Parul®
University

Data Mining and Warehousing (03105430)

Dheeraj Kumar Singh, Assistant Professor
Department of Information Technology

The Course Outline

Chapter 1 : Introduction to data mining (DM)

Chapter 2: Overview and concepts Data Warehousing and Business Intelligence

Chapter 3: Data Warehousing and Online Analytical Processing

Chapter 4: Data Pre-processing

Chapter 5: Mining Frequent Patterns, Associations, and Correlations:

Chapter 6: Classification

Chapter 7: Clustering

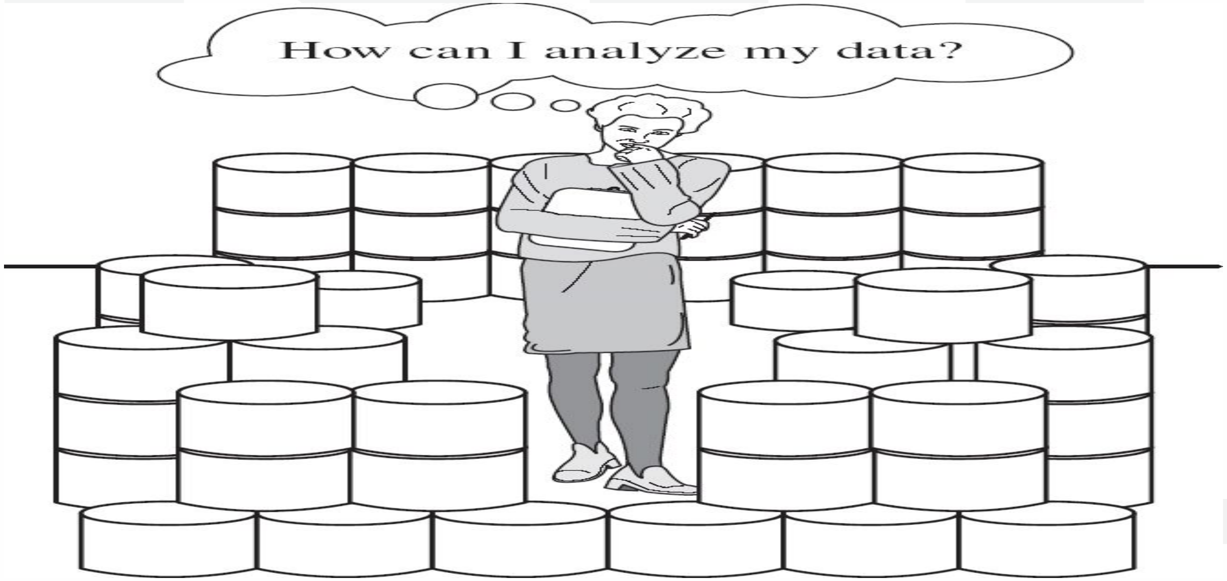
Chapter 8: Applications

CHAPTER-1

Introduction to Data Mining

Introduction of Data Mining

- Drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining
    - Automated analysis of massive data sets



Introduction of Data Mining

- Extraction of implicit, previously unknown and potentially useful information from data

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

Introduction of Data Mining

- Data Mining also known as Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archaeology, data dredging, information harvesting, business intelligence, etc

- As data is growing at very remarkable rate, there comes a need to analyze large, complex and information rich data sets to gain the hidden information. This may result into greater customer satisfaction and remarkable turn over for the firm.

Data vs. Information

Data

- raw facts
- no context
- just numbers and text

Example:

