Why we need Data Mining?

Volume of information is increasing everyday that we can handle from business transactions, scientific data, sensor data, Pictures, videos, etc. So, we need a system that will be capable of extracting essence of information available and that can automatically generate report, views or summary of data for better decision-making.

Why Data Mining is used in Business?

Data mining is used in business to make better managerial decisions by:

- Automatic summarization of data
- Extracting essence of information stored.
- Discovering patterns in raw data.

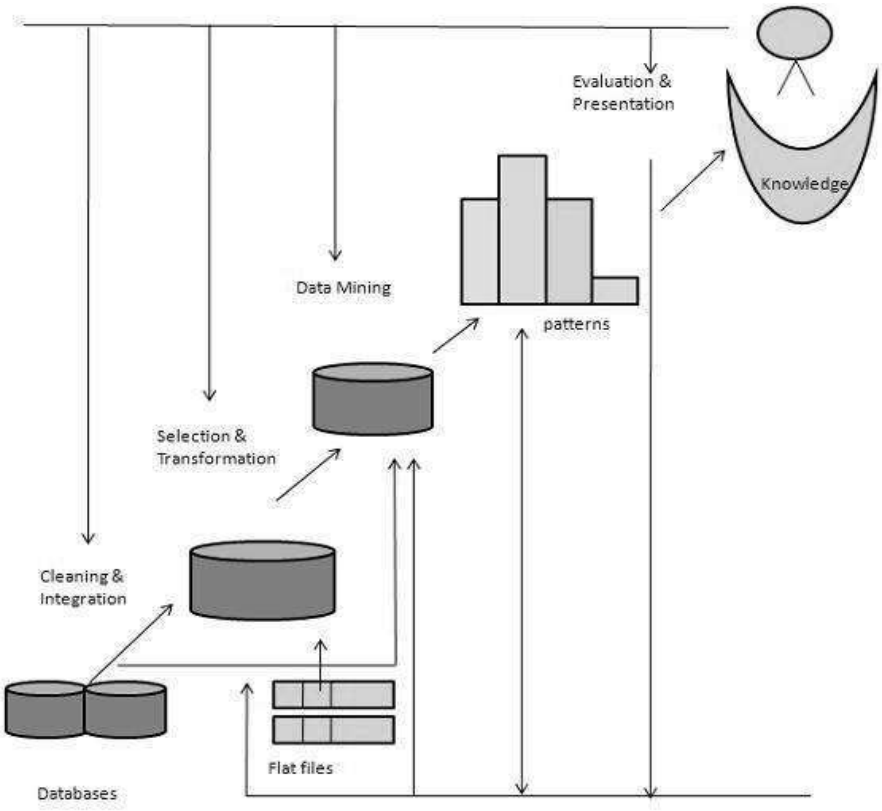
What is Data Mining?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications −

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Knowledge discovery from Data (KDD) is essential for data mining. While others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process −

- **Data Cleaning** − In this step, the noise and inconsistent data is removed.
- **Data Integration** − In this step, multiple data sources are combined.
- **Data Selection** − In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** − In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** − In this step, data patterns are evaluated.
- **Knowledge Presentation** − In this step, knowledge is represented.



What kinds of data can be mined?

1. Flat Files
2. Relational Databases
3. DataWarehouse
4. Transactional Databases
5. Multimedia Databases
6. Spatial Databases
7. Time Series Databases
8. World Wide Web(WWW)

1. ***Flat Files***
   - Flat files are defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
   - Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, and then there will be no relations between the tables.
   - Flat files are represented by data dictionary. Eg: CSV file.
   - **Application**: Used in Data Warehousing to store data, Used in carrying data to and from server, etc.
2. ***Relational Databases***
   - A Relational database is defined as the collection of data organized in tables with rows and columns.
   - Physical schema in Relational databases is a schema which defines the structure of tables.
   - Logical schema in Relational databases is a schema which defines the relationship among tables.
   - Standard API of relational database is SQL.
   - **Application**: Data Mining, ROLAP model, etc.

