CSL 407
Data Mining & Data Warehouse

CSL407 : (DE) (L-T-P-C: 3-0-2-4)

Syllabus:

Unit 1 : Introduction to Data Mining and Warehousing, real time applications, scope of mining and warehousing for various applications.

Unit 2 : Data Pre-processing- Various techniques like cleaning, integration, transformation, reduction, discretization, visualizations.

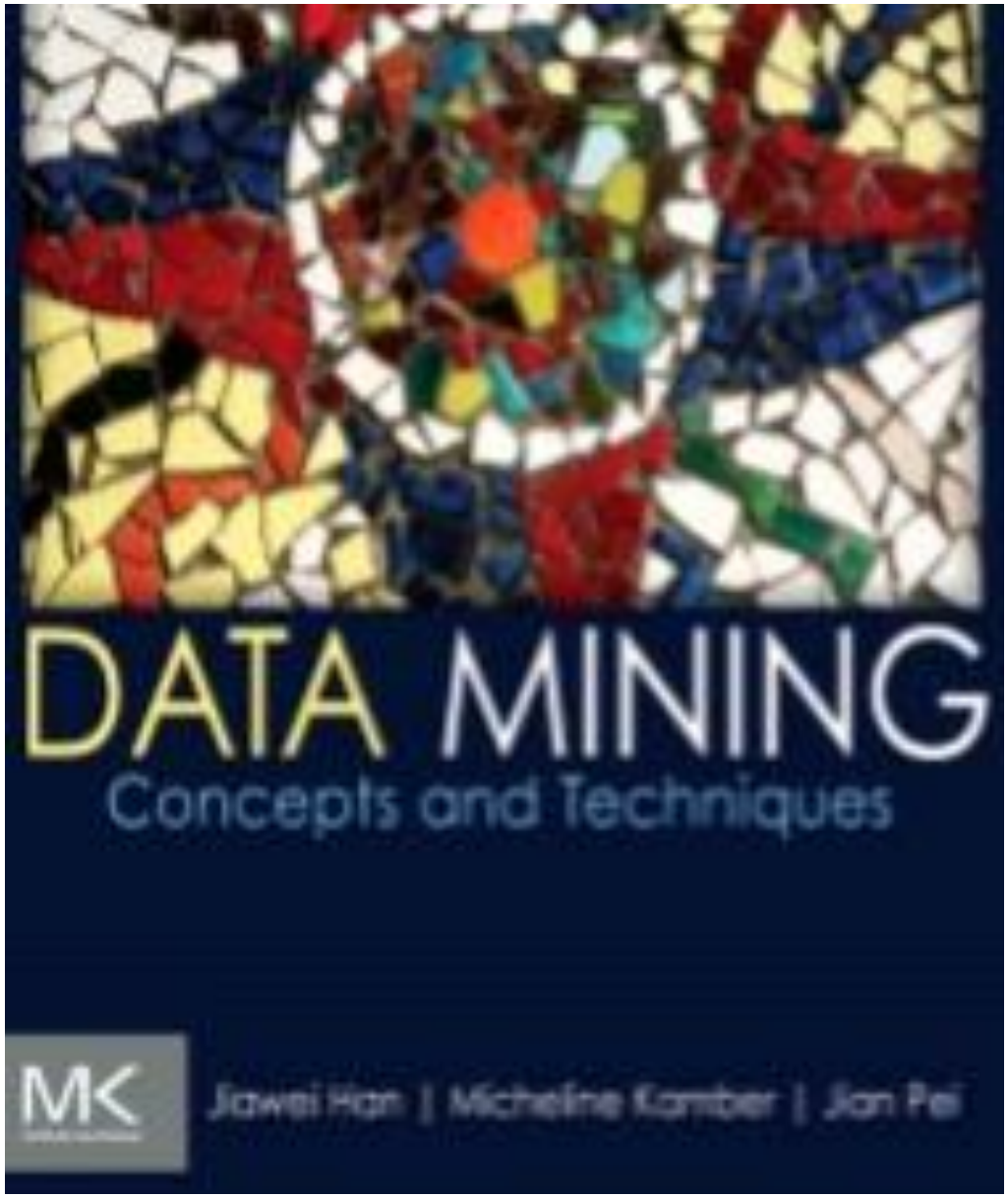
Unit 3 : Data Warehousing - Introduction to Data Mining and warehousing, real time applications, scope of mining and warehousing for various applications. Data ware house Architecture, OLAP, ROLAP and MOLAP , concepts of Fact and dimension table

Unit 4 : Data Mining Tools - Association Rules , A priori Algorithms, FP-trees Algorithms, Constraints and solution.

Unit 5: Data Mining Tools - Cluster Analysis – Paradigms , DBSCAN , Cluster algorithms.

Unit 6: Mining Tools - Decision Trees and applications.

Text Book


DATA MINING
Concepts and Techniques
MK
Jiawei Han | Micheline Kamber | Jian Pei

Course Outcomes

CO1 -Understand that discovering and extracting knowledge from a massive amount of data is a key problem in many scientific and business disciplines

CO2 - Demonstrate a thorough understanding of data mining and knowledge discovery principles and techniques

CO3 - Apply data exploration and mining techniques in their chosen disciplines.

Assessment

Internal (40)	Descriptive Exams (60)	Total (100)
Lab Evaluations / Assignments (3 or 4 for 30) + Quiz (5) + Viva(5)	MidTerm (25) + EndSem (35)	100

Lab Assignments (Python)

- **NumPy**: Base n-dimensional array package
- **SciPy**: Fundamental library for scientific computing
- **Matplotlib**: Comprehensive 2D/3D plotting
- **IPython**: Enhanced interactive console
- **Sympy**: Symbolic mathematics
- **Pandas**: Data structures and analysis

Why Data Mining? : Natural Evolution

- **Vast amount of data are collected daily. Analyzing such data is an important need.**
- **To discover the knowledge from the data, data mining tools can provides a great aid.**
- 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

- This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools.

Major source of Data Generators

- Telecommunication Network Industry
- Medical and Health industry
- Web Search Results
- Communities and Social Media

- This explosively growing, widely available, and gigantic body of data makes our time truly the **data age**.

Why Data Mining?

- Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge.
- "Necessity is the mother of invention"—**Data mining—Automated analysis of massive data sets**

Evolution of Sciences

- **Before 1600, empirical science** (Observations, Experiences)
- **1600-1950s, theoretical science** (Calculated based on theory rather than only Experiences or Observations)
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- **1950s-1990s, computational science** (Based on mathematical Computations may be using computers and other methods)
  - 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 (data-driven science)** (Collection and analysis of data)
  - 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!

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

Potential Applications

- **Data analysis and decision support**
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

- **Other Applications**
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining

What is Data Mining

- Data Mining is a set of activities used to find new, hidden, or unexpected patterns in data.
- **Patterns represents Information.**
- Using this information, the **data scientists** can often provide answers to questions that **decision makers** had previously not thought to ask.