Data: 51007

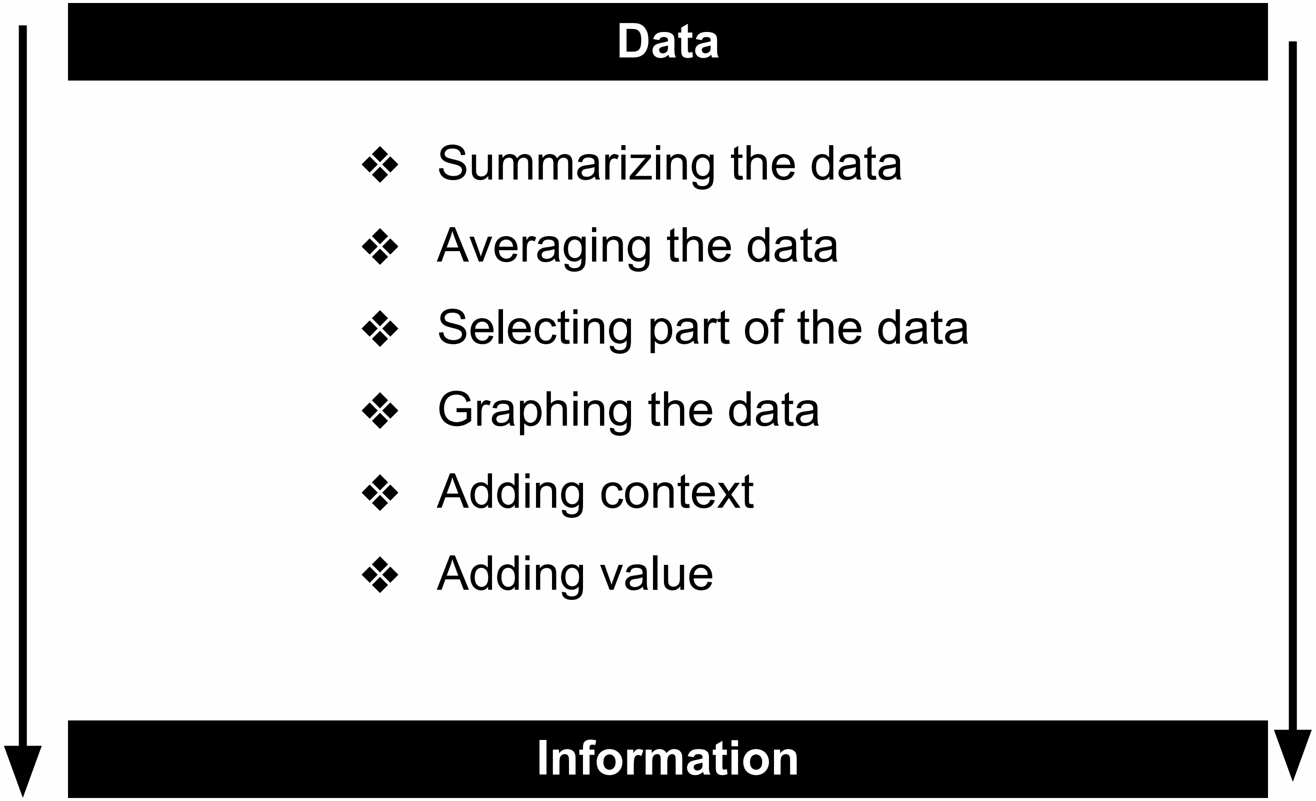
Information

- data with context
- processed data
- value-added to data
  - summarized
  - organized
  - analyzed

Example:

- 5/10/20 Date of your final exam.
- $51,007 you salary.
- 51007 Zip code of any place.

