Data Mining

Chapter 1. Introduction

- What motivated Data mining?
- Evolution of Information Technology
- What is Data Mining?
- Potential Applications.
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?

What Motivated Data Mining?

Data explosion problem

Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.

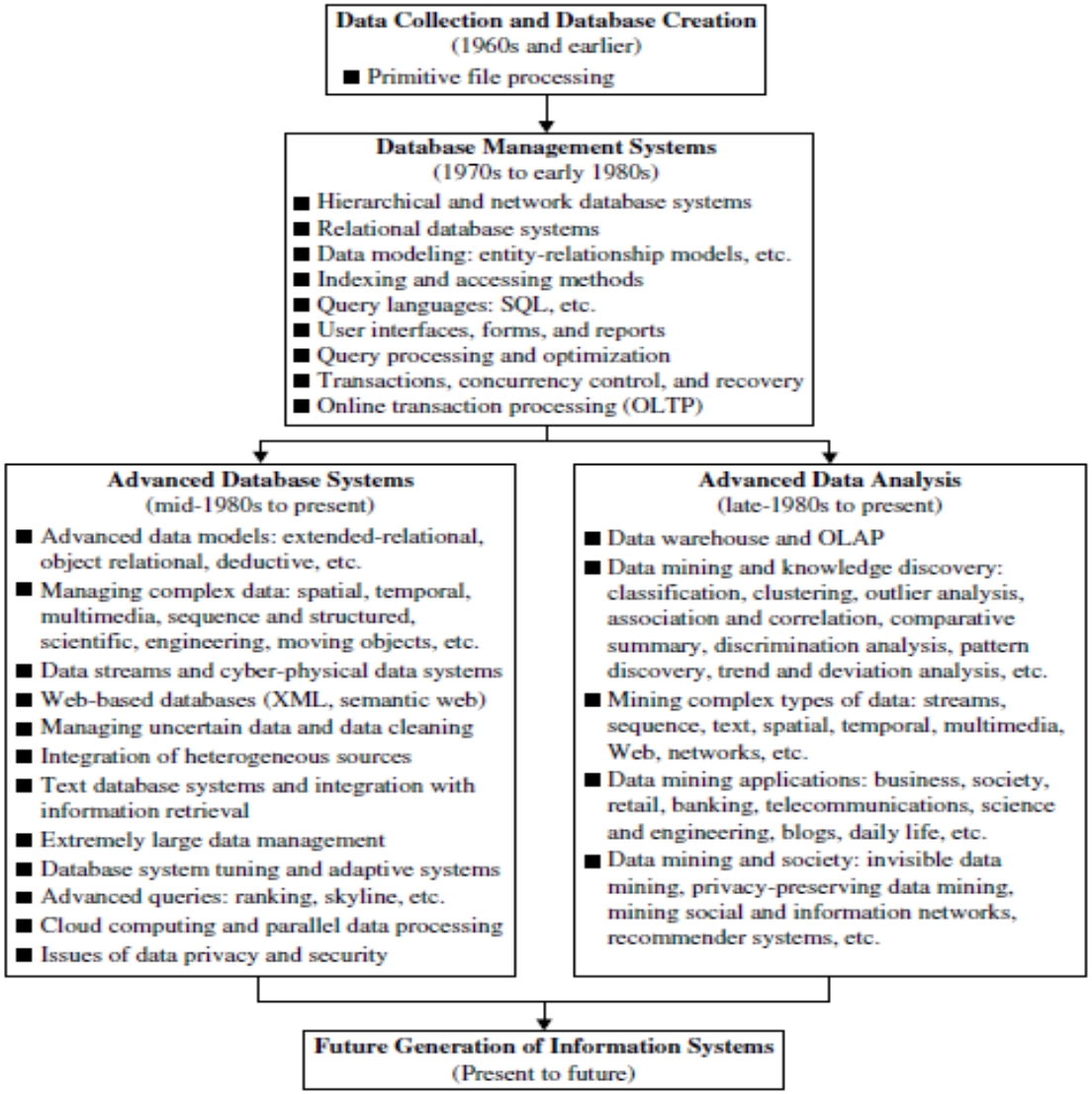
We are drowning in data, but starving for knowledge!

Powerful and versatile tools are needed to automatically extract valuable information from large amount of data and to transform into knowledge.

Solution: Data warehousing and data mining

- Data warehousing and on-line analytical processing (OLAP).
- Extraction of interesting knowledge (**rules, regularities, patterns, constraints**) from data in large databases.

Evolution of Information Technology



What Is Data Mining?

- Data mining turns large collection of data into knowledge.
- Data mining (knowledge discovery or mining from data)
  - Extraction of interesting (**non-trivial, implicit, previously unknown and potentially useful**) patterns or knowledge from huge amount of data

  (or)

  - Process that finds a small set of precious nuggets from great deal of raw materials.
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, information harvesting, business intelligence, etc.