3. ***DataWarehouse***
   - A datawarehouse is defined as the collection of data integrated from multiple sources that will query and decision making.
   - There are three types of datawarehouse: **Enterprise** datawarehouse, **Data Mart** and **Virtual** Warehouse.
   - Two approaches can be used to update data in DataWarehouse: **Query-driven** Approach and **Update-driven** Approach.
   - **Application**: Business decision making, Data mining, etc.
4. ***Transactional Databases***
   - Transactional databases are a collection of data organized by time stamps, date, etc to represent transaction in databases.
   - This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
   - Highly flexible system where users can modify information without changing any sensitive information.
   - Follows ACID property of DBMS.
   - **Application**: Banking, Distributed systems, Object databases, etc.
5. ***Multimedia Databases***
   - Multimedia databases consists audio, video, images and text media.
   - They can be stored on Object-Oriented Databases.
   - They are used to store complex information in pre-specified formats.
   - **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.
6. ***Spatial Database***
   - Store geographical information.
   - Stores data in the form of coordinates, topology, lines, polygons, etc.
   - **Application**: Maps, Global positioning, etc.
7. ***Time-series Databases***
   - Time series databases contain stock exchange data and user logged activities.
   - Handles array of numbers indexed by time, date, etc.
   - It requires real-time analysis.
   - **Application**: eXtremeDB, Graphite, InfluxDB, etc.
8. ***WWW***
   - WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
   - It is the most heterogeneous repository as it collects data from multiple resources.
   - It is dynamic in nature as Volume of data is continuously increasing and changing.
   - **Application**: Online shopping, Job search, Research, studying, etc.

What kinds of Patterns can be mined?

On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining −

a) Descriptive
b) Classification and Prediction

a) Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions −

1. Class/Concept Description
2. Mining of Frequent Patterns
3. Mining of Associations
4. Mining of Correlations
5. Mining of Clusters

1. Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways −

- **Data Characterization** − This refers to summarizing data of class under study. This class under study is called as Target Class.
- **Data Discrimination** − It refers to the mapping or classification of a class with some predefined group or class.

2. Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns −

- **Frequent Item Set** − It refers to a set of items that frequently appear together, for example, milk and bread.
- **Frequent Subsequence** − A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** − Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

3. Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

4. Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

5. Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

b) Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

1. Classification (IF-THEN) Rules
2. Prediction
3. Decision Trees
4. Mathematical Formulae
5. Neural Networks
6. Outlier Analysis
7. Evolution Analysis

The list of functions involved in these processes is as follows −

1. **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.

2. **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.

3. **Decision Trees** − A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label.

4. **Mathematical Formulae** – Data can be mined by using some mathematical formulas.

5. **Neural Networks** − Neural networks represent a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to "**learn**" from the data.

6. **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.