Data ⍰ Information ⍰ Knowledge

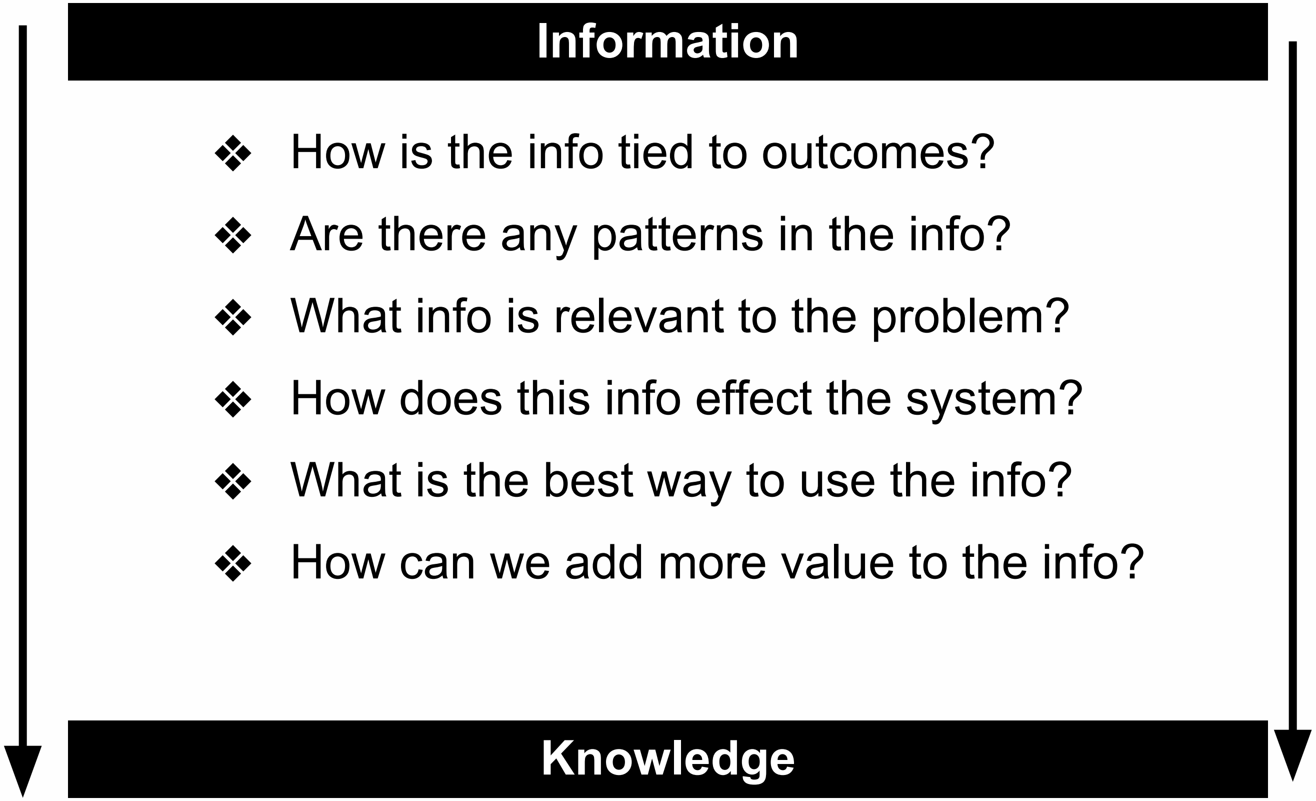


Data

- ❖ Summarizing the data
- ❖ Averaging the data
- ❖ Selecting part of the data
- ❖ Graphing the data
- ❖ Adding context
- ❖ Adding value

Information

Data ⍰ Information ⍰ Knowledge



Information

- ❖ How is the info tied to outcomes?
- ❖ Are there any patterns in the info?
- ❖ What info is relevant to the problem?
- ❖ How does this info effect the system?
- ❖ What is the best way to use the info?
- ❖ How can we add more value to the info?

Knowledge

Why do We Need Mata Mining?

- **Lots of data is being collected and warehoused**

- Web data, e-commerce

- purchases at grocery stores

- Bank/Credit Card transactions

Figure 1.1
E-Commerce

Contd……

- **Computers have become cheaper and more powerful**
- **Competitive Pressure is Strong**

 - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Data Mining Functionality

- **Concept description: Characterization and discrimination**

  - Generalize, summarize, and contrast data characteristics

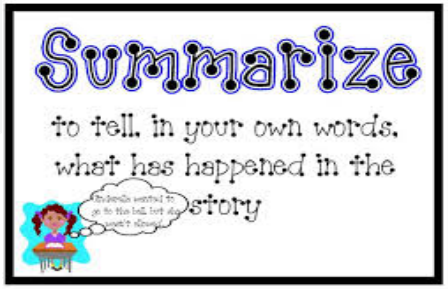
  - Example: dry vs. wet regions

Generalizations
A statement about a group that is true for MOST of the group
HOT
They are peppers, they must be hot

- **Association (correlation and causality) :**

  - Multi-dimensional vs. single-dimensional association

  - age(X, ―20..29) ^ income(X, ―20..29K) ->buys(X, ―PC) [support = 2%, confidence = 60%]



Contd.....

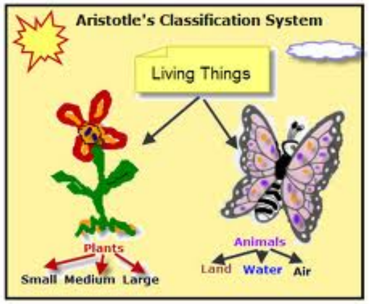
- **Classification and Prediction:**

  - Finding models (functions) that describe and distinguish classes or concepts for future prediction

  - E.g., classify countries based on climate, or classify cars based on gas mileage

  - Presentation: decision-tree, classification rule, neural network

  - Prediction: Predict some unknown or missing numerical values



Contd.....

