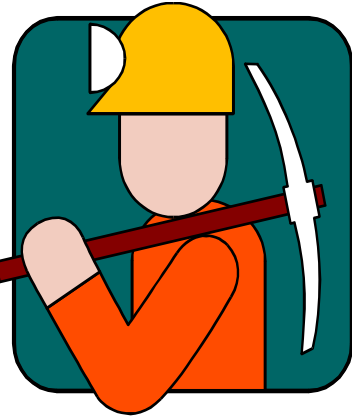


- Data cleaning
- Data integration from multiple sources
- Warehousing the data
- Data cube construction
- Data selection for data mining
- Data mining
- Presentation of the mining results
- Patterns and knowledge to be used or stored into knowledge-base

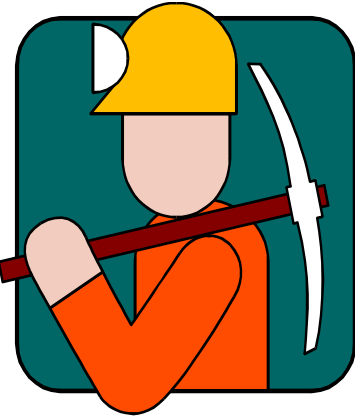


Increasing potential
to support
business decisions





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



- 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

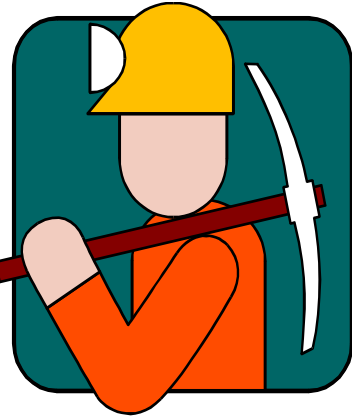




Introduction to Data Mining

- 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
- Summary





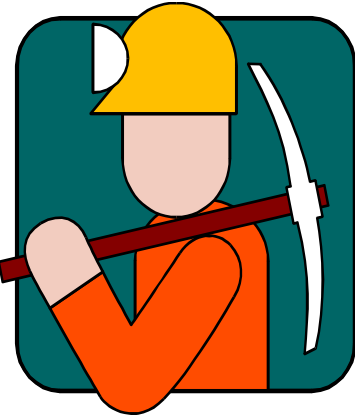
- **Data to be mined**
 - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
 - 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.



