Data warehousing and Mining

-Kiran Bhowmick

Course structure

<table>
<tr><td colspan="8">Program: Third Year B.Tech. in Computer Engineering</td><td colspan="2">Semester : V</td><td></td></tr>
<tr><td colspan="8">Course : Data Mining and Warehouse</td><td colspan="2">Course Code:DJ19CEC501</td><td></td></tr>
<tr><td colspan="8">Course : Data Mining and Warehouse Laboratory</td><td colspan="2">Course Code: DJ19CEL501</td><td></td></tr>
<tr><td colspan="4" rowspan="2">Teaching Scheme
(Hours / week)</td><td colspan="6">Evaluation Scheme</td><td></td></tr>
<tr><td colspan="3">Semester End Examination Marks (A)</td><td colspan="3">Continuous Assessment Marks (B)</td><td rowspan="2">Total marks (A+ B)</td></tr>
<tr><td rowspan="3">Lectures</td><td rowspan="3">Practical</td><td rowspan="3">Tutorial</td><td rowspan="3">Total Credits</td><td colspan="3">Theory</td><td>Term Test 1</td><td>Term Test 2</td><td>Avg.</td></tr>
<tr><td colspan="3">75</td><td>25</td><td>25</td><td>25</td><td>100</td></tr>
<tr><td colspan="3">Laboratory Examination</td><td colspan="2">Term work</td><td rowspan="2">Total Term work</td><td rowspan="3">50</td></tr>
<tr><td rowspan="2">3</td><td rowspan="2">1</td><td rowspan="2">-</td><td rowspan="2">4</td><td>Oral</td><td>Practical</td><td>Oral &Practical</td><td>Laboratory Work</td><td>Tutorial / Mini project / presentation/ Journal</td></tr>
<tr><td>-</td><td>-</td><td>25</td><td>15</td><td>10</td><td>25</td></tr>
</table>

Syllabus

1	**Introduction to Data Warehouse and Dimensional modelling:** Introduction to Strategic Information, Need for Strategic Information, Features of Data Warehouse, Data warehouse versus Data Marts, Data warehouse versus Data Lake, Top-down versus Bottom-up approach. Data warehouse architecture, metadata, E-R modelling versus Dimensional Modelling, Information Package Diagram, STAR schema, STAR schema keys, Snowflake Schema, Fact Constellation Schema, Factless Fact tables, Update to the dimension tables, Aggregate fact tables.
2	**ETL Process and OLAP:** Major steps in ETL process, Data extraction: Techniques, Data transformation: Basic tasks, Major transformation types, Data Loading: Applying Data, OLTP Vs OLAP, OLAP definition, Dimensional Analysis, Hypercubes, OLAP operations: Drill down, Roll up, Slice, Dice and Rotation, OLAP models: MOLAP, ROLAP.
3	**Introduction to Data Mining, Data Exploration and Preprocessing:** Data Mining Task and Techniques, KDD process, Issues in Data Mining, Applications of Data Mining Data Exploration: Types of Attributes, Statistical Description of Data, Data Visualization, Measuring data similarity and dissimilarity. Data Preprocessing: Major tasks in preprocessing, Data Cleaning: Missing values, Noisy data; Data Integration: Entity Identification Problem, Redundancy and Correlation Analysis, Tuple Duplication, Data Value Conflict Detection and Resolution; Data Reduction: Attribute subset selection, Histograms, Clustering and Sampling; Data Transformation & Data Discretization: Data Transformation by Normalization, Discretization by Binning, Discretization by Histogram Analysis, Concept hierarchy generation for Nominal data