7. **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

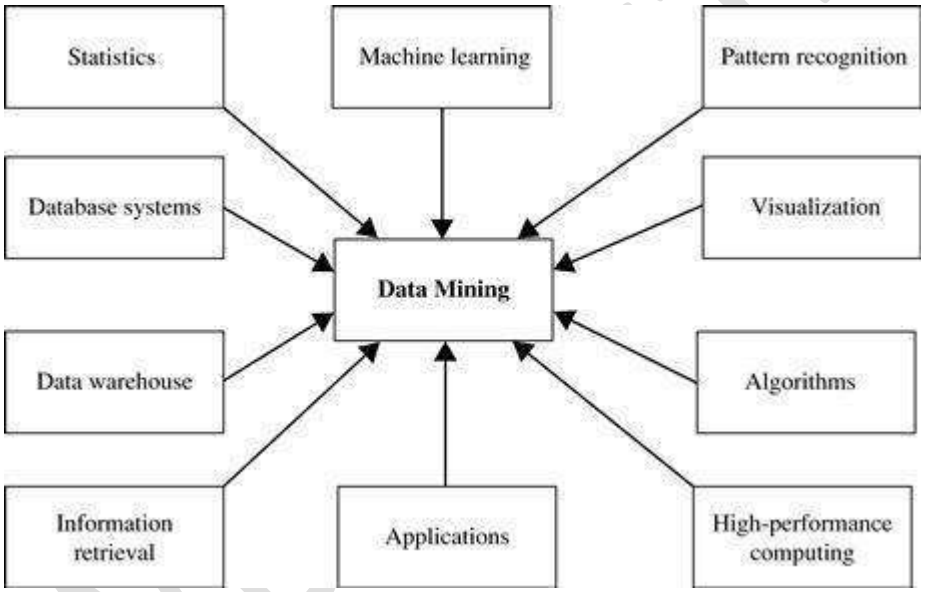
Data Mining Task Primitives

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
- A data mining query is defined in terms of data mining task primitives.

Note − These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives −

- Set of task relevant data to be mined.
- Kind of knowledge to be mined.
- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Which Technologies are used in data mining?



1. Statistics:

- It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
- Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

2. Machine learning

- **Arthur Samuel** defined machine learning as a field of study that gives computers the ability to learn without being programmed.
- When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.
- In machine learning, an algorithm is constructed to predict the data from the available database **(Predictive analysis).**
- It is related to computational statistics.

The four types of machine learning are:

a. Supervised learning

- It is based on the classification.
- It is also called as **inductive learning**. In this method, the desired outputs are included in the training dataset.

b. Unsupervised learning

- Unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

c. Semi-supervised learning

- Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs).

d. Active learning

- Active learning is a powerful approach in analyzing the data efficiently.
- The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

3. Information retrieval

- Information deals with uncertain representations of the semantics of objects (text, images).

For example: Finding relevant information from a large document.

4. Database systems and data warehouse

- Databases are used for the purpose of recording the data as well as data warehousing.
- Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.
- Data warehouses are used to store historical data which helps to take strategically decision for business.
- It is used for online analytical processing (OALP), which helps to analyze the data.

5. Pattern Recognition:

Pattern recognition is the automated recognition of patterns and regularities in data. Pattern recognition is closely related to artificial intelligence and machine learning, together with applications such as data mining and knowledge discovery in databases (KDD), and is often used interchangeably with these terms.

6. Visualization:

It is the process of extracting and visualizing the data in a very clear and understandable way without any form of reading or writing by displaying the results in the form of pie charts, bar graphs, statistical representation and through graphical forms as well.

7. Algorithms:

To perform data mining techniques we have to design best algorithms.

8. High Performance Computing:

High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business.

Are all patterns interesting?

- Typically the answer is No – only small fraction of the patterns potentially generated would actually be of interest to a given user.
- What makes patterns interesting?
  - The answer is if it is (1) easily understood by humans, (2) valid on new or test data with some degree of certainty, (3) potentially useful, (4) novel.
- A Pattern is also interesting if it is validates a hypothesis that the user sought to confirm.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Like,

- Loan payment prediction and customer credit policy analysis.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc.

A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration.

Major Issues in data mining:

Data mining is a dynamic and fast-expanding field with great strengths. The major issues can divided into five groups:

a) Mining Methodology
b) User Interaction
c) Efficiency and scalability
d) Diverse Data Types Issues
e) Data mining society

a) Mining Methodology:

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Mining knowledge in multidimensional space –** when searching for knowledge in large datasets, we can explore the data in multidimensional space.
- **Handling noisy or incomplete data** − the data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − the patterns discovered should be interesting because either they represent common knowledge or lack novelty.

b) User Interaction:

- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

c) Efficiency and scalability

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

d) Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

e) Data Mining and Society

- **Social impacts of data mining –** With data mining penetrating our everyday lives, it is important to study the impact of data mining on society.
- **Privacy-preserving data mining –** data mining will help scientific discovery, business management, economy recovery, and security protection.
- **Invisible data mining –** we cannot expect everyone in society to learn and master data mining techniques. More and more systems should have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms.

Data Objects and Attribute Types:

Data Object:

An Object is real time entity.

Attribute:

It can be seen as a data field that represents characteristics or features of a data object. For a customer object attributes can be customer Id, address etc. The attribute types can represented as follows—

1. **Nominal Attributes – related to names:** The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represent some category or state and that's why nominal attribute also referred as categorical attributes.