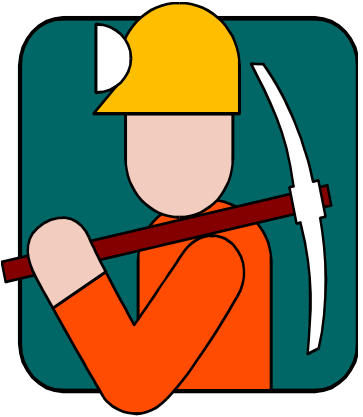
Introduction to Data Mining

- 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
- Summary

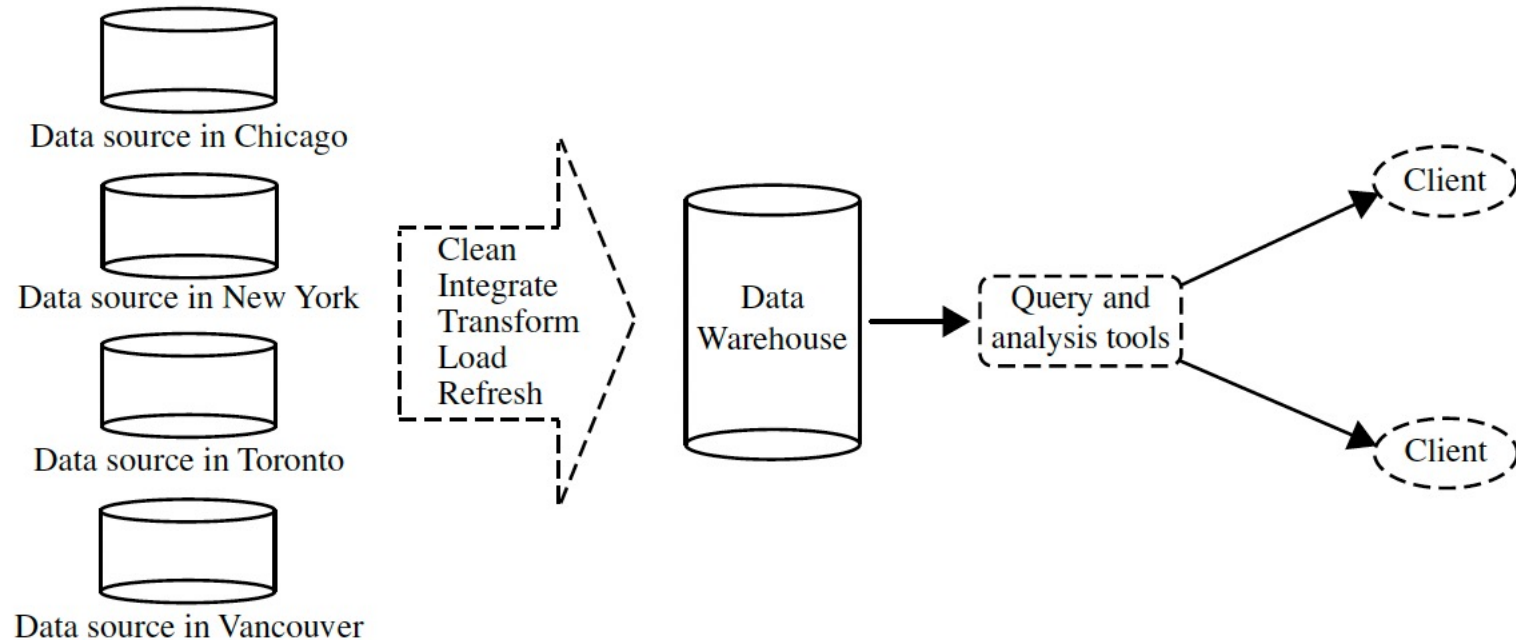


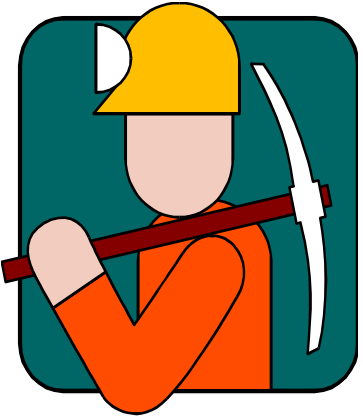


- **Database-oriented data sets and applications**
 - Relational database, data warehouse, transactional database
- **Advanced data sets and advanced applications**
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data (incl. bio-sequences)
 - Structure data, graphs, social 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

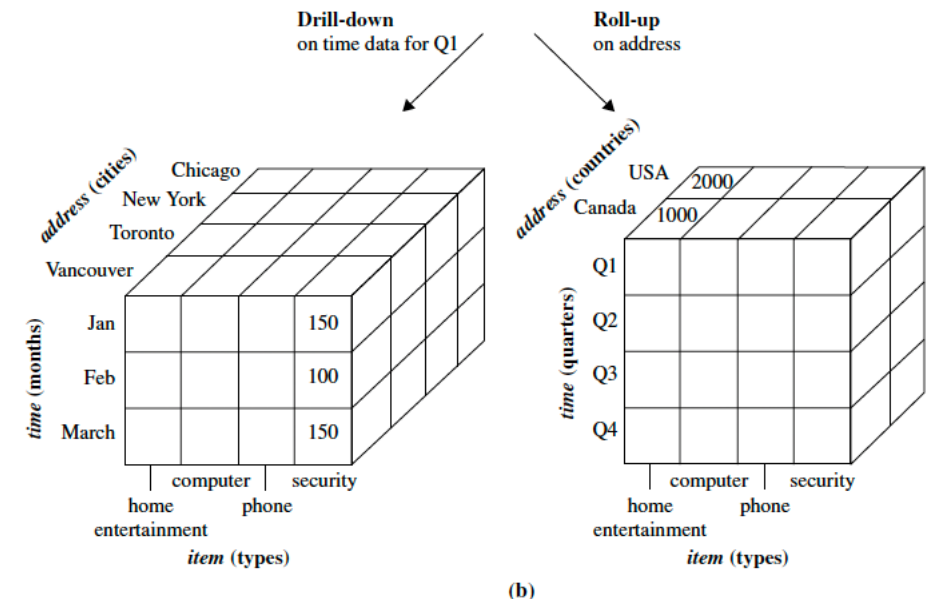
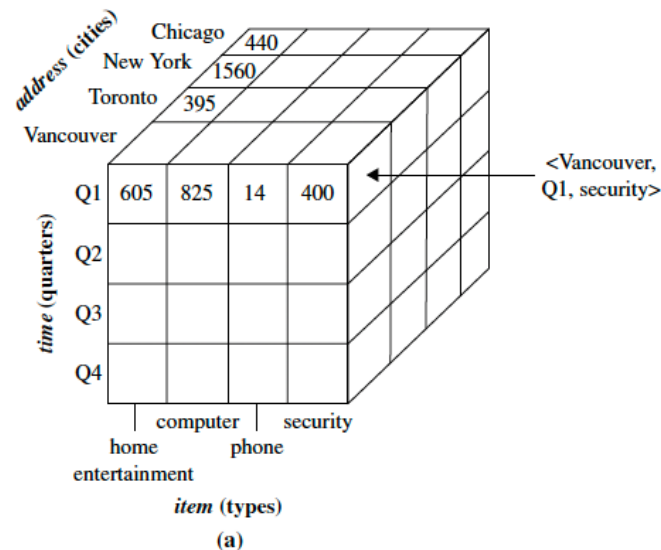