Potential Applications

Database analysis and decision support

- Market analysis and management
  - Target marketing, customer relation management, market basket analysis, cross selling, market segmentation
- Risk analysis and management
  - Forecasting, customer retention, quality control, competitive analysis
- Fraud detection and management

Other Applications

- Text mining (news group, email, documents) and Web analysis.
- Intelligent query answering.

Market Analysis and Management (1)

- Data sources for analysis:
  - Credit card, loyalty cards, discount coupons, customer complaint calls, public lifestyle studies
- Target marketing:
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales.
  - Prediction based on the association information.

Market Analysis and Management (2)

- Customer profiling
  - What types of customers buy what products (clustering or classification).
- Identifying customer requirements
  - Identifying the best products for different customers.
  - Use prediction to find what factors will attract new customers.
- Provides summary information
  - Various multidimensional summary reports.
  - Statistical summary

Corporate Analysis and Risk Management

- Finance planning and asset evaluation
  - Cash flow analysis and prediction.
  - Cross-sectional and time series analysis (financial-ratio, trend analysis, etc).
- Resource planning
  - Summarize and compare the resources and spending.
- Competition
  - Monitor competitors and market directions.
  - Group customers into classes and a class-based pricing procedure.
  - Set pricing strategy in a highly competitive market.

Fraud Detection and Management (1)

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances.
- Examples:
  - auto insurance: detect a group of people who stage accidents to collect on insurance.
  - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network).
  - medical insurance: detect professional patients and ring of doctors and ring of references.

Fraud Detection and Management (2)

- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
- Retail
  - Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat.

- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Knowledge Discovery from Data (KDD)

- The knowledge discovery process is an iterative sequence with following steps:
- **Data cleaning:** To remove noise and inconsistent data
- **Data Integration:** Where multiple data sources may be combined
- **Data selection:** Data relevant to the analysis task are retrieved from the database.
- **Data Transformation:** Data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.

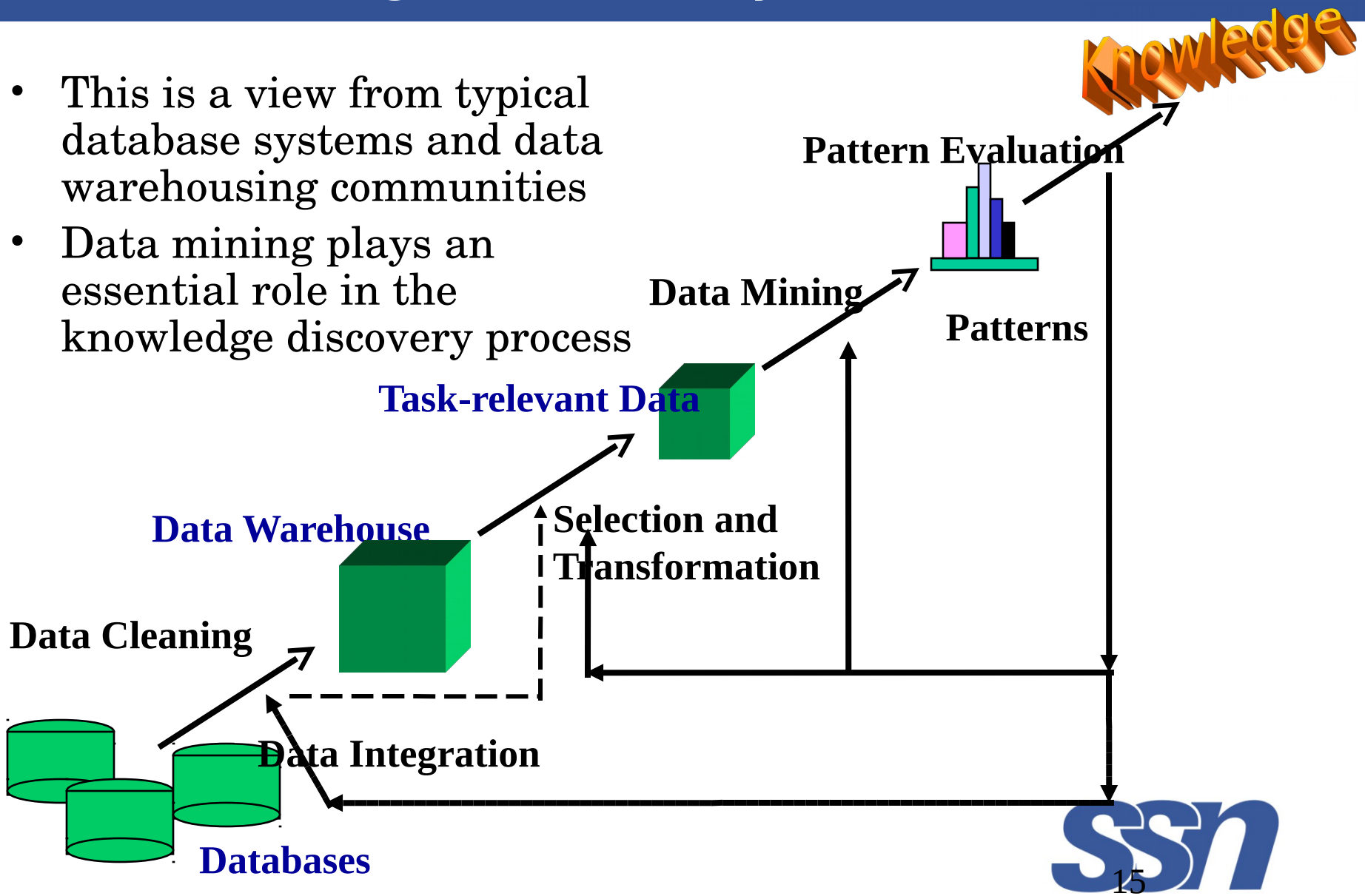
Knowledge Discovery from Data (KDD)