4	**Classification and Prediction:** Basic Concepts of classification, Decision Tree Induction, Attribute Selection Measures using Information Gain, Tree pruning Bayes Classification Methods: Bayes’ Theorem, Naïve Bayesian Classification Rule - Based Classification: Using IFTHEN Rules for classification, Rule Extraction from a Decision Tree, Rule Quality Measures, Rule Pruning Model Evaluation & Selection: Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling, Cross Validation, Bootstrap, Model Selection Using Statistical Tests of Significance, Comparing Classifiers Based on Cost–Benefit and ROC Curves Improving Classification Accuracy: Ensemble classification, Bagging, Boosting and AdaBoost, Random Forests, Improving Classification Accuracy in Class Imbalance Data Prediction: Simple Linear regression
5	**Clustering:** Cluster Analysis and Requirements of Cluster Analysis Partitioning Methods: k-Means, k-Medoids Hierarchical Methods: Agglomerative, Divisive Density Based Methods: DBScan Evaluation of Clustering: Assessing Clustering Tendency, Determining Number of Clusters and Measuring cluster quality: Intrinsic and Extrinsic methods
6	**Mining Frequent Patterns and Association Rules:** Market Basket Analysis, Frequent Item sets, Closed Item sets, and Association Rule Frequent Item set Mining Methods: Apriori Algorithm, Association Rule Generation, Improving the Efficiency of Apriori. FP growth, Mining frequent Itemsets using Vertical Data Format Introduction to Mining Multilevel Association Rules and Multidimensional Association Rules