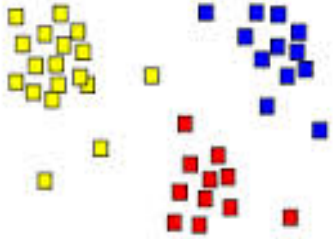
- **Cluster analysis :**

  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns

  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity20

- **Outlier analysis :**

  - Outlier: a data object that does not comply with the general behaviour of the data C



Contd.....

- It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

•**Trend and evolution analysis:**

- Trend and deviation: regression analysis

- Sequential pattern mining, periodicity analysis

- Similarity-based analysis2

•**Other pattern-directed or statistical analyses**

Data Mining Task

- **Data mining is widely divided into two parts:**
  - Predictive Data mining
  - Descriptive Data mining

Data Mining Task

- **Predictive Data mining:**

  - The objective of predictive tasks is to use the values of some variable to predict the values of other variable.

  - Ex: Web mining is used by the online marketers to predict the purchase by online user on a website

- **Classification**

  - Used to map data in a predefined groups.

- **Regression**

  - Maps a data item to a real valued prediction variable.

Data Mining Task

- **Clustering**

  - Form a similar data together.

- **Summarization**

  - It is used to map data in a subsets. Link Analysis defines relationships among data.

Data Mining Task

- **Discriptive Data mining:**

  - The objective of descriptive tasks is to find human readable patterns which describes the relationships between data.

Origin of Data Mining System

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

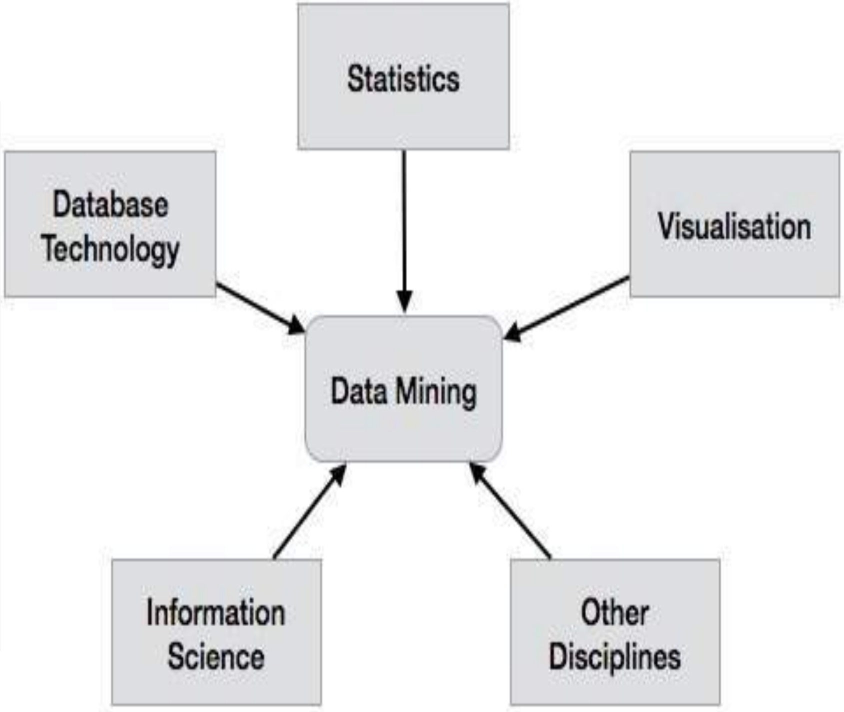


Figure 1.2
Origin of Data
Mining

Origin of Data Mining System (Contd.....)

- Traditional Techniques may be unsuitable due to:

- Enormity of data Statistics/ Machine Learning/ AI Pattern

- High dimensionality Recognition of data

- Heterogeneous, Data Mining distributed nature of data Database systems

Classification of Data Mining System

- **Data mining system classified into four kind of data mining**

  - Databases to be mined

  - Knowledge to be mined

  - Applications adapted

  - Techniques utilized

Classification of Data Mining System (Contd.....)

- **Databases to be mined:**

  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc

- **Knowledge to be mined:**

  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.

  - Multiple/integrated functions and mining at multiple levels⍰ Techniques utilized

Classification of Data Mining System (Contd.....)