- **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.

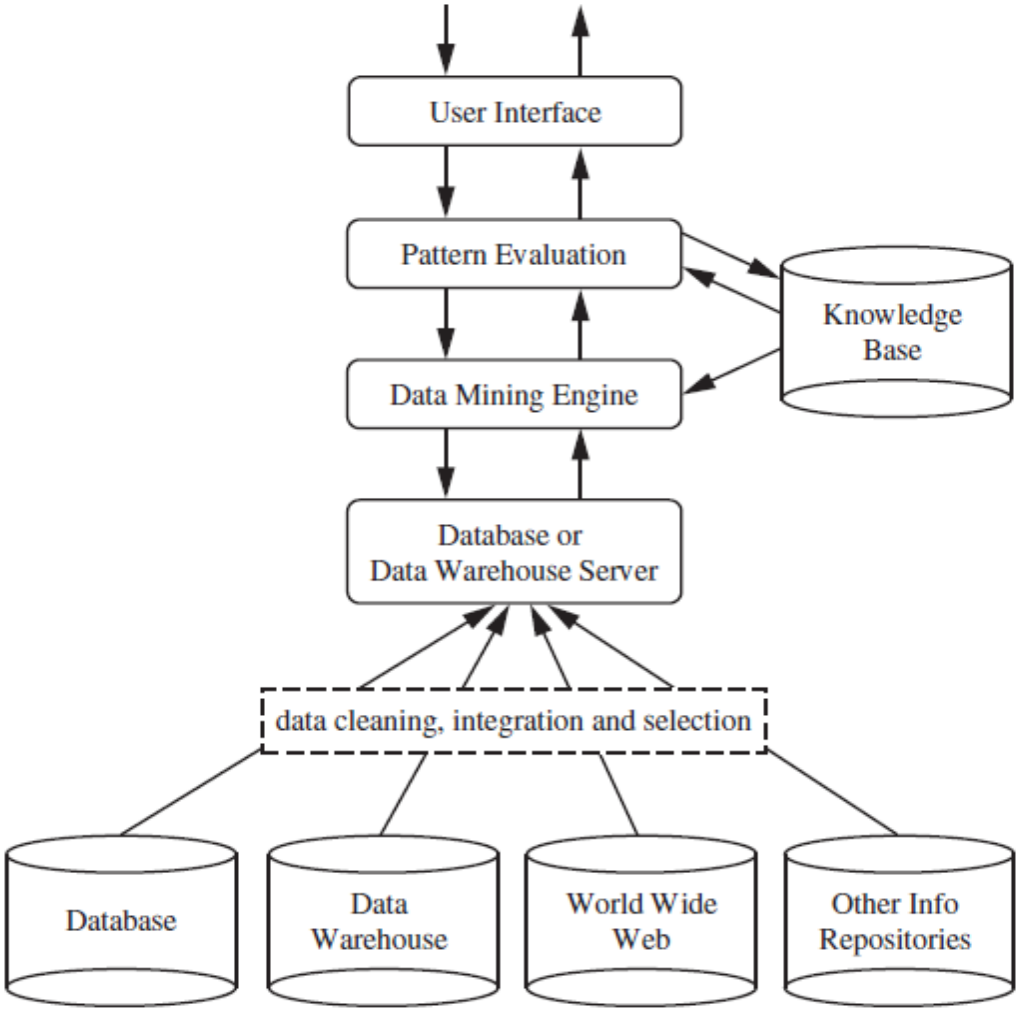
"**Data mining is the process of discovering interesting patterns and knowledge from large amounts of data**"

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



Architecture of Data Mining



Architecture of Data Mining

- Database, Data warehouse, www or other information repository:
  - One or more set of databases, DW, spreadsheets or other kinds of information repositories.
  - Data cleaning and data integration techniques may be performed on the data.
- **Database or Data warehouse server:**
  - It is responsible for fetching the relevant data based on the user's data mining request

