UNIT III INTRODUCTION TO DATA MINING 9

Data mining-KDD versus datamining, Stages of the Data Mining Process-task primitives, Data Mining Techniques -Data mining knowledge representation – Data mining query languages, Integration of a Data Mining System with a Data Warehouse – Issues, Data preprocessing – Data cleaning, Data transformation, Feature selection, Dimensionality reduction, Discretization and generating concept hierarchies-Mining frequent patterns- association-correlation

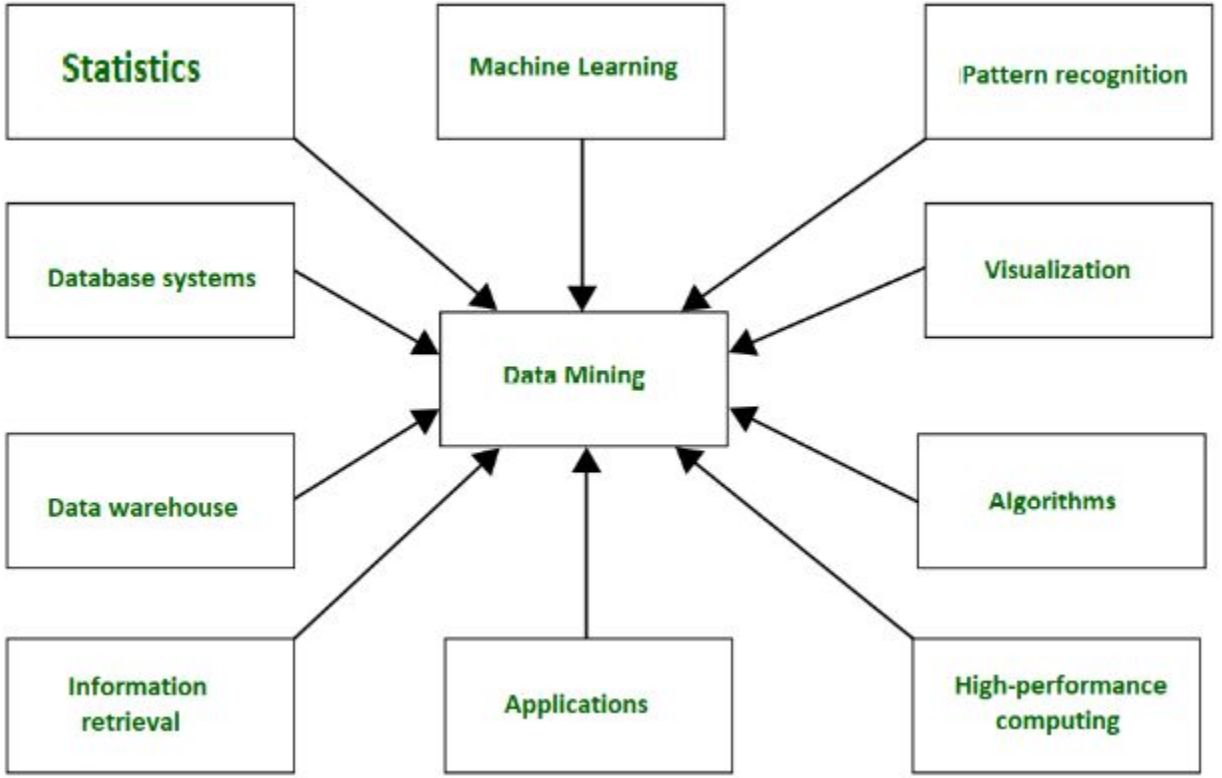
Data mining

Data Mining

- "**Mining**" is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining, etc. In the context of computer science, "**Data Mining"** can be referred to as **knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging**. It is basically the process carried out for the extraction of useful information from a bulk of data or data warehouses. One can see that the term itself is a little confusing. In the case of coal or diamond mining, the result of the extraction process is coal or diamond. But in the case of Data Mining, the result of the extraction process is not data!! Instead, data mining results are the patterns and knowledge that we gain at the end of the extraction process. In that sense, we can think of Data Mining as a step in the process of Knowledge Discovery or Knowledge Extraction.

- **Gregory Piatetsky-Shapiro** coined the term **“Knowledge Discovery in Databases”** in 1989. However, the term **‘data mining’** became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.
- Nowadays, data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use ‘data mining’ to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