Example:

Attribute	Values
Colors	Black, Green, Brown, red

2. **Binary Attributes:** Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.
   i) Symmetric: Both values are equally important (Gender).
   ii) Asymmetric: Both values are not equally important (Result).

Attribute	Values
Gender	Male, Female

Attribute	Values
Result	Pass, Fail

3. **Ordinal Attributes:** The Ordinal Attributes contains values that have a meaningful sequence or ranking (order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Values
Grade	O, S, A, B, C, D, F

4. **Numeric:** A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types.
   i. An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point or we can call zero point. Data can be added and subtracted at interval scale but cannot be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice than the other day we cannot say that one day is twice as hot as another day.
   ii. A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range and five number summaries can be given.

5. **Discrete:** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.
   Example

Attribute	Values
Profession	Teacher, Business man, Peon
ZIP Code	521157, 521301

6. **Continuous:** Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.
   Example:

Attribute	Values
Height	5.4, 5.7, 6.2, etc.,
Weight	50, 65, 70, 73, etc.,

Basic Statistical Descriptions of Data:

Basic Statistical descriptions of data can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

1. Measuring the central Tendency:

There are many ways to measure the central tendency.

a) Mean: Let us consider the values are $x_1, x_2, x_3, \ldots\ldots, x_N$ be a set of N observed values or observations of X.

The Most common numeric measure of the "center" of a set of data is the *mean.*

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1+x_2+x_3+------+x_N}{N}$$

Sometimes, each values x_i in a set maybe associated with weight w_i for i=1, 2, …., N. then the mean is as follows

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{N} = \frac{w_1x_1+w_2x_2+w_3x_3+------+w_Nx_N}{N}$$

This is called **weighted arithmetic mean** or **weighted average**.

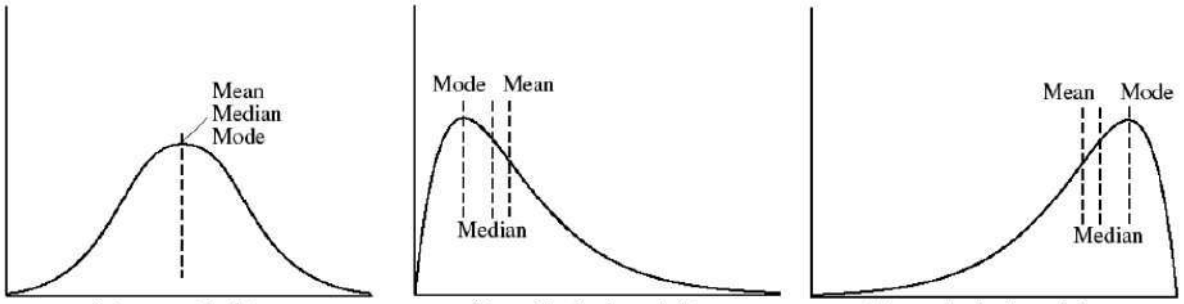
b) Median: Median is middle value among all values.

For N number of odd list median is $\frac{N+1}{2}$ th value.

For N number of even list median is $\frac{N+(N+1)}{2}$ th value.

c) Mode:

- The *mode* is another measure of central tendency.
- Datasets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal.
- A dataset with two or more modes is multimodal. If each data occurs only once, then there is no mode.



(a) symmetric data (b) positively skewed data (c) negatively skewed data

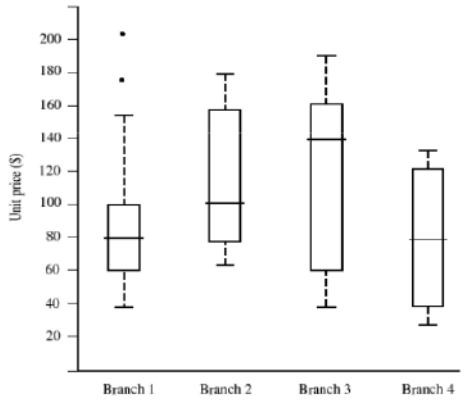
2. Measuring the Dispersion of data:

- The data can be measured as range, quantiles, quartiles, percentiles, and interquartile range.
- The k^{th} percentile of a set of data in numerical order is the value x_i having the property that k percent of the data entries lay at or below x_i.
- The measures of data dispersion: Range, Five-number summary, Interquartile range (IQR), Variance and Standard deviation

Five Number Summary:

This contains five values Minimum, Q1 (25% value), Median, Q3 (75% value), and Maximum. These Five numbers are represented as Boxplot in graphical format.