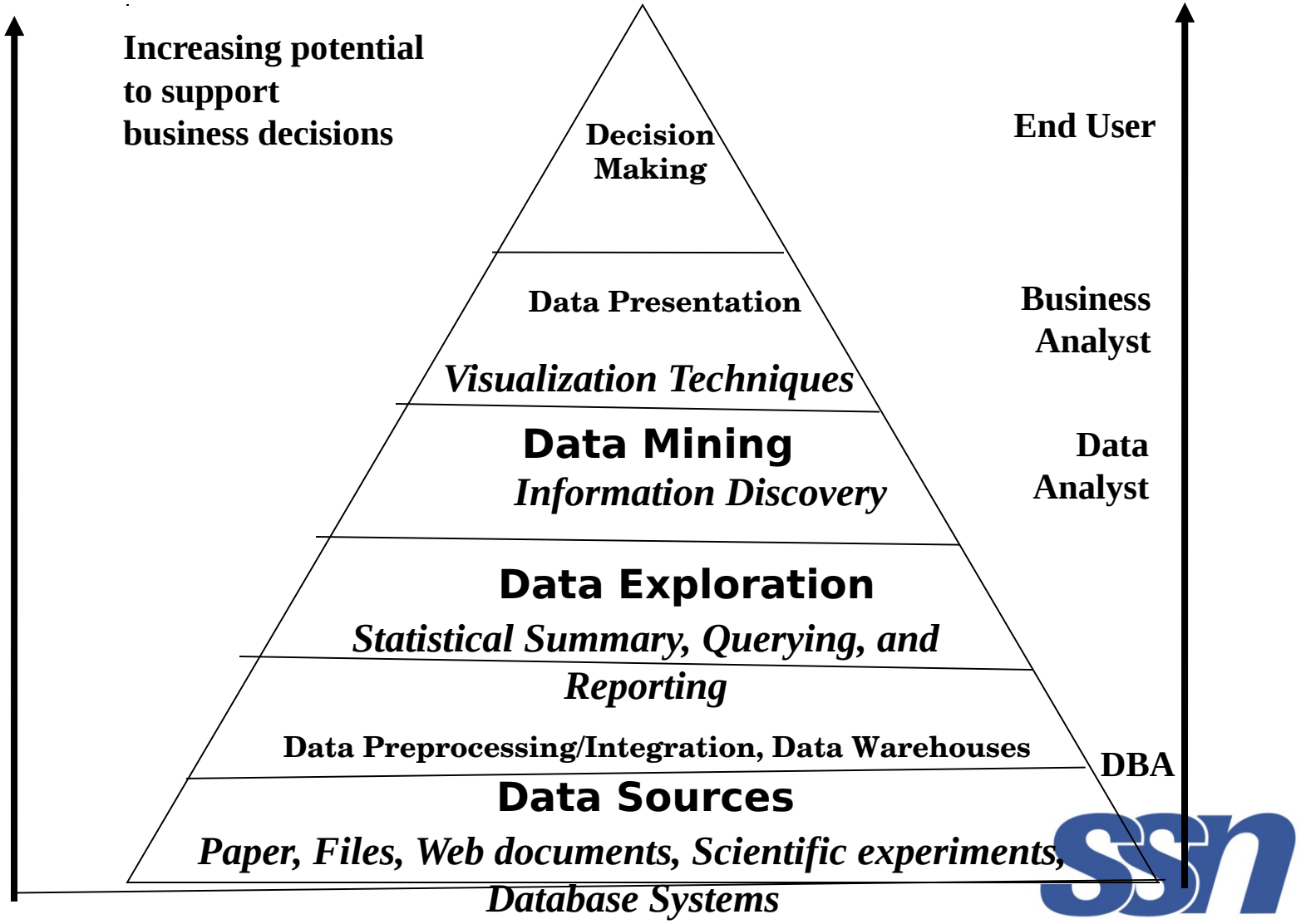
Architecture of Data Mining

- **Knowledge base:**
  - Domain knowledge used to guide the search or evaluate the interestingness of resulting patterns
  - Knowledge includes concept hierarchies, user beliefs, interestingness constraints or thresholds etc.
- **Data mining engine:**
  - Essential to the DM system consists of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outliers analysis and evolution analysis

Architecture of Data Mining

- **Pattern Evaluation module:** Component employs interestingness measures and interacts with the data mining modules to **focus the search toward interesting patterns.**
- **User interface:** Module communicates between users and the data mining system.
  - Allows the user to interact with the system by providing information to focus the search and perform exploratory DM.
  - Allows user to browse db and dw, evaluate mined patterns and visualize the patterns in different forms.