Data Mining Applications

- Data mining: A young discipline with broad and diverse applications
- Some application domains (briefly discussed here)
  - Data Mining for Financial data analysis
  - Data Mining for Retail and Telecommunication Industries
  - Data Mining in Science and Engineering
  - Data Mining for Intrusion Detection and Prevention
  - Data Mining and Recommender Systems

APPLICATIONS OF DATA MINING

- **Data mining applications in banking / finance:**

There are numerous fields in which data mining can be used like in financial and banking sector for credit analysis, to detect fraudulent transactions, customer segmentation and profitability, optimizing stocks portfolios, predicting payment default, ranking investments, marketing, high risk loan applicants, cash management & forecasting operations, and most profitable credit card customers & cross selling.

APPLICATIONS OF DATA MINING

- **Data Mining for the Retail Industry:**

Retail industry assemble huge amount of data related to sales and customer history of shopping. Retail data mining helps in analyzing client behaviour, client patterns of shopping and trends which increases the quality of client service, enhance things consumption ratios, design more effective goods transportations and distribution policies achieve better customer retention and satisfaction and to minimize the cost of business.

APPLICATIONS OF DATA MINING

- **Data mining applications in sales/ marketing:**

---Data mining is the process of extracting unknown patterns from database which help in planning, organizing, managing and launching new market in a cost effective way.

---Data mining plays an important role in Market Basket Analysis. It gives information relevant to item sets that are purchased together, their sequence and when they were bought.

---This information helps business encouragement and to make it most profitable.

APPLICATIONS OF DATA MINING

- **Data mining applications in Health Care and Insurance:**

Insurance industry growth is completely depends on the ability of transforming data into information regarding customers, competitors and its market.

The insurance industries have implemented the Data Mining successfully and have achieved tremendous competitive advantages.

The data mining applications in insurance industry can be used in the form that, data mining is applied in claims analysis such as identifying the medical procedures which are claimed together. Data mining enables to forecasts the potential customers who will buy new schemes. This data mining also proactive insurance companies to detect risky customer's behaviour patterns. Data mining also helps in detecting fraudulent behaviour.

APPLICATIONS OF DATA MINING

- **Data mining for the Telecommunications industry:**

Telecommunication industries generally generate and store large amount of high quality data, having a very huge customer base, and operate in rapidly changing and highly competitive environment.

Telecommunication companies use data mining to enhance their marketing efforts to detect fraud and to betterment of their telecommunication networks.

APPLICATIONS OF DATA MINING

- **Data Mining Application in Higher Education:**

Data mining can be effectively used to address students and alumni challenges. Data mining facilitate organizations to use their current reporting capabilities to uncover and understand hidden patterns in huge databases. These patterns are then built into data mining models and used to predict individual behaviour accurately.

As a result of their insight, institutions are able to allocate resources and staff efficiently. This data mining can provide an entity the information necessary to take action before a student drops out, or to efficiently allocate resource with an accurate estimate of how many students will take a particular course.

APPLICATIONS OF DATA MINING

- **Intrusion Detection** techniques using data mining have attracted more and more interests in recent years. Data mining techniques used for intrusion detection are frequent modalities for mining, classification, clustering and mining data streams etc. Fields where data mining technology can be applied for intrusion detection are development of data mining algorithms for intrusion detection, aggregation to help select and build discriminating attributes, Association and Correlation analysis, Analysis of stream data, Visualization, Distributed data mining and Querying tools.

APPLICATIONS OF DATA MINING

- Data mining for instruction Detection:

Instructions are the set of actions that threatens the availability and integrity of a network resource. Network instruction detection has been considered to be one of the most promising method for defending complex and dynamic intrusion behaviours.

Trends of Data Mining

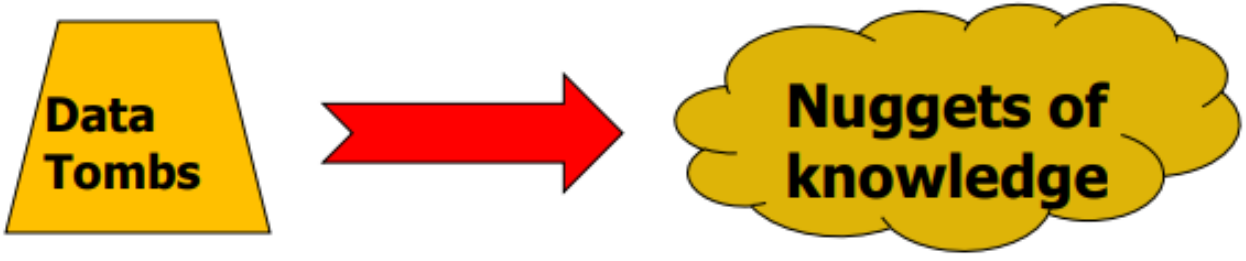
- Application exploration: Dealing with application-specific problems
- Scalable and interactive data mining methods
- Integration of data mining with Web search engines, database systems, data warehouse systems and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving objects and cyber-physical systems
- Mining multimedia, text and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Why Data Mining/ Popularity of DM

Summary:

- Abundance of data and data archives are seldom visited.
- Far exceeded human ability for comprehension
- Intuitive decisions are prone to biases and errors, and is extremely time-consuming and costly
- Data mining tools perform data analysis and uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.

Mining tools are cheaper and affordable as compare to before



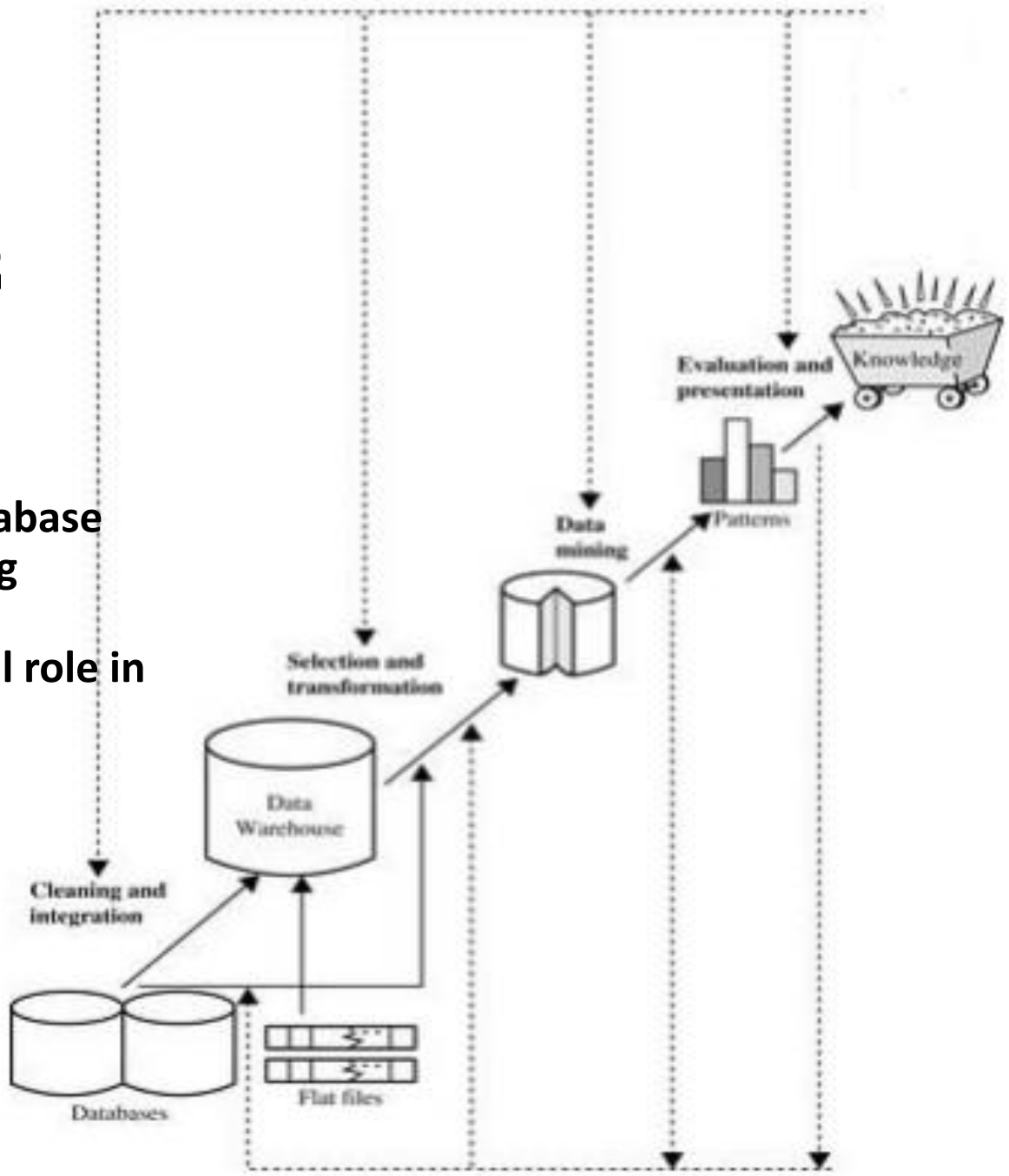
What is Data Mining?