7	**Spatial and Web Mining:** Spatial Data, Spatial Vs. Classical Data Mining, Spatial Data Structures, Mining Spatial Association and Co-location Patterns, Spatial Clustering Techniques: CLARANS Extension, Web Mining: Web Content Mining, Web Structure Mining, Web Usage mining, Applications of Web Mining

Reference Books

- ***Paulraj Ponniah, —Data Warehousing: Fundamentals for IT Professional, Wiley India.***
- ***Reema Theraja —Data warehousing, Oxford University Press.***
- ***Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd edition.***
- ***M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education.***

Course outcomes

- Understand Data Warehouse fundamentals and data mining principles.
- Design data warehouse with dimensional modelling
- Understand ETL process and apply OLAP operations.
- Apply appropriate pre-processing techniques.
- Identify appropriate data mining algorithms to solve real world problems.
- Compare and evaluate different data mining techniques like classification, clustering and association rule mining

Chapter 1. Introduction

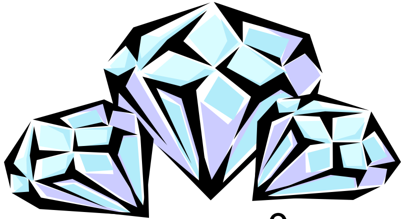
- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Classification of data mining systems
- Top-10 most popular data mining algorithms
- Major issues in data mining
- Overview of the course

Why Data Mining?

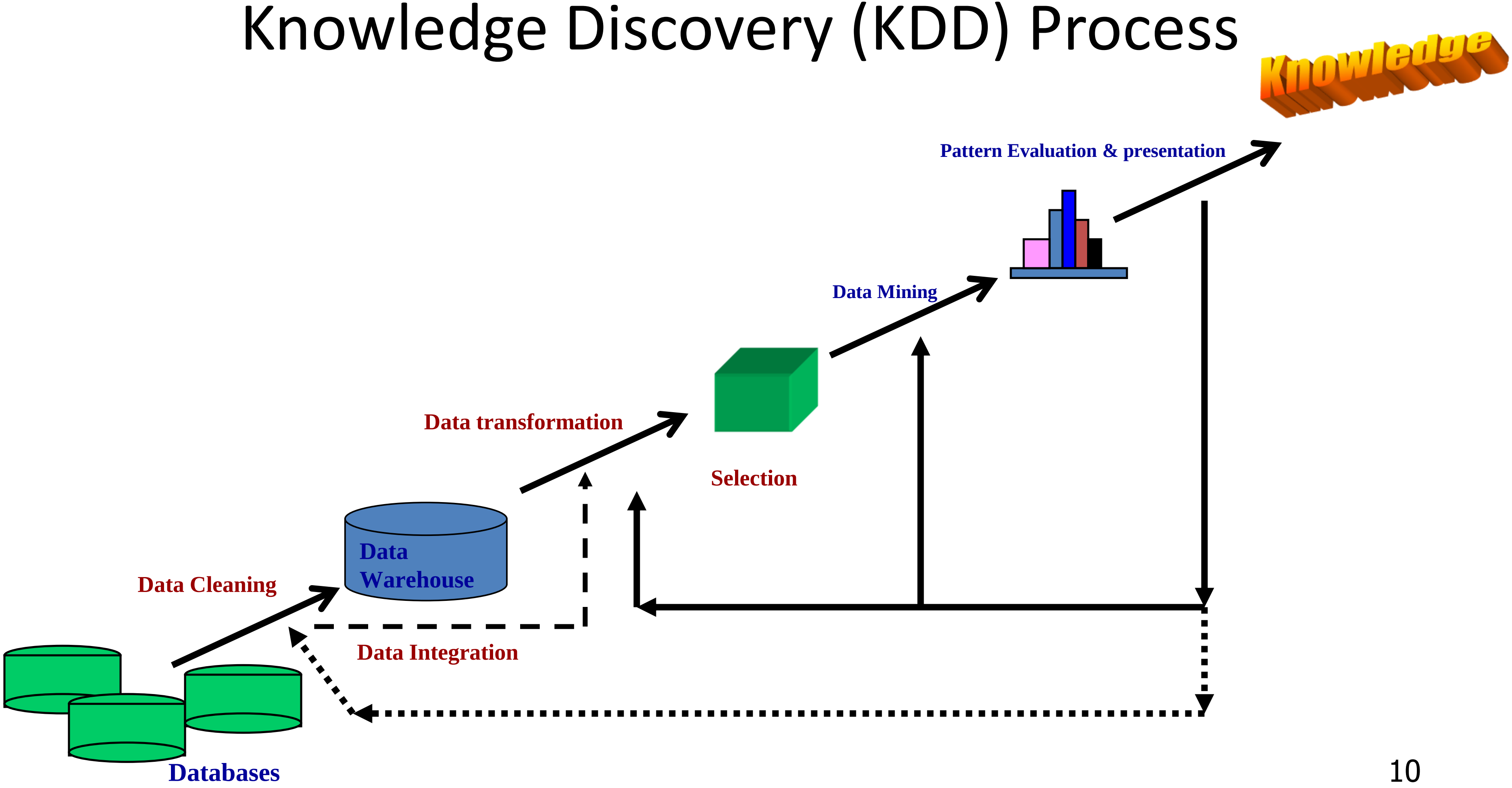
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