- A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.





- A **data warehouse** is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or *sum(sales amount)*.
- A **data cube** provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

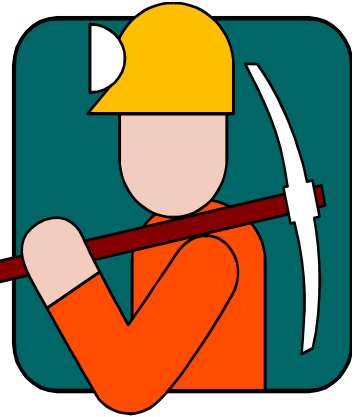




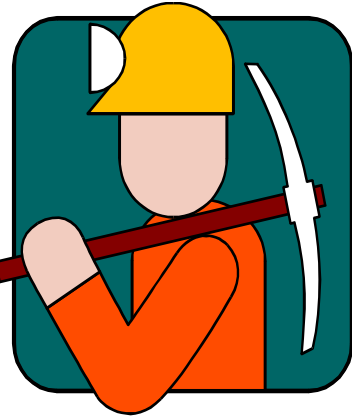
Introduction to Data Mining

- 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
- Summary

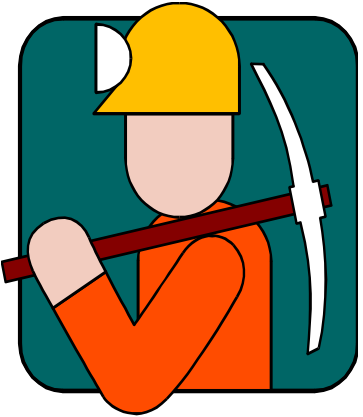




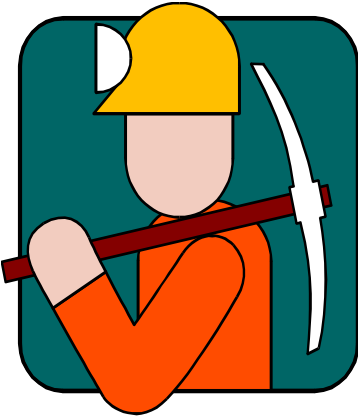
- **Information integration and data warehouse construction**
 - Data cleaning, transformation, integration, and multidimensional data model
- **Data cube technology**
 - Scalable methods for computing (i.e., materializing) multidimensional aggregates
 - OLAP (online analytical processing)
- **Multidimensional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



- **Data characterization** is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.
- **For example**, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.
- **Data discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
- **For example**, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.