Data Mining in Business Intelligence



Data Mining

- **Data to be mined**
  - Database data (**e**xtended-relational, object-oriented, heterogeneous), data warehouse, transactional data,
  - Stream, spatial temporal, 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,
  - Descriptive : Characterize properties of the data in a target data set
  - Predictive data mining : Perform induction on the current data in order to make predictions.
  - Multiple/integrated functions and mining at multiple levels

Data Mining

- **Techniques utilized**
  - Data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, 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?

- Most basic forms of data for mining are :
  - Relational database, data warehouse, transactional database

Data Mining: On What Kinds of Data?
Relational database

- DBMS consists collection of interrelated data- database and a set software programs to manage and access the data.
- Relational db: Collection of interrelated data stored in tables with set of attributes and tuples.
- Eg: **AllElectronics** store having branches: customer, item employee and branch

Cust_ID	Name	Address	Age	Income	Occupation	Credit_info
123	M.Kannan	123, south st,	34	34000	Manager	-----

Data Mining: On What Kinds of Data?
Relational database

- Relational database can be accessed by queries using SQL.
- Analyze the database using **aggregate functions.**
- Mining relational databases results in searching for **trends or data patterns.**

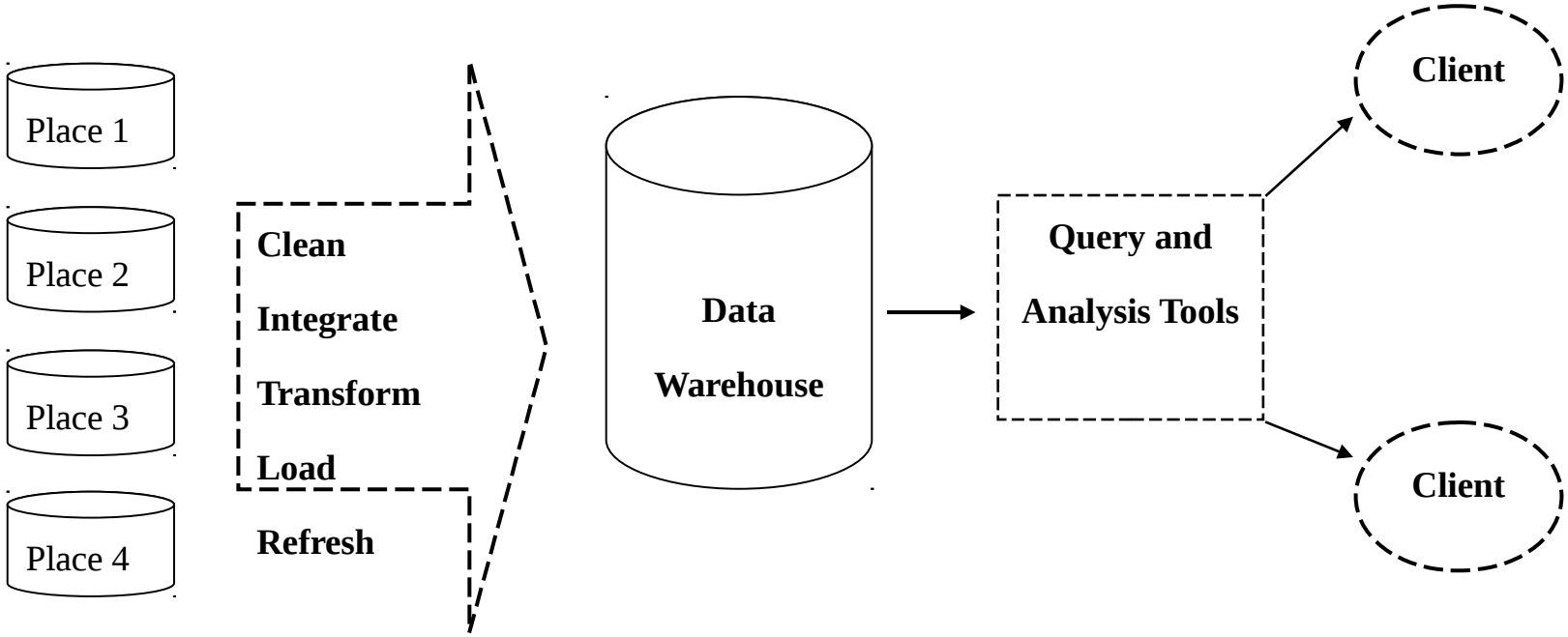
Eg: Customer data can be analyzed by data mining systems to **predict the credit card risk** of new customers based on their income,age and previous credit card information.