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?
- 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



Knowledge Discovery (KDD) Process



Knowledge Discovery from Data (KDD)

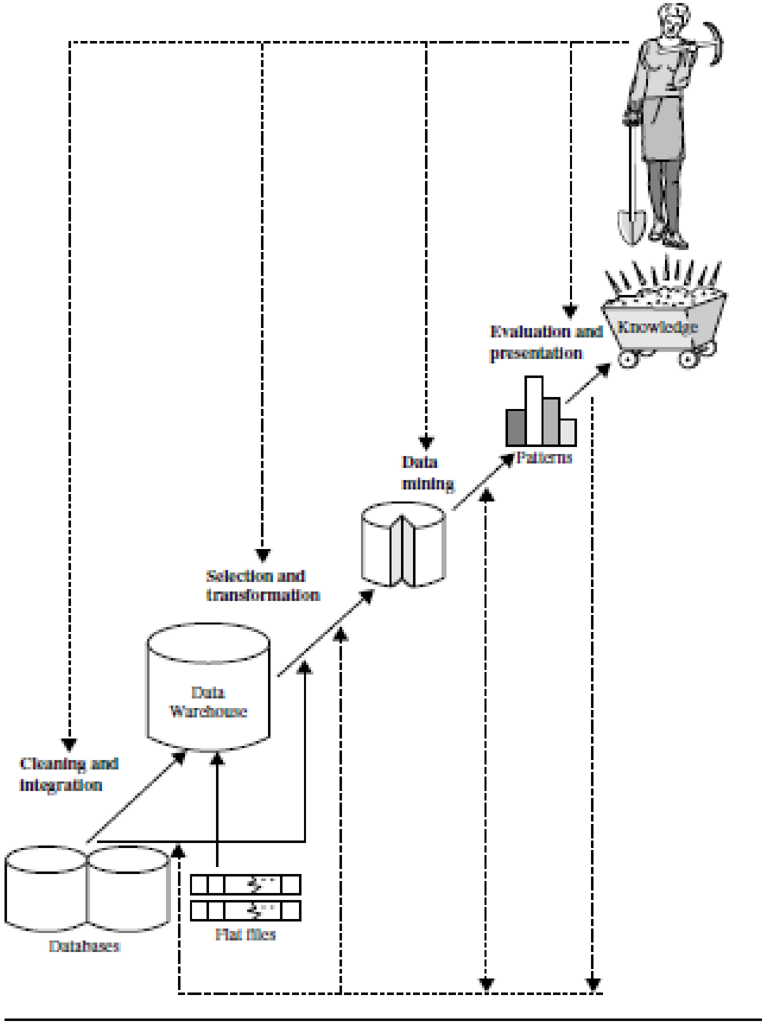
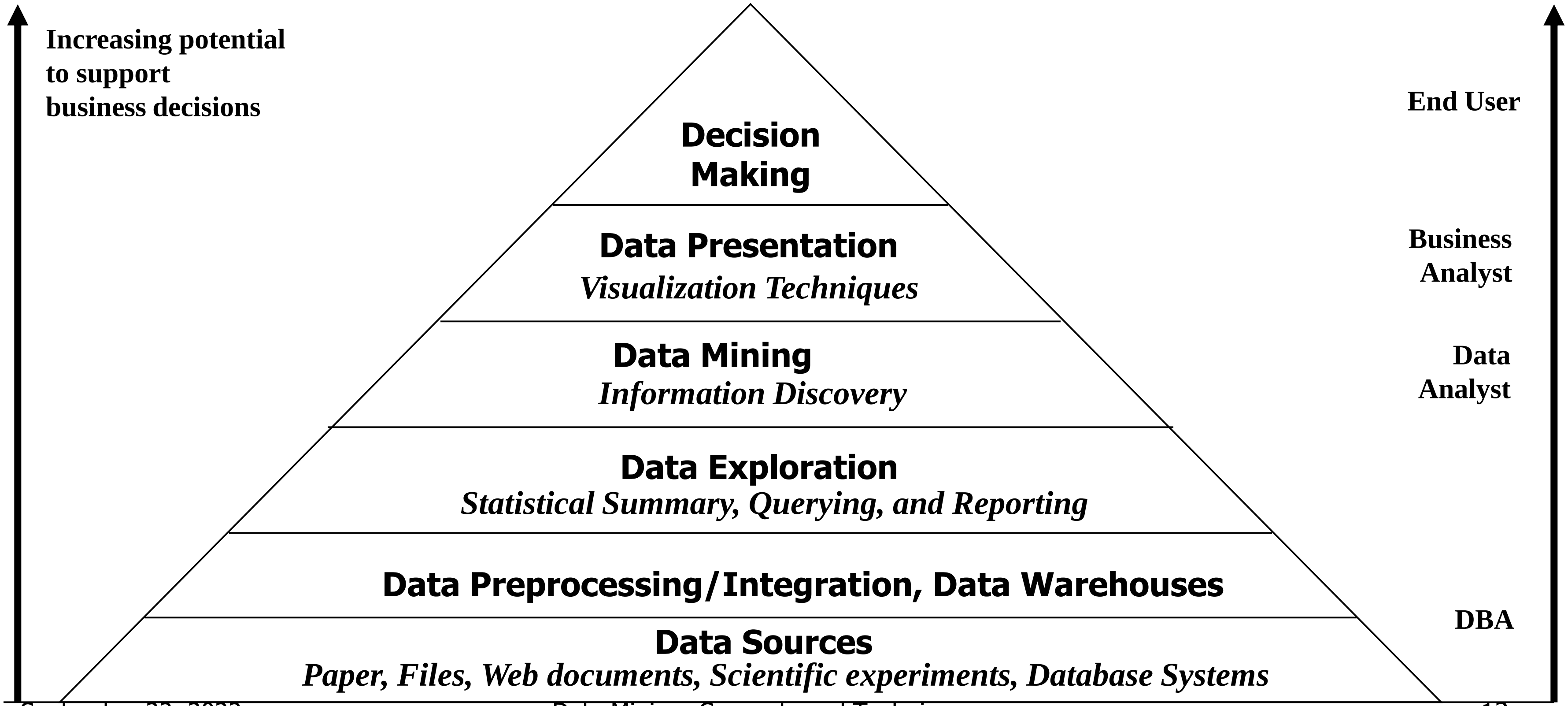


Figure 1.4 Data mining as a step in the process of knowledge discovery.