Main Purpose of Data Mining

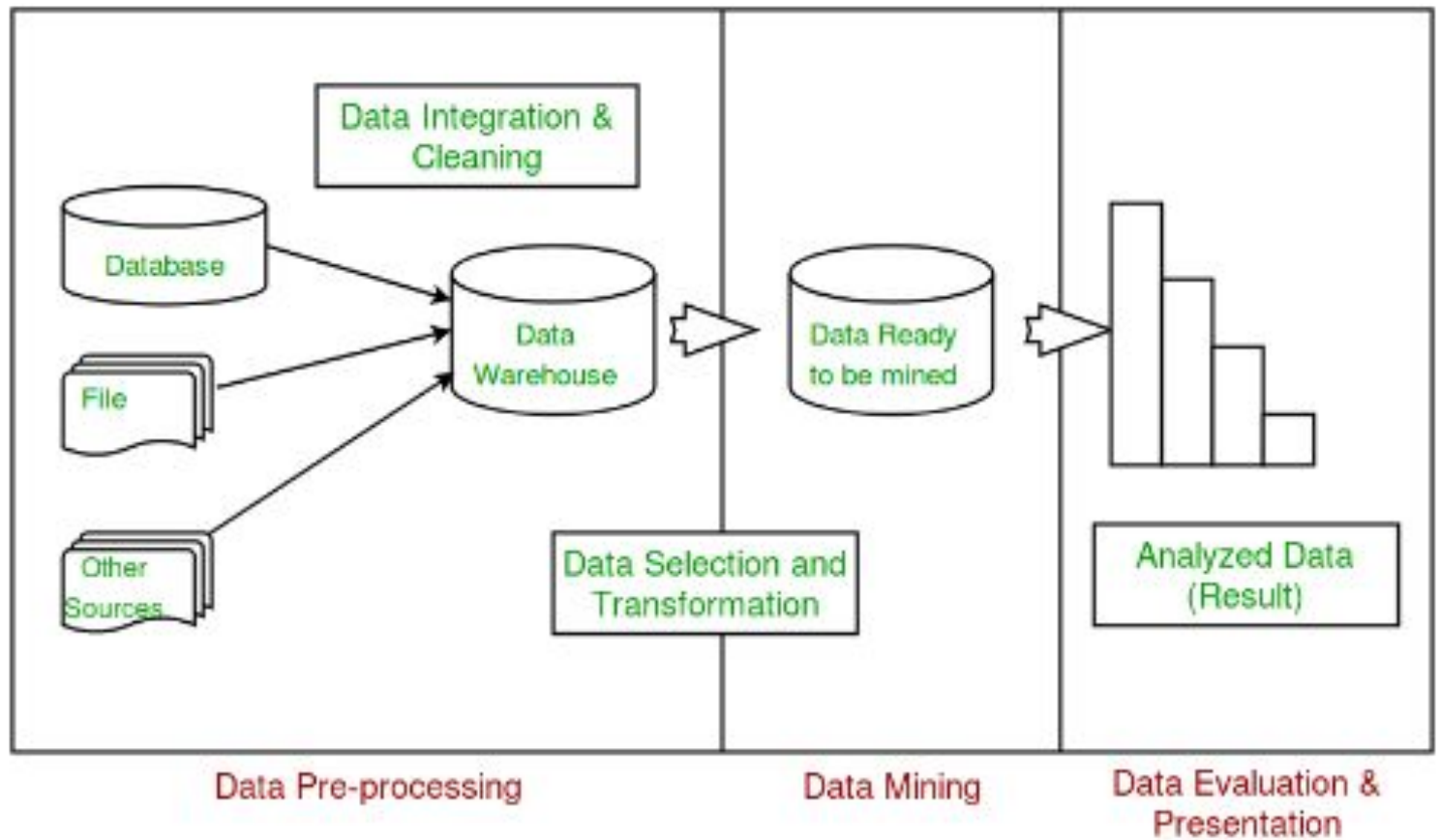


- Basically, Data mining has been integrated with many other techniques from other domains such as **statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization,** etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions.
- Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.
- Data Mining can be applied to any type of data e.g. **Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.**

Data Mining as a whole process

- The whole process of Data Mining consists of three main phases:

1. Data Pre-processing – Data cleaning, integration, selection, and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results


Data Integration &
Cleaning
Database
File
Other
Sources
Data
Warehouse
Data Ready
to be mined
Data Selection and
Transformation
Analyzed Data
(Result)
Data Pre-processing
Data Mining
Data Evaluation &
Presentation

Applications of Data Mining

1. Financial Analysis
2. Biological Analysis
3. Scientific Analysis
4. Intrusion Detection
5. Fraud Detection
6. Research Analysis

Real-life examples of Data Mining

- **Market Basket Analysis**: It is a technique that gives the careful study of purchases done by a customer in a supermarket. The concept is basically applied to identify the items that are bought together by a customer. Say, if a person buys bread, what are the chances that he/she will also purchase butter. This analysis helps in promoting offers and deals by the companies. The same is done with the help of data mining.

- **Protein Folding:** It is a technique that carefully studies the biological cells and predicts the protein interactions and functionality within biological cells. Applications of this research include determining **causes and possible cures for Alzheimer's, Parkinson's,** and cancer caused by Protein misfolding.

- **Fraud Detection:** Nowadays, in this land of cell phones, we can use data mining to analyze cell phone activities for comparing suspicious phone activity. This can help us to detects calls made on cloned phones. Similarly, with credit cards, comparing purchases with historical purchases can detect activity with stolen cards.
- Data mining also has many successful applications, such as business intelligence, Web search, bioinformatics, health informatics, finance, digital libraries, and digital governments.