- Relational databases are richest information helps in study of DM

Data Mining: On What Kinds of Data? Data Warehouse (DW)

- It is repository of information collected from multiple sources stored under unified schema and residing in a single site.
- Constructed via a process of data cleaning, integration, transformation, loading and periodic refreshening.
- Data are stored to provide information from historical perspective and are typically summarized.
- DW is Modeled by multidimensional data structure called **data cube**

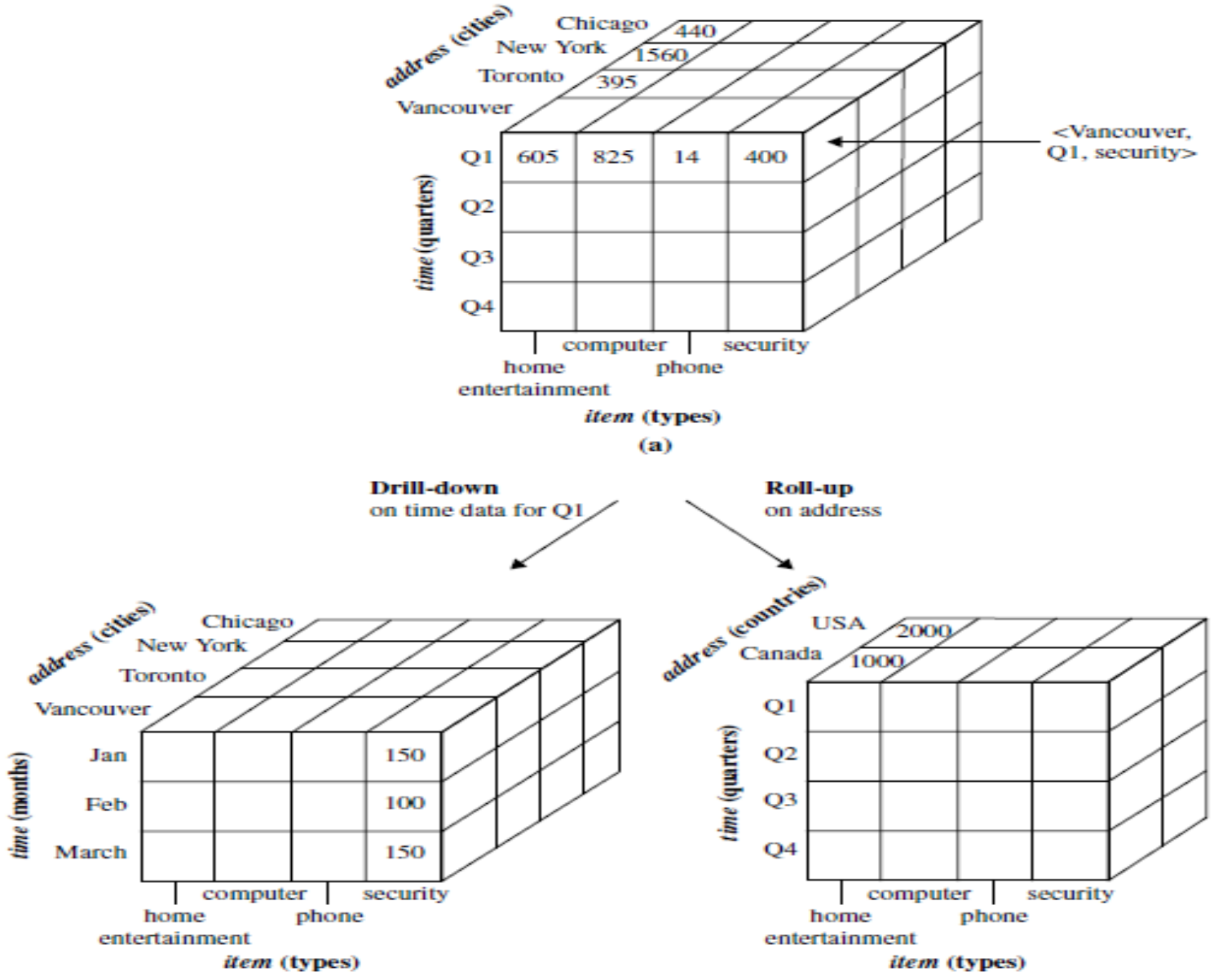
On What Kinds of Data?- Data Warehouse (DW)



- Summarized data of the transaction either per item type for each store or for each sales region.

Data Mining: On What Kinds of Data? Data Warehouse (DW)

- **Data cube**
  - helps in fast access of summarized data.
  - Each dimension corresponds to set of attribute or an attribute.
  - Each cell stores value of aggregate measure.
- Provide support to Online Analytical processing (OLAP) by providing multidimensional data views and summarized data.
- OLAP allow data to be presented in different abstraction levels.
- OLAP accommodate different user view points using operations drill-down and roll-up


address (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
time (quarters)
<Vancouver, Q1, security>
computer
security
home entertainment
phone
item (types)
(a)
Drill-down
on time data for Q1
Roll-up
on address
address (cities)
Chicago
New York
Toronto
Vancouver
Jan
Feb
March
150
100
150
time (months)
computer
security
home entertainment
phone
item (types)
address (countries)
USA
Canada
2000
1000
Q1
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)
(b)

Data Mining: On What Kinds of Data?