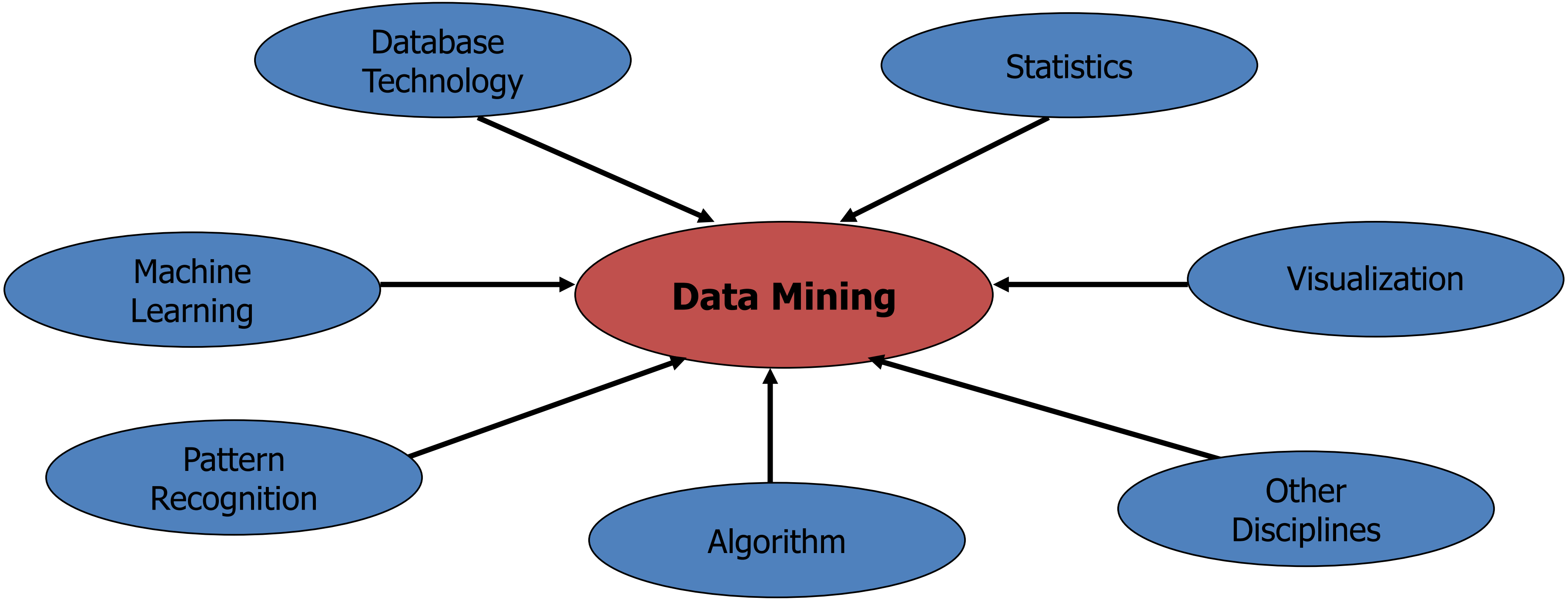
Data Mining and Business Intelligence


Increasing potential
to support
business decisions
End User
Decision
Making
Business
Analyst
Data Presentation
Visualization Techniques
Data
Analyst
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
DBA
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

Why Not Traditional Data Analysis?

- 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

Data Mining: Confluence of Multiple Disciplines



Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining Definition

- Finding hidden information in a database
- Fit data to a model
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

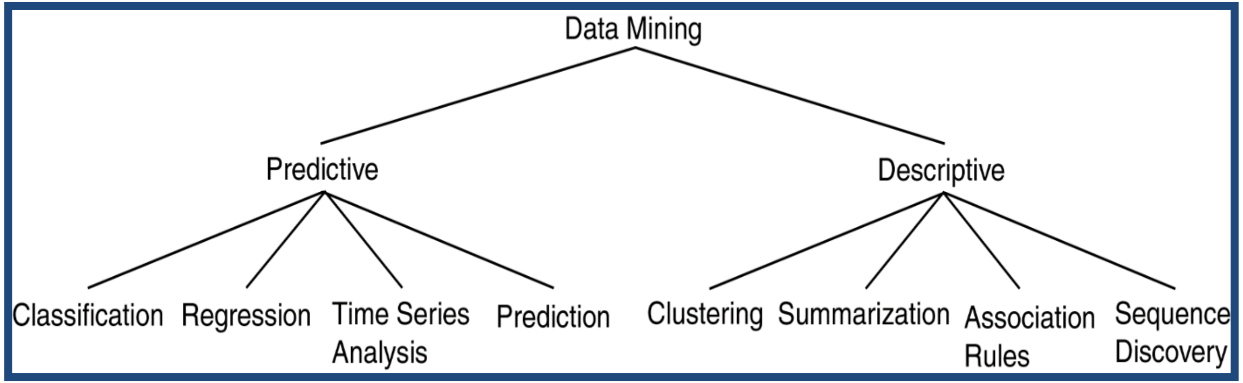
Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preference – Technique to choose the best model
- Search – Technique to search the data
  - “Query”