- **Techniques utilized**

  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

- **Applications adapted**

  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Architecture of Data Mining System

- **Four kind of data mining architecture**

  - No- Coupling
  - Loose Coupling
  - Semi tight Coupling
  - Tight coupling

No- Coupling

- In this architecture, data mining system doesn’t use any functionality of a database or data warehouse system.
- Data is retrieved from data sources like file system and processed using data mining algorithms which are stored into file system.
- This architecture is considered as a poor architecture for data mining system as it does not take any advantages of database or data warehouse.
- However it is used for simple data mining processes

Loose Coupling

- The loose coupling data mining system uses database or data warehouse for data retrieval.
- In this architecture, data mining system retrieves data from database or data warehouse, processes data using data mining algorithms and stores the result in those systems.
- Loose coupling architecture is for memory-based data mining system which does not require high scalability and high performance.

Semi- tight Coupling

- In semi-tight coupling data mining architecture, it not only links it to database or data warehouse system, but it also uses several features of database or data warehouse systems which perform some data mining tasks like sorting and indexing etc.

- Moreover the intermediate result can also be stored in database or data warehouse system for better performance

Tight Coupling

- In this architecture, database or data warehouse is treated as an information retrieval component.

- Tight-coupling data mining architecture provides scalability, high performance and integrated information.