- **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**? (Knowledge Mining from data)
- **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

The goal of Data Mining is the extraction of patterns and knowledge from large amounts of data, **not the extraction of data itself.**

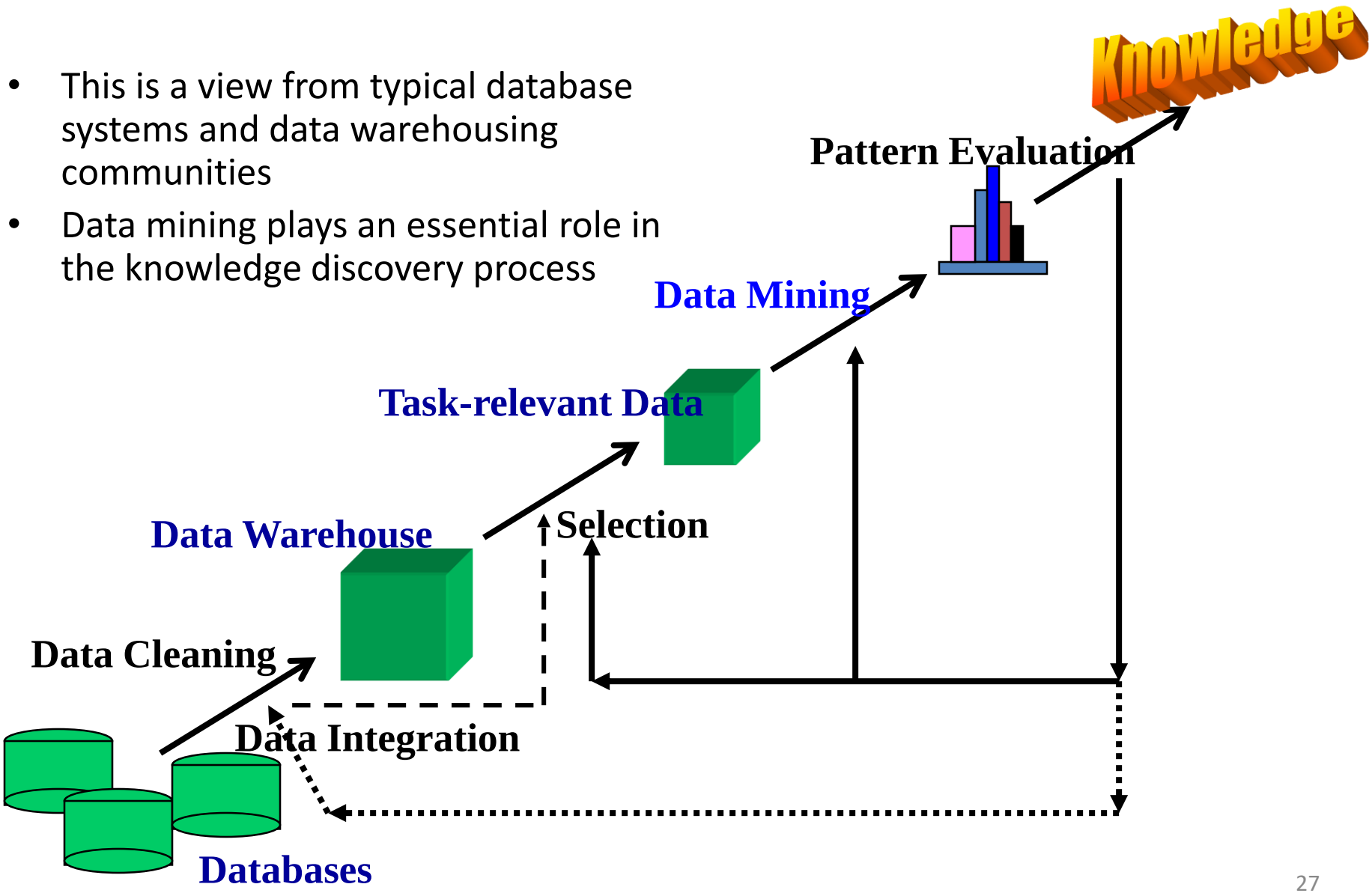
Knowledge Discovery (KDD)Process

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



Knowledge Discovery (KDD) Process

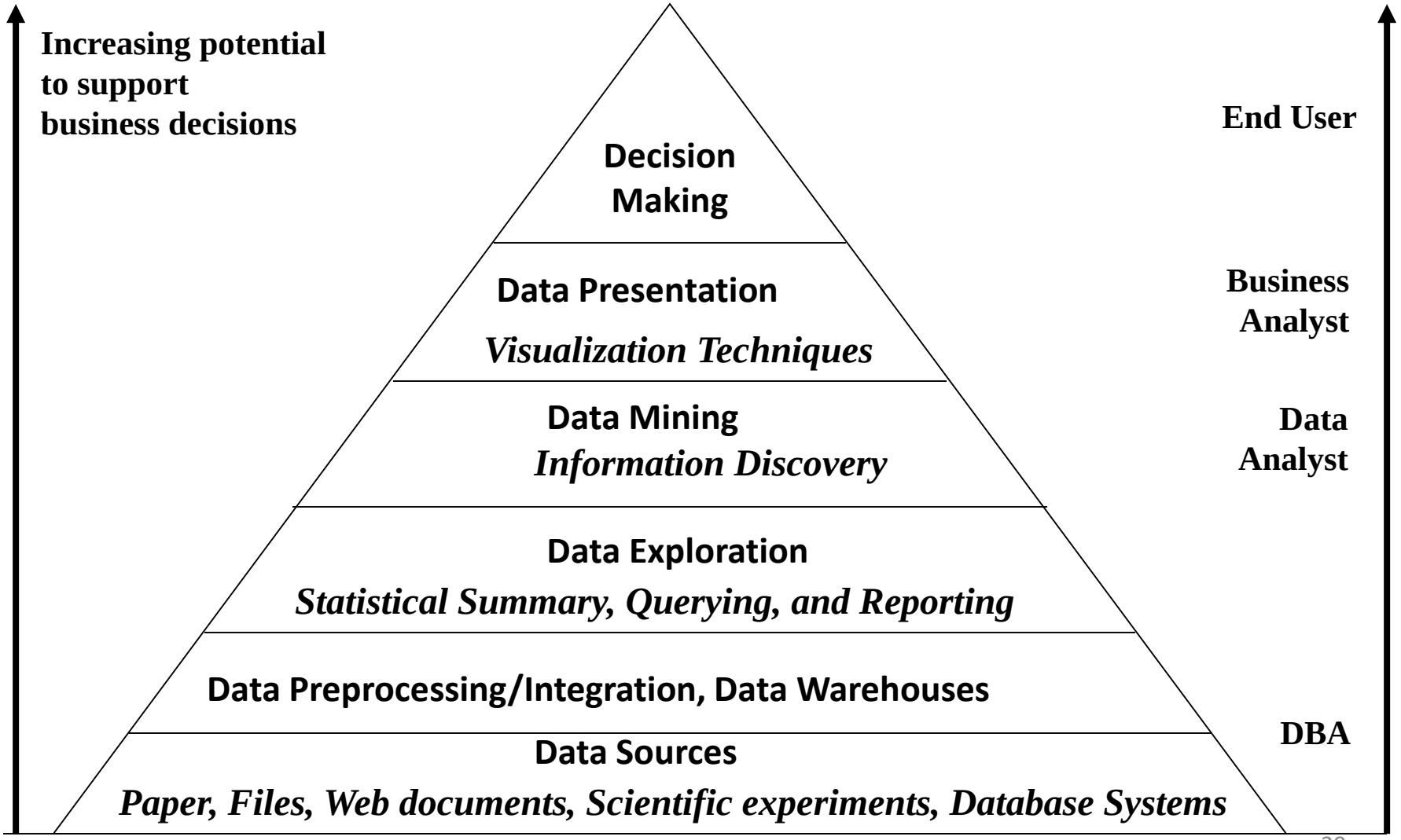
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



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



Data Mining: On What Kinds of Data?

- **Structured and semi-structured data**
  - Relational database/ Object-relational data
  - Data Warehouse,
  - Transactional Database
- **Unstructured data**
  - Data streams and sensor data
  - Text data and web data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Graphs, social networks and information networks
  - Spatial data, spatiotemporal data and multimedia data

Relational Database