- **Frequent patterns (or frequent itemsets)**
 - What items are frequently purchased together in your Walmart?
- **Association, correlation vs. causality**
 - A typical association rule
 $\text{Buys}(X, \text{"computer"}) \rightarrow \text{buys}(X, \text{"software"})$ [support=1%, confidence= 50%]
 - where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together
 - 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?**

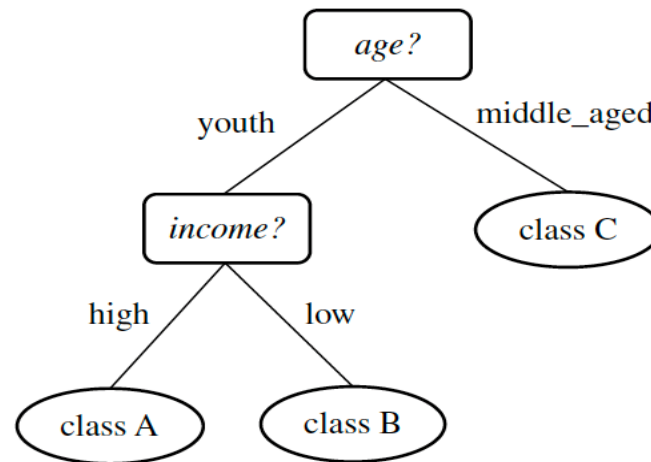


- **Classification and label prediction**
 - Construct models (functions) based on some training examples
 - Describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
 - Predict some unknown class labels
- **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, ...

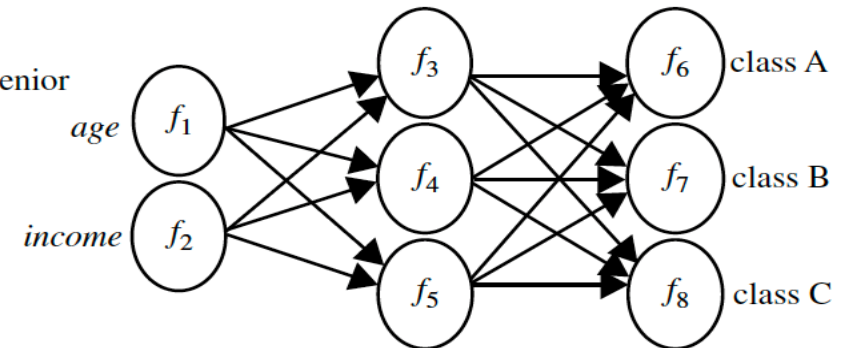


$age(X, \text{"youth"}) \text{ AND } income(X, \text{"high"}) \longrightarrow class(X, \text{"A"})$
 $age(X, \text{"youth"}) \text{ AND } income(X, \text{"low"}) \longrightarrow class(X, \text{"B"})$
 $age(X, \text{"middle_aged"}) \longrightarrow class(X, \text{"C"})$
 $age(X, \text{"senior"}) \longrightarrow class(X, \text{"C"})$

(a)

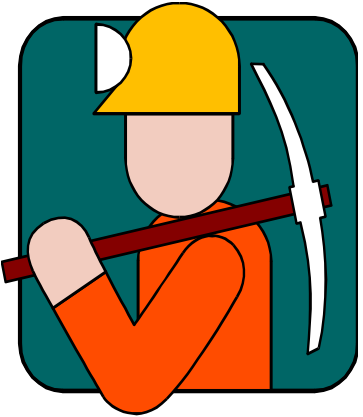


(b)