Architecture of Data Warehouse

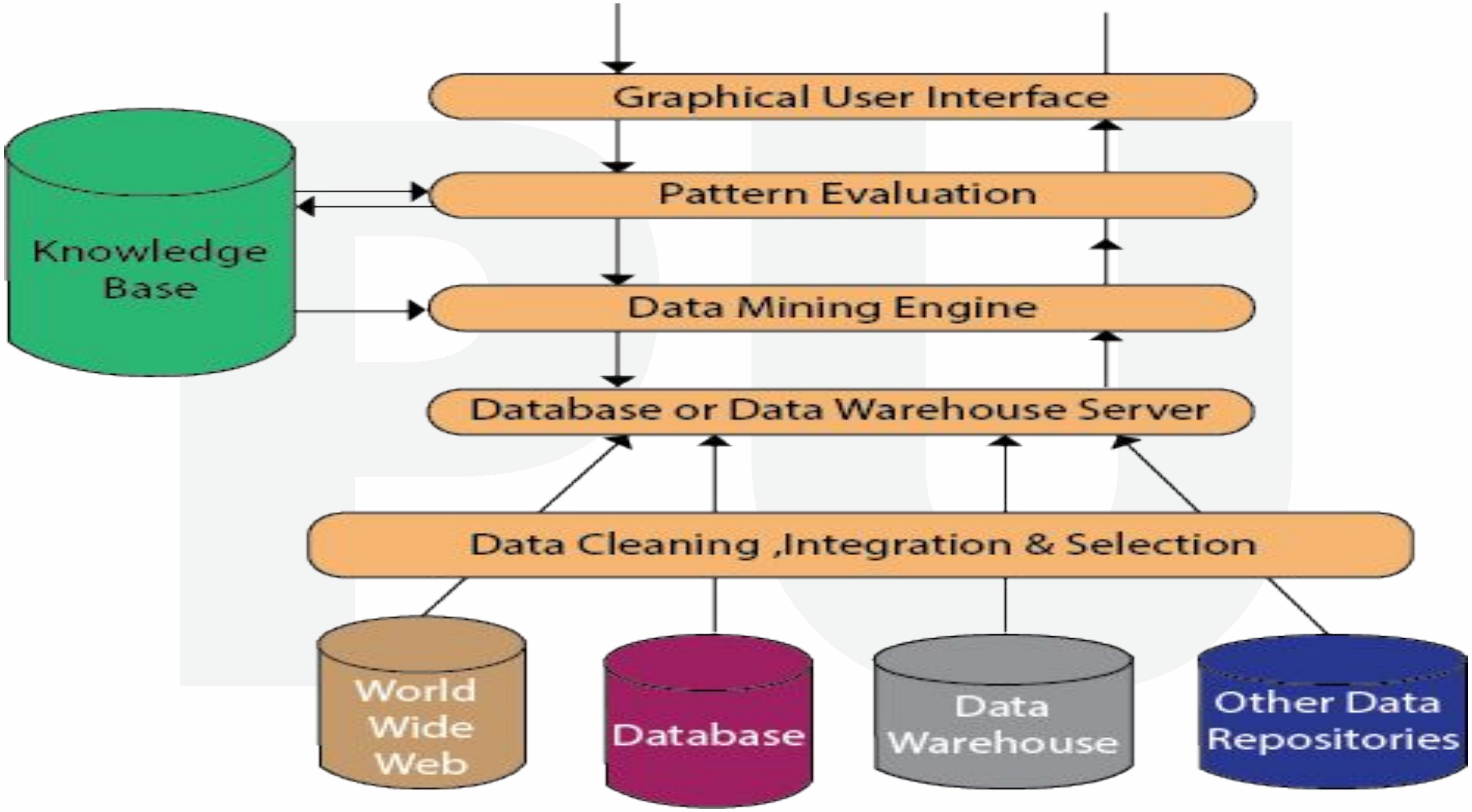


Figure 1.4

Data mining: A KDD Process

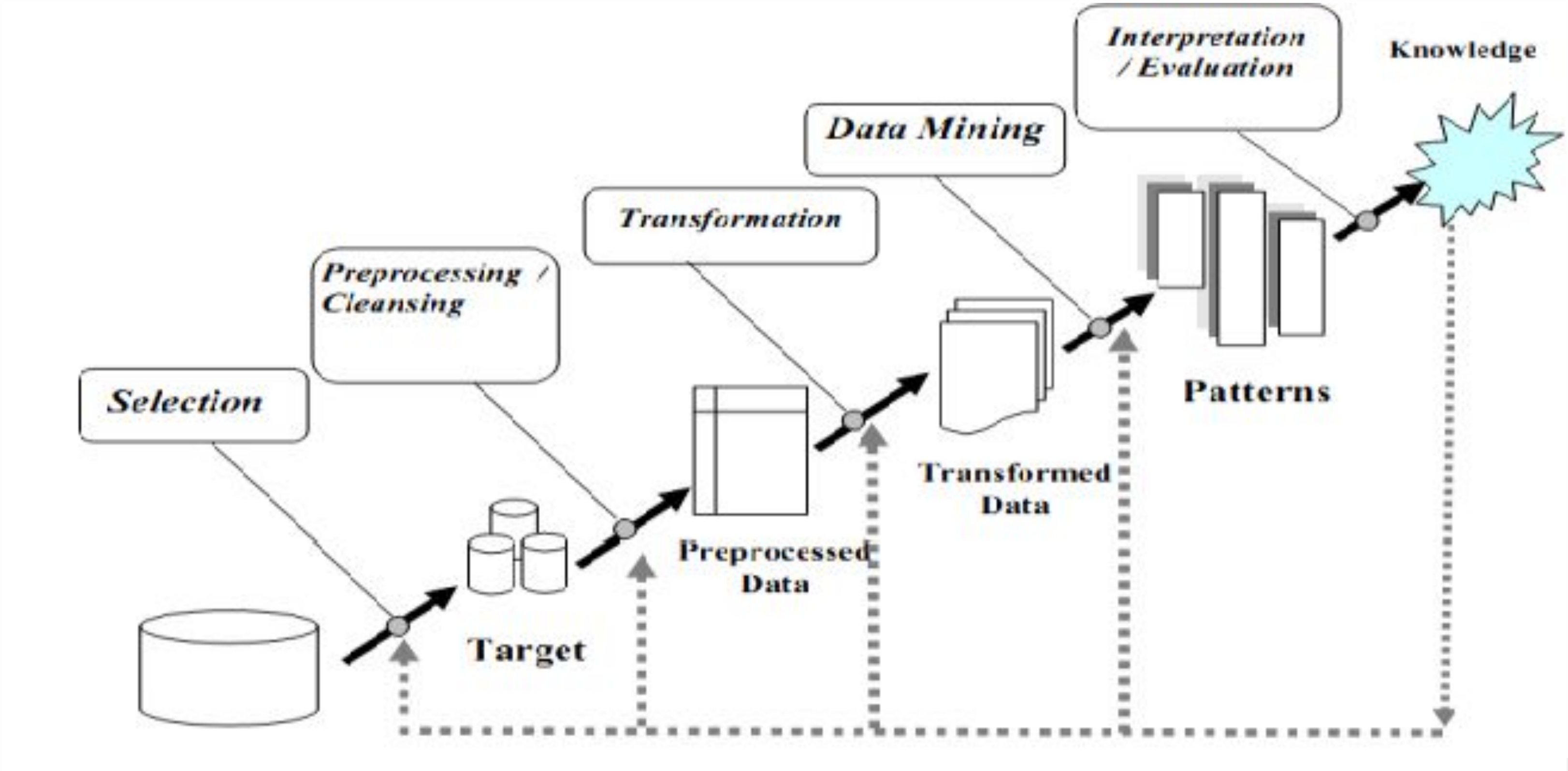


Figure 1.4 A kDD Process

Data mining: A KDD Process (Contd......)