- A relational database is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Each tuple in a relational table represents an object identified by unique key and described by a set of attribute values.

Relational Database

- **Four relational tables: *customer, item, employee and branch.***
- **Each relation consists of a set of attributes.**

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

works_at

empl_ID	*branch_ID*
E55	B1
...	...

Relational Database

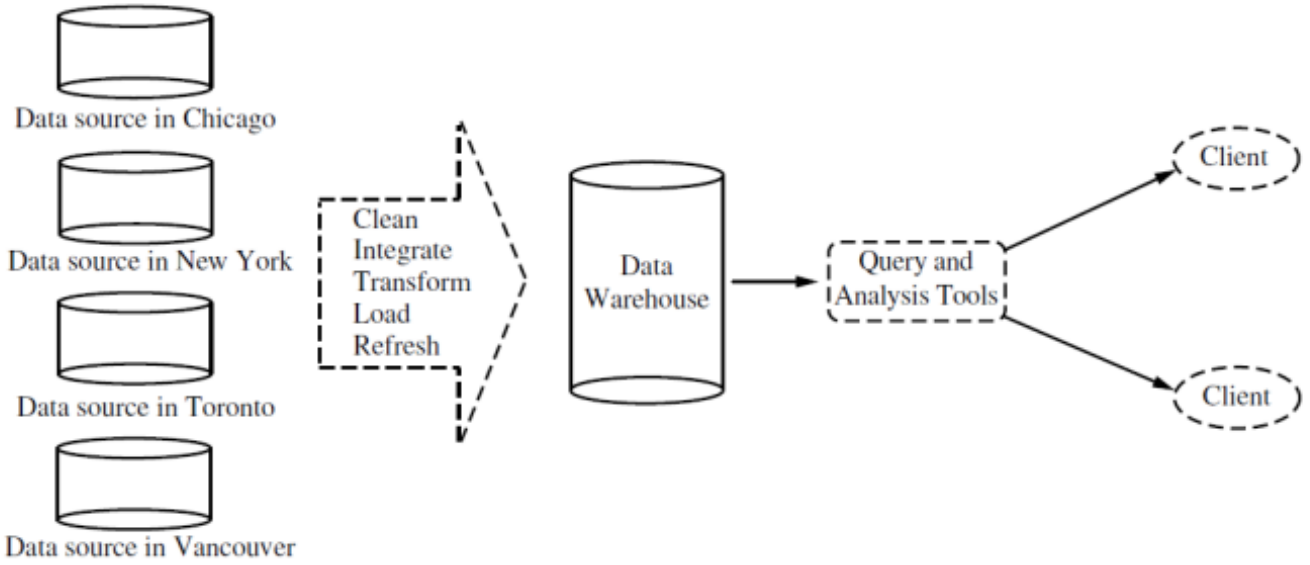
- Show me a list of all items that were sold in the last quarter
- Show me the total sales of the last month, grouped by branch
- Which sales person has the highest amount of sales?
- How many sales transactions occurred in the month of September?

Purpose of relational databases

- The main purpose of a relational database is to store data correctly and retrieve data on demand.
- This type of data processing is sometime called Online Transaction Processing (OLTP).
- Relational databases are passive data repositories in the sense that a query only shows you what is stored in the database, but cannot tell you much about the meaning or trend of the data.