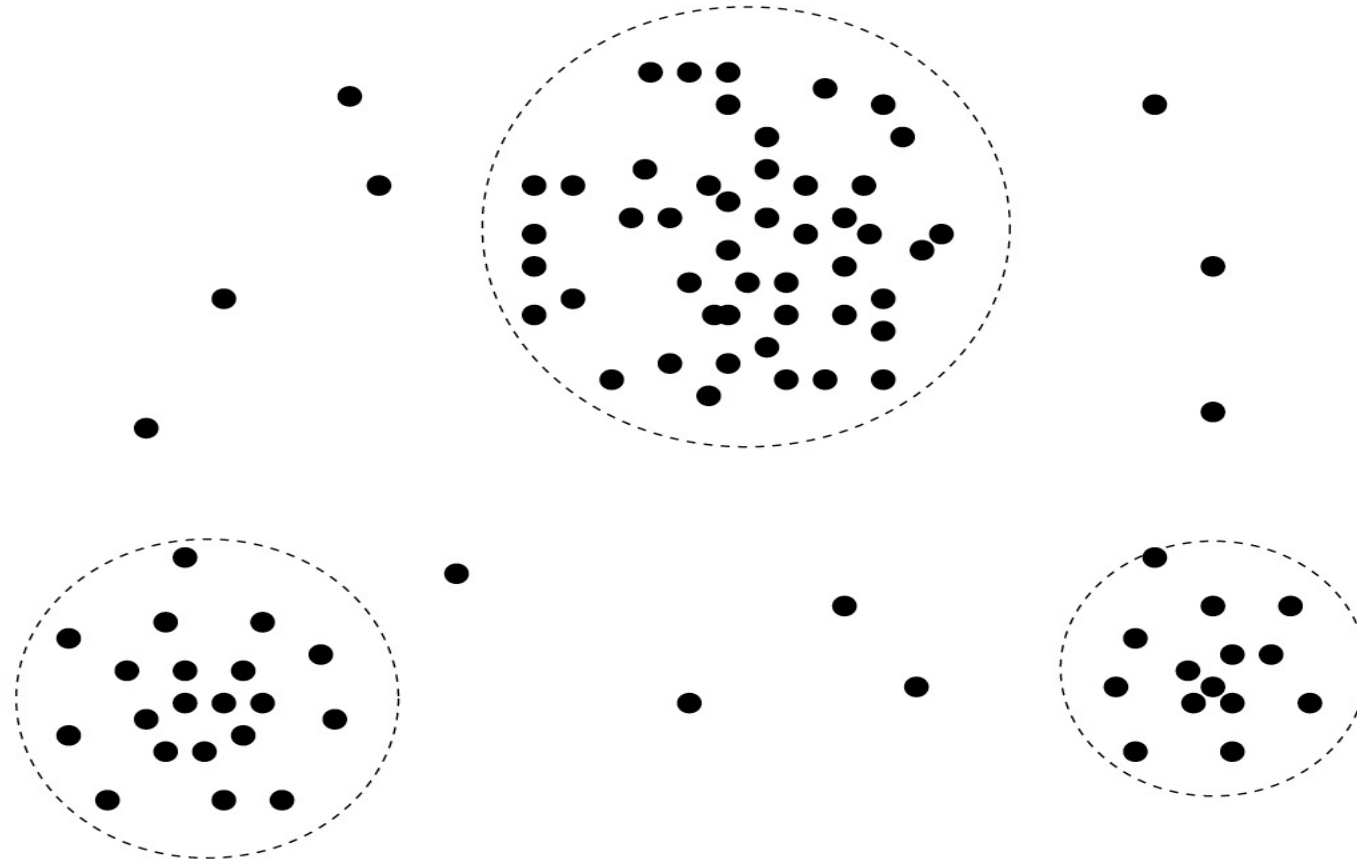
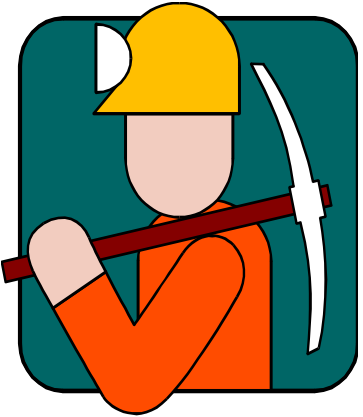
(c)

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.



- **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

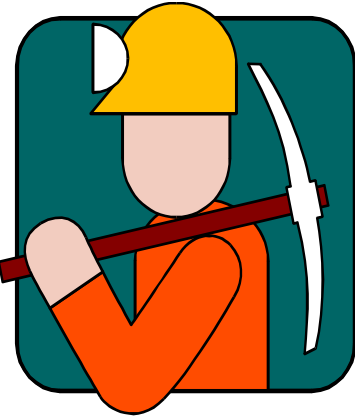




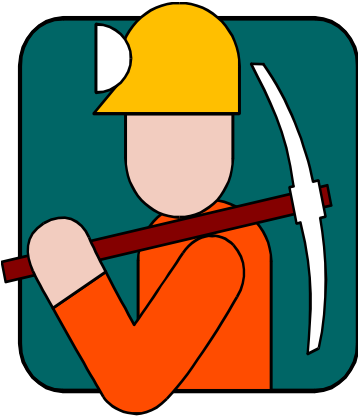
A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.