- The KDD process comprises of a few steps leading from raw data collections to some form of new knowledge.
- The iterative process consists of the following steps:

  - Data cleaning
  - Data integration
  - Data selection
  - Data transformation
  - Data mining
  - Pattern evaluation
  - Knowledge representation

Data mining: A KDD Process (Contd.....)

- **Data cleaning**

  - noise data and irrelevant data are removed from the collection

- **Data integration**

  - multiple data sources (heterogeneous) may be combined in a common source

- **Data selection**

  - data relevant to the analysis is decided on and retrieved from the data collection

Data mining: A KDD Process (Contd.....)

- **Data transformation**

  - Also known as data consolidation

  - it is a phase in which the selected data is transformed into forms appropriate for the mining procedure

- **Data mining**

  - clever techniques are applied to extract patterns potentially useful.

- **Pattern evaluation**

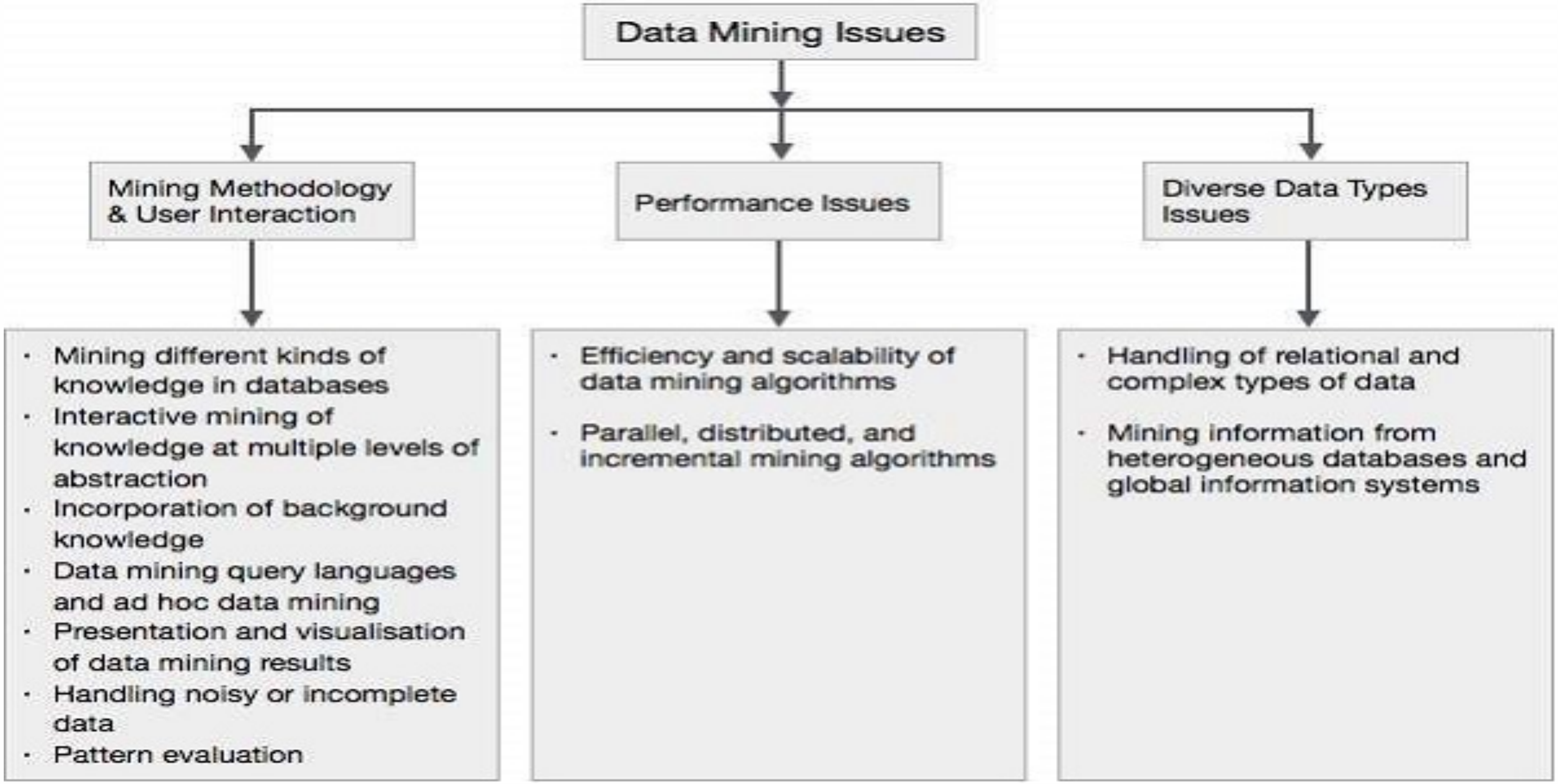
  - interesting patterns representing knowledge are identified based on given measures