Transactional data

- Transactional database captures a transaction.
- It consists of unique transaction id, list of items making up the transactions.
- Transactions can be stored in a table with one record per transactions.
- Market basket analysis or mining of frequent item sets can be done on transactional data using **association mining algorithms.**
  - Eg: Which items sold well together?

Data Mining: On What Kinds of Data?

Advanced data sets and advanced applications

- **Data streams:** Data flow in and out of an observation platform (or window) dynamically.
  - Network traffic, stock exchange, telecommunication, web click streams video surveillance, and weather or environment monitoring.
  - **Time-series data, sequence data and Temporal databases:**
    - historical records, stock exchange data and time series and biological sequence data

Data Mining: On What Kinds of Data?

- **Advanced data sets and advanced applications**
  - Spatial Databases and Spatiotemporal Databases: Spatial database contain spatial-related information.
  - Eg: maps, geographical database and satellite image database
- **Engineering design data:** design of buildings and components or integrated circuits
- **Multimedia database:** text,image,videos and audio data
- **Graph and networked data**: social and information networks
- **The World-Wide Web:** Information from the web

Applications –Data mining

Various knowledge can be mined :

- Eg: stock exchange data can be mined that help us to plan investment strategies.
- Computer network data can be mined to detect intrusions based on the anomaly of message flows.
- By text mining we can identify the evolution of hot topics in the field.
- By mining customer comments on products we can access customer sentiments .
- By mining video data of a hockey game we can detect video sequences corresponding to goals

What kind of pattern to be mined?

- Number of data mining functionalities are available that specifies the kind of patterns to be found in data mining tasks.
- **Data Mining is generally divided into two tasks.**

1. Predictive tasks
2. Descriptive tasks

- **Descriptive Tasks:** Derive patterns which summarizes the underlying relationship between data.
- **Predictive tasks:** Perform induction on the current data in order to make predictions

Data Mining Functions: (1) Generalization

Multidimensional class/concept description: Characterization and discrimination

- Data entries can be associated with **class /concepts.**
- Describe individual classes and concepts in summarized, concise, and precise terms.
- Such descriptions of a class or concept are called class/concept description

Data Mining Functions: (1) Generalization

Data characterization: Summarizing the general characteristics or features of the target class.

Eg: Characteristics of software products with sales increased by 10% in the previous year.

Methods: Statistical measures, plots, multidimensional data cube based OLAP operation.

Output: Pie-charts, bar charts, curves, multidimensional cubes and tables, generalized relations and rules

Data Mining Functions: (1) Generalization

- **Data discrimination:**
  - Comparing target class with one or set of comparative classes
  - Output helps to distinguish target class with comparative class.
- Eg : Compare the characteristics of products that has increased the sale by 10% against product with decreased in sale by 30% in the same period
- **Methods**: same as characterization
- **Output:** same as characterization, Discrimination descriptions expressed in rule form are referred to as discriminant rules.

Data Mining Functions: (2) Pattern Discovery

- **Frequent patterns:** Patterns occurs frequently in data
- Different kinds of frequent patterns: frequent item sets, frequent subsequences and substructures.
- **Frequent item sets:** Sets of items that often appear together. Eg: milk and bread.
- **Frequent subsequences:** Frequently occurring subsequence (pattern the customers follow in buying) Eg: laptop, digital cameras and memory card.
- **Frequent substructures:** Substructure refers to different structural forms. if a substructure occurs frequently it is called as structured pattern.
- **Mining different pattern leads to the discovery of interesting associations and correlations within data**

Data Mining Functions: (2) Pattern Discovery

- Association Analysis: Helps us to know frequently purchased together itemsets.
- **Single-dimensional association:**
  - buys(T, “computer”) => buys(T, “software”) [support = 1%, confidence = 75%]
- **Multi-dimensional association:**
  - age(X, “20..29”) & income(X, “20..29K”) => buys(X, “PC”) [support = 2%, confidence = 60%]
- Association rules are discarded as uninteresting if they do not satisfy both a min support threshold and min confidence threshold

Data Mining Functions: (3) Classification

- **Classification and label prediction**
  - Process of finding a model that describes and distinguishes classes or concepts.
  - Classification predicts categorical labels.
  - Given data set is divided into training and testing set
  - Construct models (functions) based on the features of the training set
  - Model used to predict unknown class labels of test set.

  Ex. Classify the customers as class “A”, “B”, “C” based on the age and income.