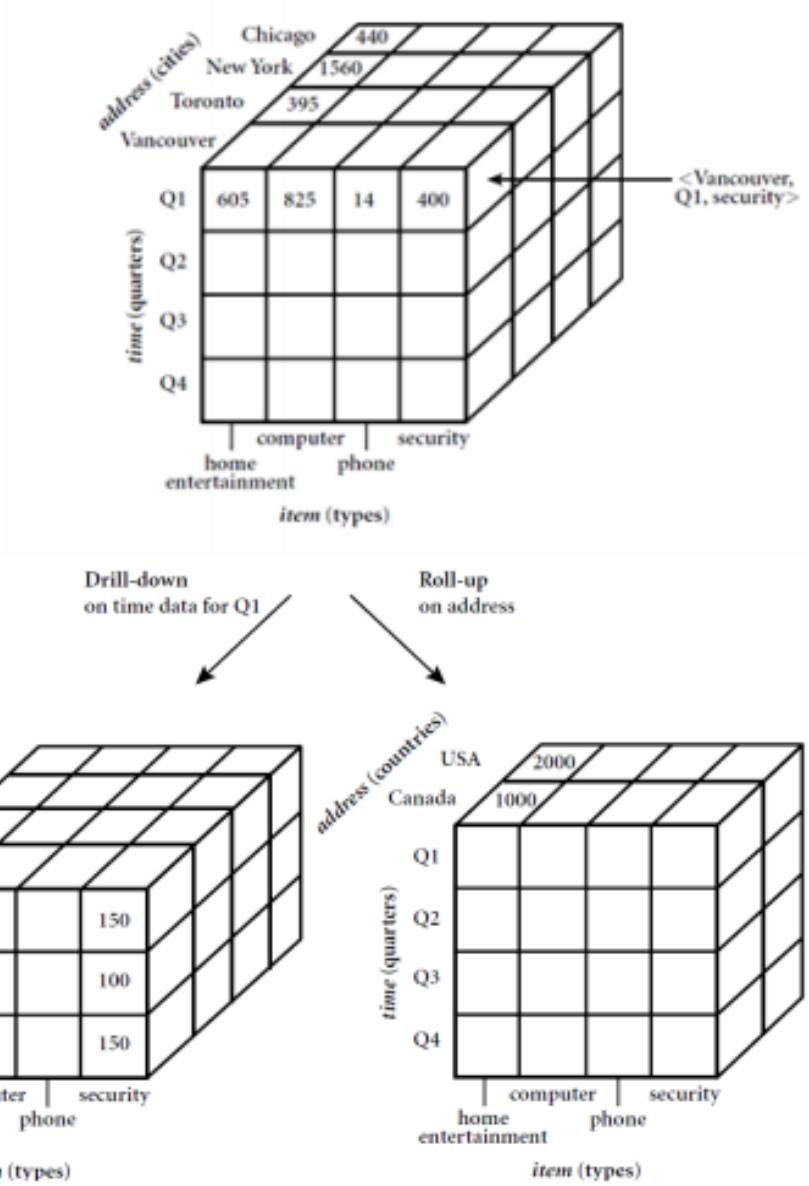
Data from the Data Warehouse

- **A data warehouse is a *repository* of information collected from multiple sources, stored under a *unified schema*, and that usually resides at a single site.**
- **Need is to provide an analysis of the company's sales per item type per branch for the a specified period.**



Data Warehouse

- **The data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region.**



Transactional Database

- A transactional database consists of a file where each record represents a transaction.
- In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
- A transaction typically includes a unique transaction identity number (trans_ID) and a list of the items making up the transaction, such as the items purchased in the transaction.

Ex. of Transactional Database

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Sample Queries: – Show me all the items purchased by ‘X’ – How many transactions include item number ‘Y’? – market basket data analysis: Which items sold well together? (Frequent item set)

Other Kinds of Data

- Besides relational database data, data warehouse data, and transaction data, there are many other kinds of data that have versatile forms and structures and rather different semantic meanings.

Such kinds of data can be seen in many applications:

- Time-related or sequence data (e.g., historical records, stock exchange data, and timeseries and biological sequence data)
- Data streams (e.g., video surveillance and sensor data, which are continuously transmitted), **Spatial data** (e.g., maps), Engineering design data (e.g., the design of buildings, system components, or integrated circuits)

Other Kinds of Data

- Hypertext and Multimedia data (including text, image, video, and audio data), Graph and Networked data (e.g., social and information networks), and the Web (a huge, widely distributed information repository made available by the Internet).
- These applications bring about new challenges, like how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and how to mine patterns that carry rich structures and semantics.

Knowledge Mined from other types of data

- With spatial data, we may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways.
- The relationships among a set of spatial objects can be examined in order to discover which subsets of objects are spatially autocorrected or associated.
- By mining text data, such as literature on data mining from the past ten years, we can identify the evolution of hot topics in the field.

Knowledge Mined from other types of data

- By mining user comments on products (which are often submitted as short text messages), we can assess customer sentiments and understand how well a product is embraced by a market. From multimedia data, we can mine images to identify objects and classify them by assigning semantic labels or tags.
- By mining video data of a hockey game, we can detect video sequences corresponding to goals.
- Web mining can help us learn about the distribution of information on the WWW in general, characterize and classify web pages, and uncover web dynamics and the association and other relationships among different web pages, users, communities, and web-based activities.

Data Science and Analytics

- Data Science is the most important component of analytics.
- It consists of Statistical & Operations research techniques, Machine learning and Deep learning algorithms.
- The objective is to identify the **most appropriate** statistical model/machine learning algorithm that can be used to generate required knowledge from the databases.

Data Science and Analytics (Data Mining)

- Analytics can be grouped into three types:

1. Descriptive analytics (Data Mining).

2. Predictive analytics (Data Mining).

3. Prescriptive analytics (Data Mining).

Descriptive Analytics (Data Mining)

- Descriptive analytics (DA) is the simplest form of analytics that mainly uses simple descriptive statistics, data visualization techniques and queries to understand past data.
- The primary objectives of DA is innovative ways of data summarization.
- It is generally used for understanding the trends in past data which can be useful for generating insights.

Descriptive Analytics (Data Mining)

- Trends obtained through DA can be used to derive actionable items.
- Ex1- Walmart's Chief Information officer wanted to understand the purchasing behaviour of their customer when Hurricane Charley struck the US, these insights were used by Walmart when the next Hurricane struck.
- Ex2 – Dr. John Snow, 1983, using spot map, data visualization tools used to predict the reason of killer cholera is water.
- Tools – **Tableau** and **Qlik Sense** are popular visualization tools

Descriptive Analytics (Data Mining)

- One of the important tools of DA is **Query**.
- In 2014, China Eastern Airline found that a man had booked a first class ticket more than 300 times in a year and cancelled it before its expiry for full refund so that he could eat free food at the airport's VIP lounge.

Predictive Analysis (Data Mining)

- It aims to predict the probability of occurrence of future events.
- Descriptive analytics is used for finding what has happened in the past, while predictive analytics is used for predicting what is likely to happen in future.
- The most frequently used tools are Regression, Logistic regression, Classification, Clustering, Morkov Chains, Rondom Forest, Boosting, and Neural Networks

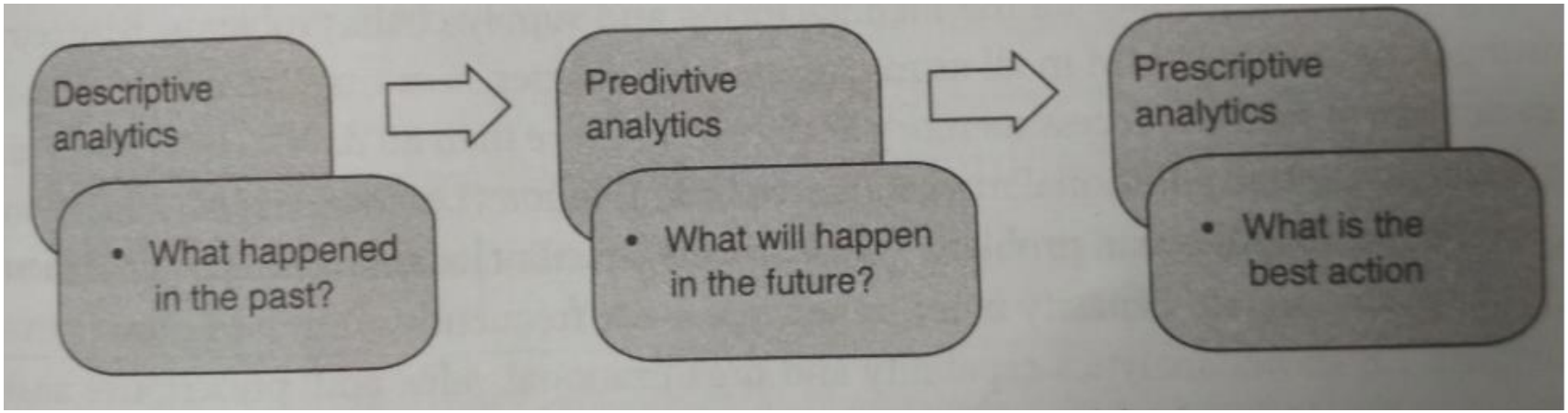
TABLE 1.2 List of predictive analytics applications