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
  - Find all credit applicants who are poor credit risks. (classification)
- Data Mining
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Models and Tasks



Descriptive data mining:
Descriptive data mining offers a detailed description of the data, for example- it gives insight into what's going on inside the data without any prior idea. This demonstrates the common characteristics in the results. It includes any information to grasp what's going on in the data without a prior idea.

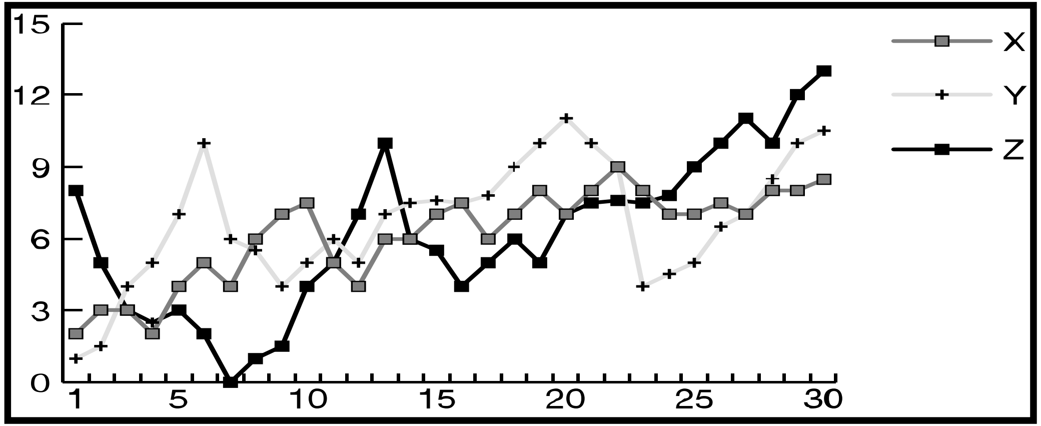
Predictive Data Mining:
This allows users to consider features that are not specifically available. For example, the projection of the market analysis in the next quarters with the output of the previous quarters, In general, the predictive analysis forecasts or infers the features of the data previously available. For an instance: judging by the outcomes of medical records of a patient who suffers from some real illness.

Basic Data Mining Tasks

- ***Classification*** maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Prediction*** predicting future states
  - Weather forecasts, earthquakes, floods etc
  - E.g. medical diagnosis, fraud detection etc.
- ***Regression*** is used to map a data item to a real valued prediction variable.
  - Regression involves predicting continuous, real-value quantities
  - regression involves the learning of the function that does this mapping.
  - E.g. Predicting house prices
- ***Time Series Analysis*** the value of an attribute is examined as it varies over time. Three basic functions:
  - Distance measures: determine similarity
  - Structure of line
  - Historical time series plot to predict future values

Ex: Time Series Analysis

- Example: Stock Market
- Predict future values
- Determine similar patterns over time



Basic Data Mining Tasks

- ***Clustering*** groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning
- ***Summarization*** maps data into subsets with associated simple descriptions.
  - Characterization or Generalization
  - It extracts or derives representative information about the database
  - E.g. Average CET score taking admission in an Engg college. Summarization will help estimate the type and intellect of student in the college
- ***Association rules*** uncovers relationships among data.
  - Also called link analysis, Affinity Analysis
  - An association rule is a model that identifies specific types of data associations.
  - E.g. Market-basket analysis

Basic Data Mining Tasks (cont'd)

- ***Sequence Discovery*** to determine sequential patterns in data.
  - Patterns are based on a time sequence of actions.
  - Similar to association – data are found that are related.
  - Difference than association – relationship is based on time. For AR – items must be purchased at same time whereas for sequence discovery items can be purchased over a period of time.
  - E.g. Analyzing web logs of a website to understand what sequence of pages are frequently visited by users.
  - (A, B, C) or (A, D, B, C) or (A, E, B, C). Then add a link directly from page A to page C.