Data Mining Functions: (3) Classification

- **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, …

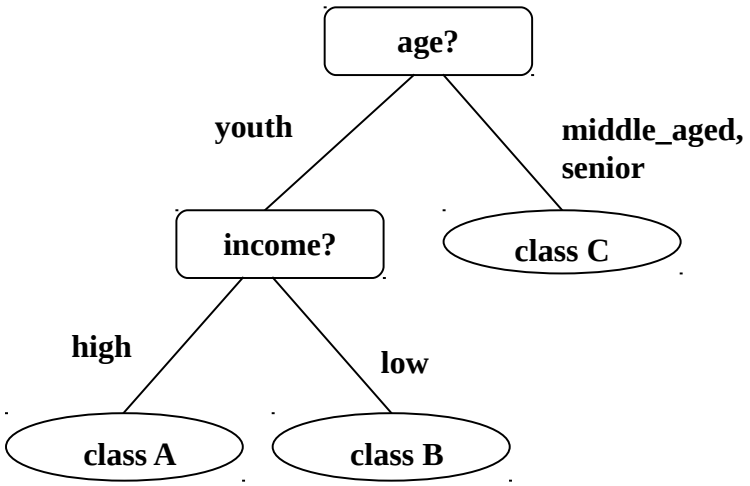
Data Mining Functions: (3) Classification
Decision trees & Rule based

Age(X,"youth") AND income(X,"high") ==> class(X,"A")

Age(X,"youth") AND income(X,"low") ==> class(X,"B")

Age(X,"middle_aged") ==> class(X,"C")

Age(X,"senior") ==> class(X,"C")



Data Mining Functions: (3) Classification

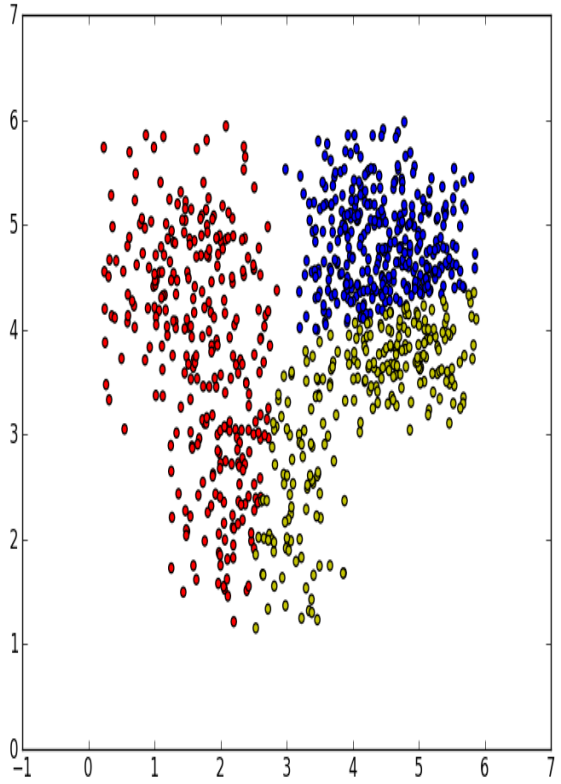
- **Classification Methods:**
  - A decision tree is a flow-chart-like tree structure, where
    - each node denotes a test on an attribute value,
    - each branch represents an outcome of the test,
    - tree leaves represent classes or class distributions.
  - Decision trees can easily be converted to classification rules.

Data Mining Functions: (3) Classification

- **Prediction :** Used to predict missing or unavailable numerical data values rather than class labels.
  - Regression analysis is a statistical methodology that is most often used for numeric prediction.
  - Eg: To predict the amount of revenue each item will generate during the upcoming sale based on the previous sale data.

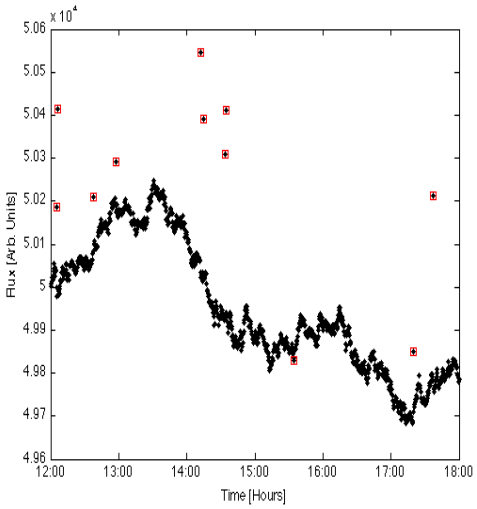
Data Mining Functions: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Clustering used to generate class labels for group of data.
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Each cluster can be viewed as class of objects.
- Eg: Cluster the customer data with respect to customer locations in city



Data Mining Functions: (6) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior or model of the data
  - Noise or exception?— Discard outliers
  - Rare objects are more interesting than the regular ones
  - Analysis of outlier data or anomaly mining.
  - Methods: statistical tests, distance measures, density based methods
  - Useful in fraud detection, rare events analysis



Data Mining Functions: (6) Evolution Analysis

- Evolution analysis:
  - Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.