Organization	Predictive Analytics Model
Polyphonic HMI	Predicts whether a song will be a hit using machine learning algorithms. Their product 'Hit Song Science' uses mathematical and statistical techniques to predict the success of a song on a scale of 1 to 10 (Anon, 2003).
Okcupid	Predicts which online dating message is likely to get a response from the opposite sex (Siegel, 2013).
Amazon.com	Uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie *et al.*, 2013).
Hewlett Packard (HP)	Developed a flight risk score for its employees to predict who is likely to leave the company (Siegel, 2013).
University of Maryland	Claimed that dreams can predict whether one's spouse will cheat (Whitelocks, 2013).
Flight Caster	Predicts flight delays 6 hours before the airline's alerts.
Netflix	Predicts which movie their customer is likely to watch next (Greene, 2006). 75% of what customer watch at Netflix is from product recommendations (MacKinzie *et al.*, 2013).
Capital One Bank	Predicts the most profitable customer (Davenport, 2007).
Google	Predicted the spread of H1N1 flu using the query terms (Carneiro and Mylonakis, 2010).
Farecast	Developed a model to predict airfare, whether it is likely to increase or decrease, and the amount of increase/decrease.[4]

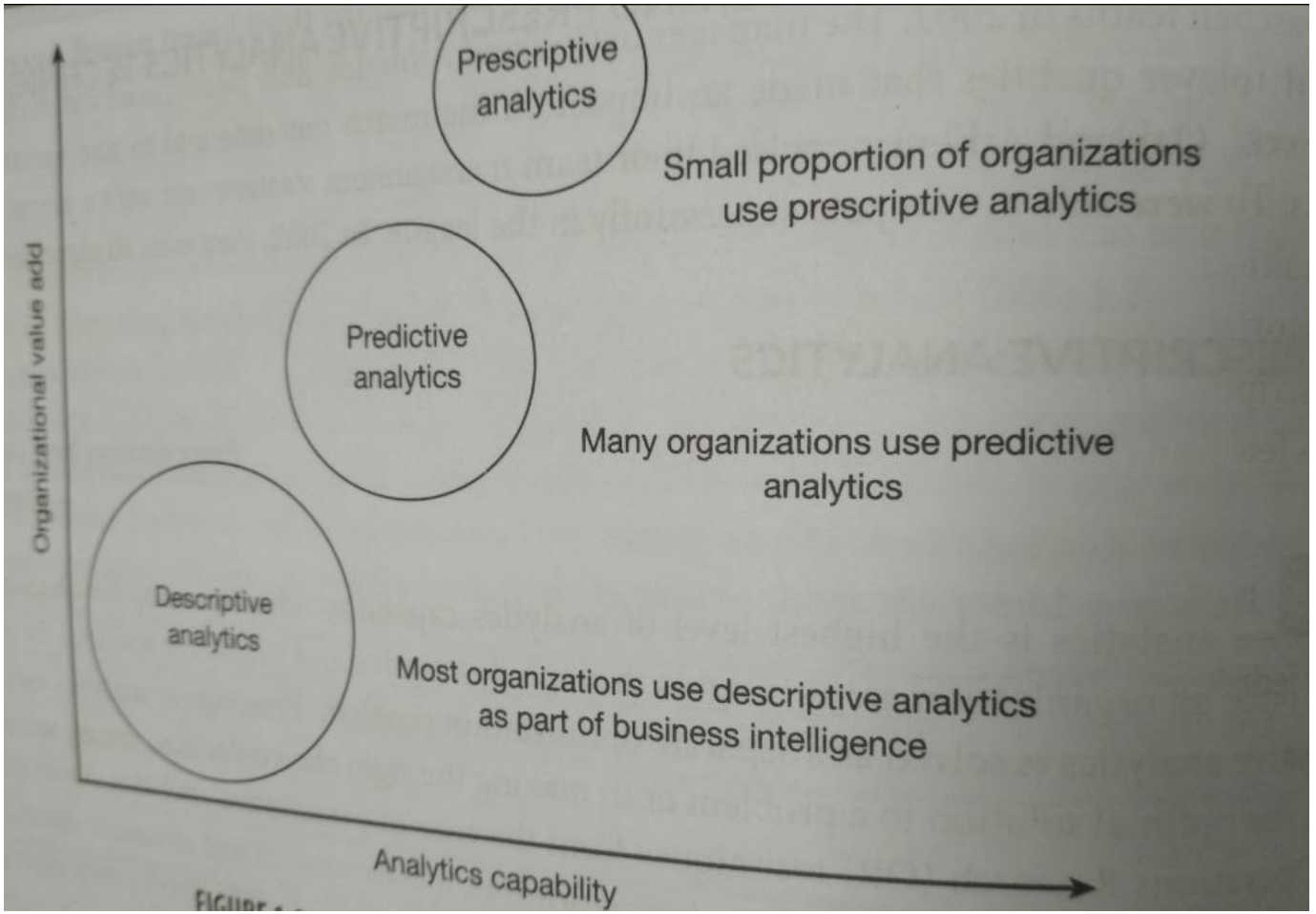
Prescriptive Analytics (Data Mining)

- Prescriptive Analytics is used to choose the optimal actions using Operations Research (OR) techniques after descriptive and predictive analytics brought the insights of an organization.
- The tools used are linear programming, integer programming, multi-criteria decision making models, combinatorial optimizations, non-linear programming, and meta heuristics.

Link between three analytics **(Data Mining)**



Analytic Capability vs Value add



Example

- Ordering a pizza in 2022

Data Mining Functionalities

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

- Data Characterization and Discrimination
- Mining of frequent patterns, associations, and correlations (Association Rule Mining)
- Classification and Regression
- Clustering Analysis and
- Outlier Analysis

Class/Concept Description: Characterization and Discrimination

- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
- Such descriptions of a class or a concept are called class/concept descriptions.
- These descriptions can be derived using-

(1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms (**Ex. Amazon store, classes of items for sale include Computers and Printers and Concept could be Concept of customers include bigSpenders and budgetSpenders**) , or

(2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes) Ex.) , or

(3) both data characterization and discrimination.

Data Characterization

- 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 Characterization

- A customer relationship manager at AllElectronics may order the following data mining task: Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics.
- The result is a general profile of some customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.

----The data mining system should allow the customer relationship manager to drill down on any dimension, such as on occupation to view these customers according to their type of employment.

Data Discrimination

- 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.
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through **database queries**.