Data mining: A KDD Process (Contd.....)

- **Knowledge representation**

  - final phase in which the discovered knowledge is visually represented to the user

Issue in Data Mining



Data Mining Issues

Mining Methodology & User Interaction

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualisation of data mining results
- Handling noisy or incomplete data
- Pattern evaluation

Performance Issues

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

Diverse Data Types Issues

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information systems

Figure 1.5 Data Mining

Application of Data Mining

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, improved underwriting, quality control, competitive analysis
  - Fraud detection and management
- **Other Applications:**
  - Text mining (news group, email, documents) and Web analysis
  - Intelligent query answering

Advantages of Data Mining (Contd.....)

- **Marketing /Retail**

- Data mining helps marketing companies to build models based on historical data which will precisely predict responders to the new marketing campaigns.

- Marketers will have appropriate approach for targeted customers

- Data mining helps retail companies as well. By using market basket analysis, a store can have an appropriate arrangement in such a way that customers can purchase frequent buying products together with pleasant. It also helps the retail companies to offer certain discounts which will attract more customers.

Advantages of Data Mining (Contd.....)

- **Finance / Banking**

  - By building a model from historical customer's data of loans, the bank officials and financial institution can determine good and bad loans.

  - Data mining also helps banks to detect fraudulent credit card transactions

- **Manufacturing**

  - Data mining is useful in operational engineering data which can detect faulty equipments and determines optimal control parameters.

  - Data mining can determine the range of control parameters which leads to the production of perfect product. Hence optimal control parameters can provide the desired quality.

Advantages of Data Mining (Contd.....)

- **Governments**

  - Data mining helps government agency to analyze records of financial transaction which will help in building patterns that can detect money laundering or criminal activities.

- **Market segmentation**

  - Data mining helps to identify the common characteristics of customers who buy the same products from your company.

Advantages of Data Mining (Contd.....)

- **Customer anticipation**

  - It helps to predict which customers may leave your company and go to a competitor.

- **Fraud detection**

  - It indentifies which transactions are most likely to be fraudulent.

Advantages of Data Mining (Contd.....)

- **Direct marketing**

  - Direct marketing identifies which prospects should be included to obtain the highest response rate.

- **Interactive marketing**

  - It is useful for predicting what each user on a Web site is most likely interested in seeing.

- **Market basket analysis**

  - It helps to understand what products or services are commonly purchased together.

- **Trend analysis**

  - Trend analysis identifies the difference between a typical customer this month and last.

Advantages of Data Mining

- **Market basket analysis**

  - It helps to understand what products or services are commonly purchased together.

- **Trend analysis**

  - Trend analysis identifies the difference between a typical customer this month and last.

Disadvantages of Data Mining

- **Privacy Issues**

- The internet is booming with social networks, ecommerce, blogs etc, the concerns about the personal privacy has been increasing.

- This worries the users as the information might be collected and used in unethical way which can potentially cause a lot of troubles.

- Businesses collect the information of its users for setting up the marketing strategies but there are chances that business might be taken by other firms or gets shut down and that's where a concern of misusing or leaking the personal information arises.

Disadvantages of Data Mining (Contd.....)

- **Security Issues**

  - Security is the biggest concern in data mining. Businesses own all the information of their employees which even includes personal and financial information, there are the chances of misusing data by hackers and which cause serious trouble to the organization and its employees.

Disadvantages of Data Mining (Contd.....)

- **Misuse of information/inaccurate information**

  - The information collected by the data mining may be exploited by unethical people or businesses in order to take benefits of vulnerable people.

  - Data mining techniques is not totally accurate. So inaccurate information may lead the wrong decision-making which may cause serious consequences.

DIGITAL LEARNING CONTENT

Parul® University