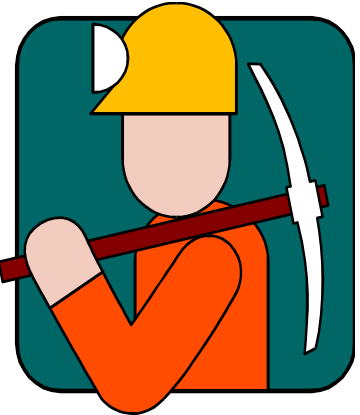
- **Outlier analysis**



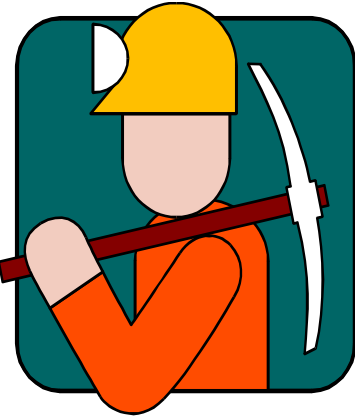
- **Outlier:** A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- **Methods:** by product of clustering or regression analysis, ...
- Useful in fraud detection, rare events analysis



- **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



- **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, ...



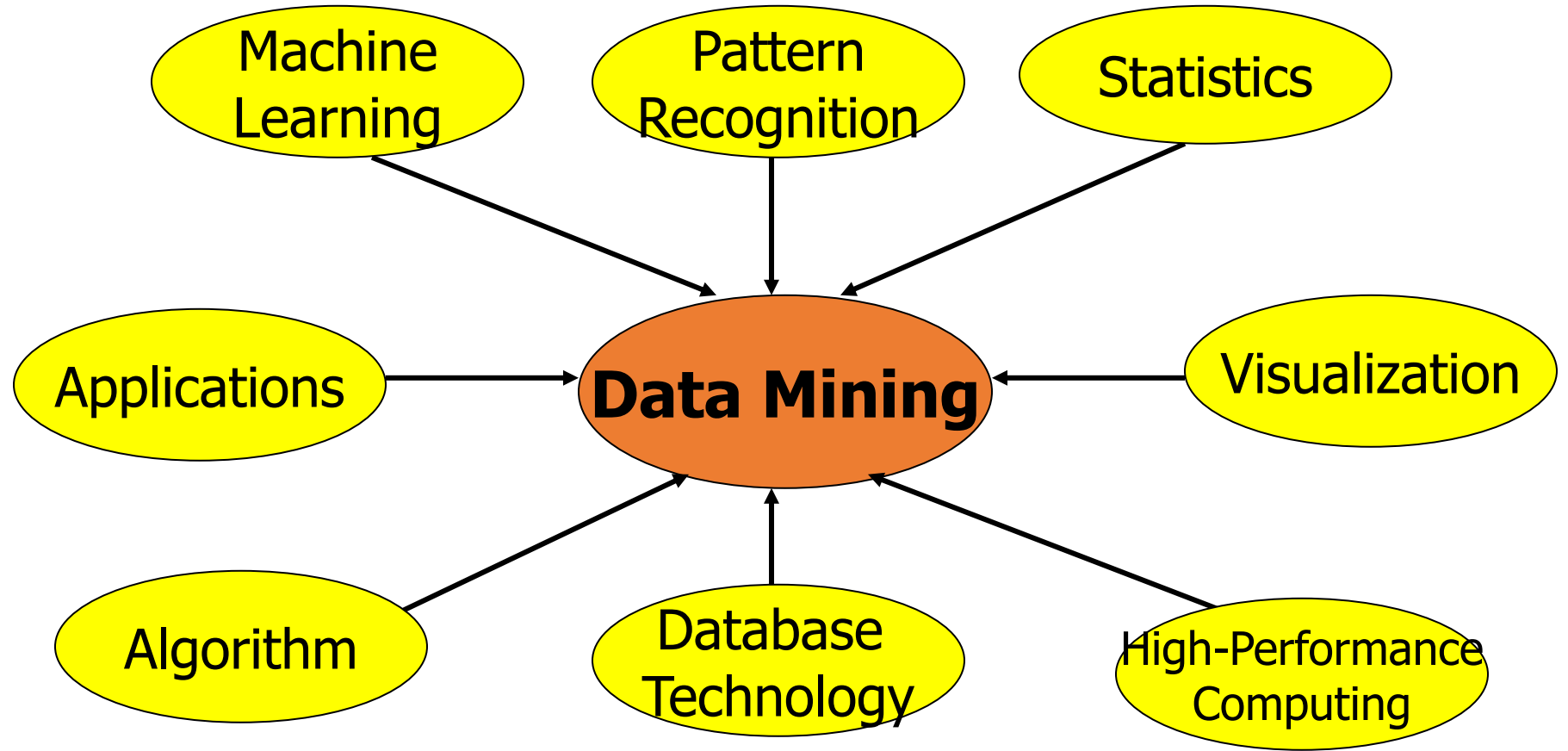
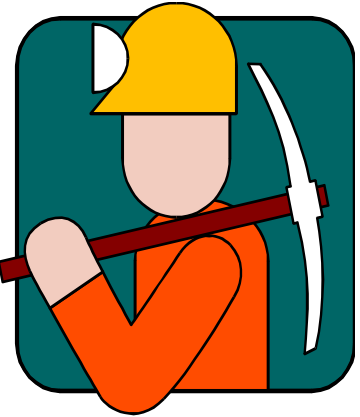
- **Are all mined knowledge interesting?**
 - One can mine tremendous amount of “patterns” and knowledge
 - Some may fit only certain dimension space (time, location, ...)
 - Some may not be representative, may be transient, ...
- **Evaluation of mined knowledge** → directly mine only interesting knowledge?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