Data Discrimination

- 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. **The methods used for data discrimination are similar to those used for data characterization.**

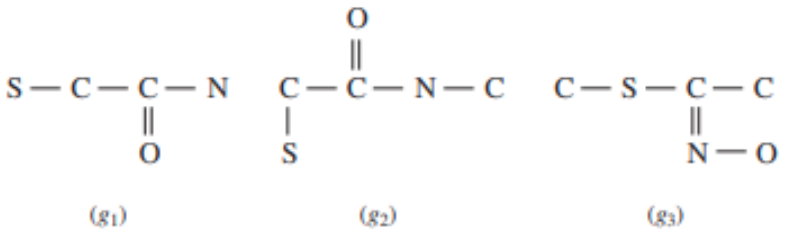
- A customer relationship manager at AllElectronics may want to compare two groups of customers— those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).
- The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.

---Drilling down on a dimension like occupation, or adding a new dimension like income_level, may help to find even more discriminative features between the two classes.

Mining Frequent Patterns, Associations and Correlations

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.
- A frequent itemset typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers.
- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop , followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.
- If a substructure occurs frequently, it is called a (frequent) structured pattern. Many scientific and commercial applications need patterns that are more complicated than frequent itemsets and sequential patterns and require extra effort to discover. Such sophisticated patterns go beyond sets and sequences, toward trees, lattices, graphs, networks, and other complex structures.
- Mining frequent patterns leads to the **discovery of interesting associations and correlations within data.**

Example of Substructure Mining



A sample graph data set.

Classification

- Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is not known.
- **A decision tree or set of classification rules is based on such type of mechanism of classification which can be retrieved for identification of future data.**

- for example one may classify the employee's potential salary on the bases of salary classification of similar employees in the company.

Regression

- Whereas classification predicts **categorical** (discrete, unordered) **labels**, regression models uses **continuous-valued** functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.

Classification and Regression (Ex)

- Suppose as a sales manager of AllElectronics you want to classify a large set of items in the store, based on three kinds of responses to a sales campaign: **good response, mild response and no response**.
- You need to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place_made, type, and category

Classification and Regression (Ex)

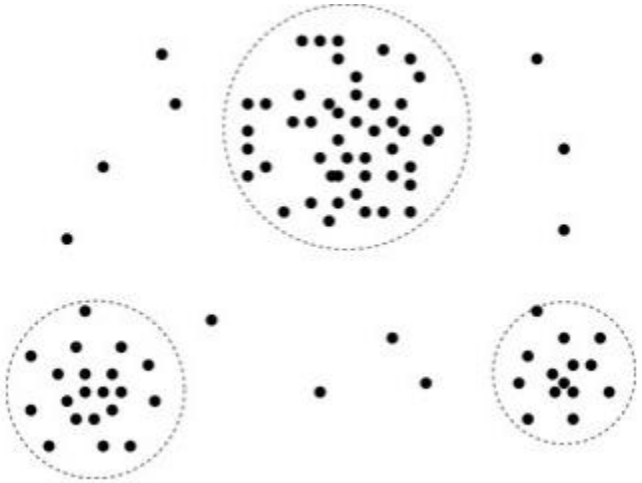
- you would like to **predict the amount of revenue** that each item will generate during an upcoming sale at AllElectronics, based on the previous sales data. This is an example of regression analysis because the regression model constructed, will predict a continuous function (or ordered value.)

Clustering

- Clustering is the process of partitioning a set of object or data in a same group called a cluster.
- These objects are more similar (in some sense or another) to each other than to those in other groups (clusters).
- Clustering is used in many fields, including machine learning, patterns recognition, bioinformatics, image analysis and information retrieval.

Clustering

- Cluster analysis can be performed on AllElectronics customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing



Outer Analysis

- A data set may contain objects that do not comply with the general behaviour or model of the data. These data objects are outliers.
- Deviants, Abnormalities, Discordant and Anomalies are also referred as outliers in data mining and statistics literature.
- The outlier can be diagnosed with the help of various statistical tests, distance model analysis, cluster and regression analysis.
- The analysis of outlier data is referred to as outlier analysis or anomaly mining.

Outer Analysis

- Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

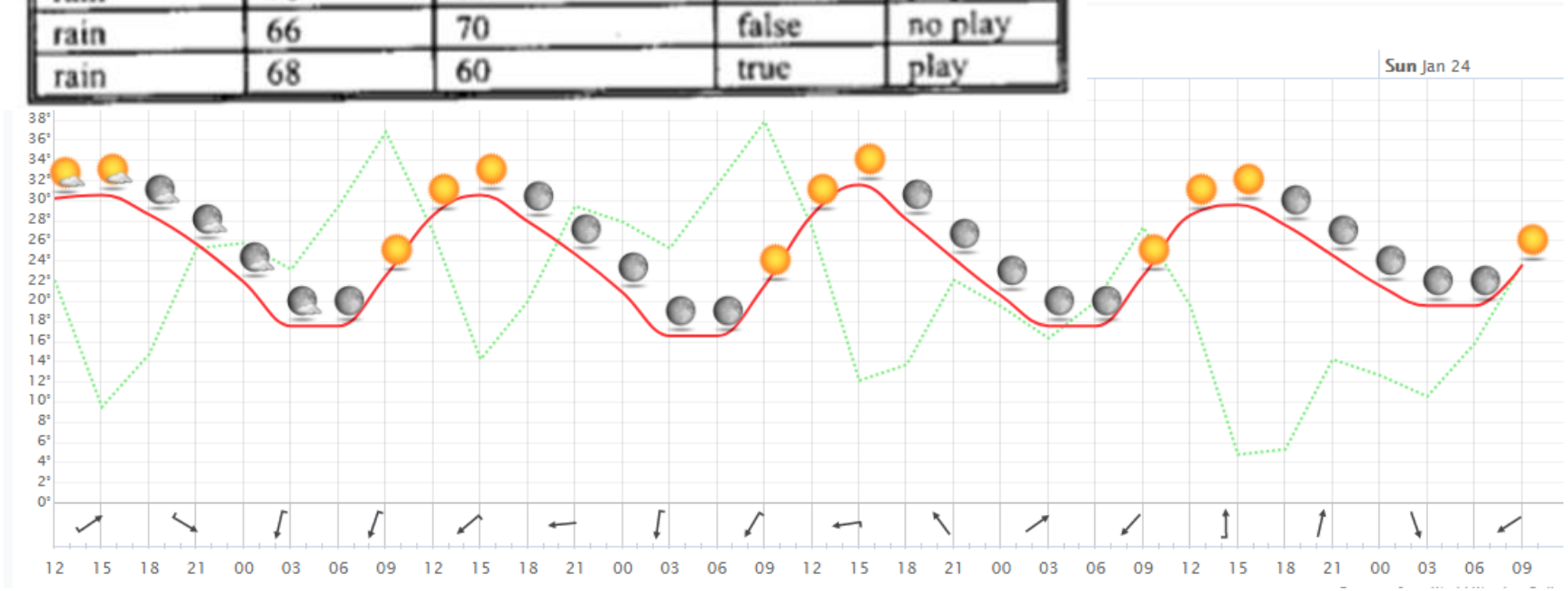
Test Datasets

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Ex. Whether Forecasting

Test Data Set

OUTLOOK	TEMP(F)	HUMIDITY(%)	WINDY	CLASS
sunny	79	90	true	play
sunny	56	70	false	play
sunny	79	75	true	no play
sunny	60	90	true	no play
overcast	88	88	false	no play
overcast	63	75	true	play
overcast	88	95	false	play
rain	78	60	false	play
rain	66	70	false	no play
rain	68	60	true	play