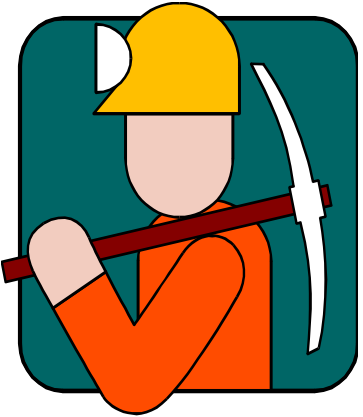


Introduction to Data Mining

- 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
- Summary







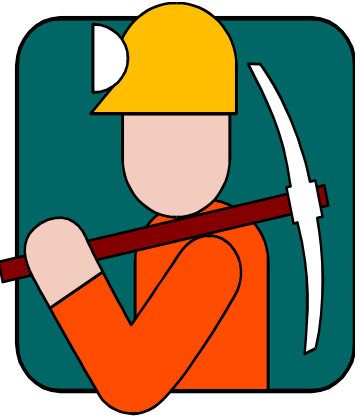
- **Tremendous amount of data**
 - Algorithms must be highly scalable to handle such as tera-bytes of data
- **High-dimensionality of data**
 - Micro-array may have tens of thousands of dimensions
- **High complexity of data**
 - Data streams and sensor data
 - Time-series data, temporal data, sequence 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**



Introduction to Data Mining

- 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
- Summary





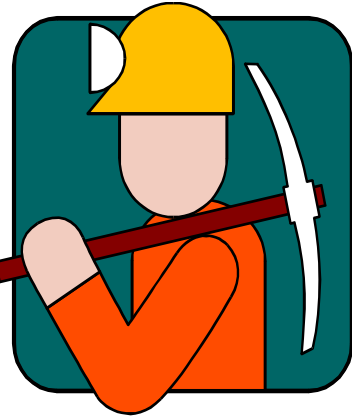
- **Web page analysis:** from web page classification, clustering to PageRank & HITS algorithms
- **Collaborative analysis & recommender systems**
- **Basket data analysis to targeted marketing**
- **Biological and medical data analysis:** classification, cluster analysis (microarray data analysis), 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



Introduction to Data Mining

- 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 
- Summary





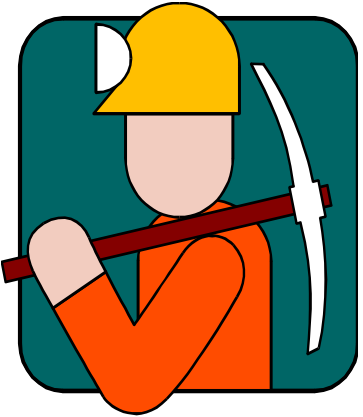
- **Mining Methodology**

- 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





- **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

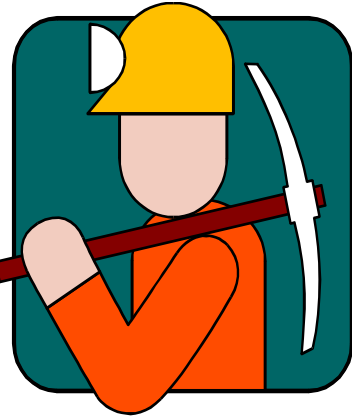




Introduction to Data Mining

- 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
- 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



Thank You