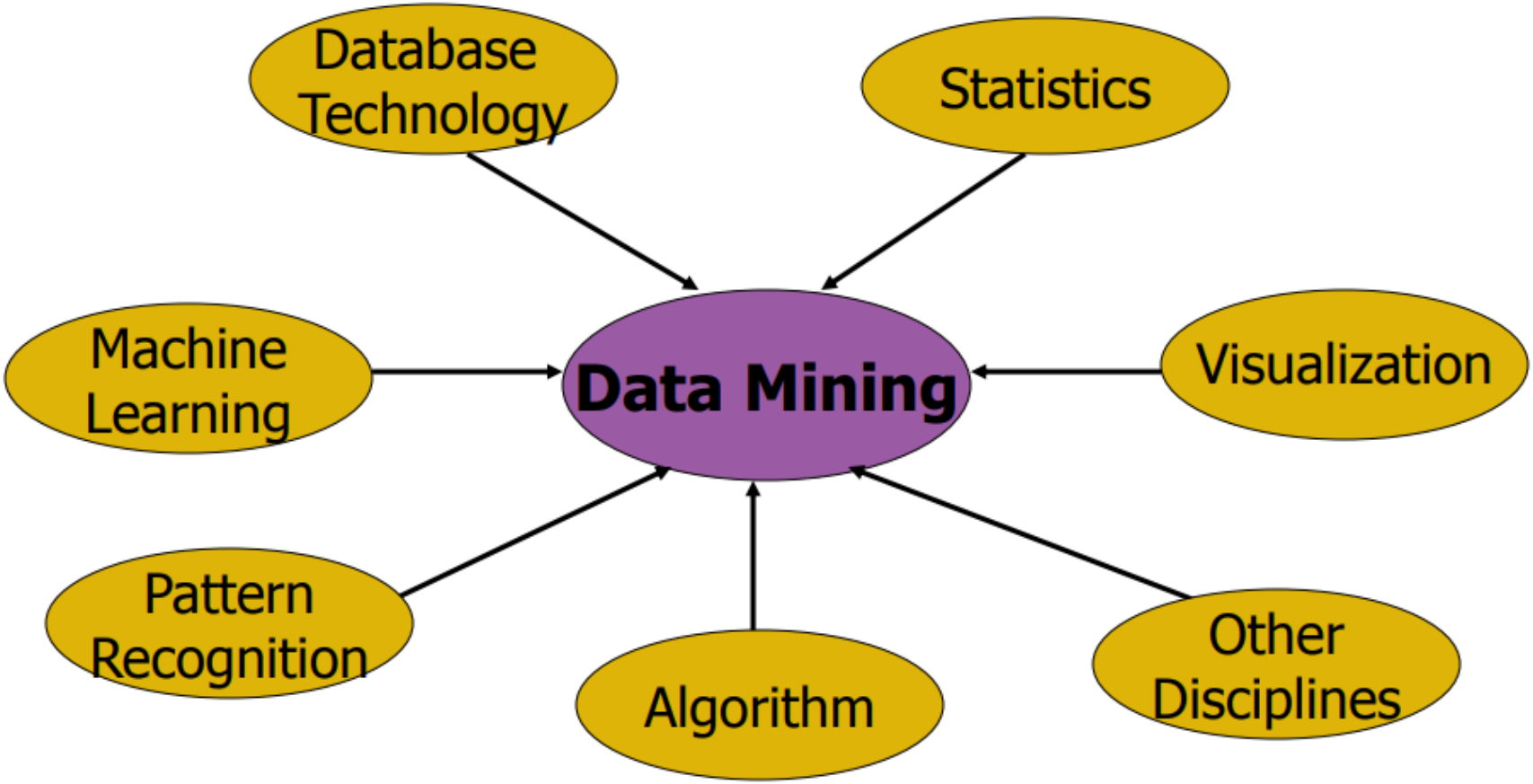
KDD Process: Several Key Steps

- **Learning the application domain**
  - relevant prior knowledge and goals of application
- **Creating a target data set: data selection**
- **Data cleaning and preprocessing: (may take 60% of effort!)**
- **Data reduction and transformation**
  - Find useful features, dimensionality/variable reduction, invariant representation
- **Choosing functions of data mining**
  - summarization, classification, regression, association, clustering
- **Choosing the mining algorithm(s)**
- **Data mining: search for patterns of interest**
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- **Use of discovered knowledge**

Data Mining: Confluence of Multiple Disciplines

- **Tremendous amount of data (terabyte-petabyte)**
- **High-dimensionality and high complexity of data**
  - Structured, un-structured, heterogeneous data
- **Scalable**
- **Data mining involves integration of multiple disciplines:**
  - Machine learning
  - Pattern recognition
  - Statistics
  - Databases
  - Business Intelligence
  - Big data
- **Efficient: Derived knowledge is new, interesting, informative and can be used for sophisticated application (decision making, process control, information management....)**

Data Mining: Confluence of Multiple Disciplines



CHALLENGES IN DATA MINING

- **Scalability & Efficiency** :

Efficiency and scalability are always considered when comparing data mining algorithms. As data amounts continue to multiply, these two factors are especially critical.

Parallel, distributed, and incremental mining algorithms are needed for current society.

Cloud computing and cluster computing, which use computers in a distributed and collaborative way to tackle very large-scale computational tasks

CHALLENGES IN DATA MINING

- **High Dimensional Data and High Data Streams:**

One challenge is to design techniques to control ultra high dimensional classification problems for mining vast, enormous and high dimensional data. Set out-of-memory, parallel and distributed algorithms, algorithm is need to be developed. The traditional data analysis techniques developed for low-dimensional data do not work for high dimensional data.

CHALLENGES IN DATA MINING

- **Complex and Heterogeneous Data:**

Another challenge erupted in these years is emergence of more data complex. A good system must scale the complexity from users. Previous analysis data mining method deals with the data set consisting attribute of similar type i.e. continuous or categorical. Due to increasing role of data mining in different areas, a need is arisen to develop techniques which can handle heterogeneous attributes. Such developed techniques for mining such complex objects ought to have taken care the relationships in data, like temporal and spatial auto-correction, graph connectivity and parent-child relationships between the components in semi-structures text and XML documents.

CHALLENGES IN DATA MINING

- **Data Ownership, Security and Privacy:**

It is a big challenge to find out data for an analysis at one location or to be owned by one location or to be owned by one entity. An automatic data mining in distributed environment can develop serious issues in terms of data privacy or its security. These issues can be addressed by developing of an efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data and further to measure the impact on the modification of data values on discovered pattern's statistical significance.

CHALLENGES IN DATA MINING

- **Data Distribution:**

This challenge in data mining is very important, generally in network problems. This can be addressed by the development of distributed data mining techniques. The key challenges in distributed data mining are: a) **To minimize the amount of communication needed to perform the distributed computation**. b) **To consolidate the data mining results obtains from multiple sources in a efficient manner.** c) **To tackle data security issues.**