- Boxplot is Data is represented with a box.
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ.
- The median is marked by a line within the box.
- Whiskers: two lines outside the box extend to Minimum and Maximum.
- To show outliers, the whiskers are extended to the extreme low and high observations only if these values are less than 1.5 * IQR beyond the quartiles.



Variance and Standard Deviation:

Let us consider the values are $x_1, x_2, x_3, \ldots\ldots, x_N$ be a set of N observed values or observations of X. The variance formula is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Standard Deviation is the square root of variance σ^2.
- σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- σ =0 only when there is no spread, that is, when all observations have the same value.

3. Graphical Displays of Basic Statistical data:

There are many types of graphs for the display of data summaries and distributions, such as:

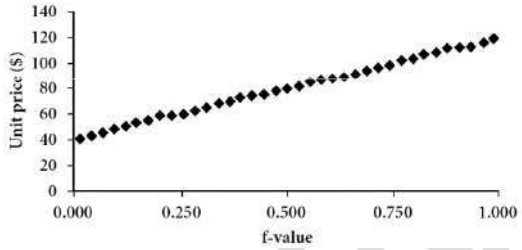
- Bar charts
- Pie charts
- Line graphs
- Boxplot
- Histograms
- Quantile plots, Quantile - Quantile plots
- Scatter plots

The data values can represent as Bar charts, pie charts, Line graphs, etc.

Quantile plots:

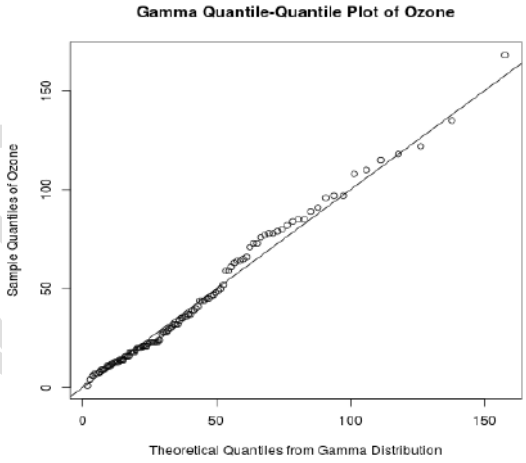
- A quantile plot is a simple and effective way to have a first look at a univariate data distribution.
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i
- Note that
  - the 0.25 quantile corresponds to quartile Q1,
  - the 0.50 quantile is the median, and
  - the 0.75 quantile is Q3.

A quantile plot for the unit price data of AllElectronics.



Quantile - Quantile plots:

In statistics, a Q-Q plot is a portability plot, which is a graphical method for comparing two portability distributions by plotting their quantiles against each other.



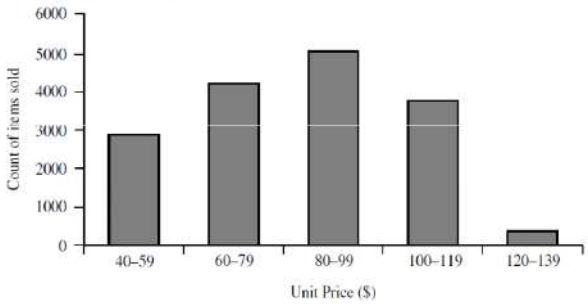
Histograms *or* frequency histograms:

- A univariate graphical method
- Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data
- If the attribute is categorical, then one rectangle is drawn for each known value of A, and the resulting graph is more commonly referred to as a bar chart.
- If the attribute is numeric, the term histogram is preferred.

- **Example:** A set of unit price data for items sold at a branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

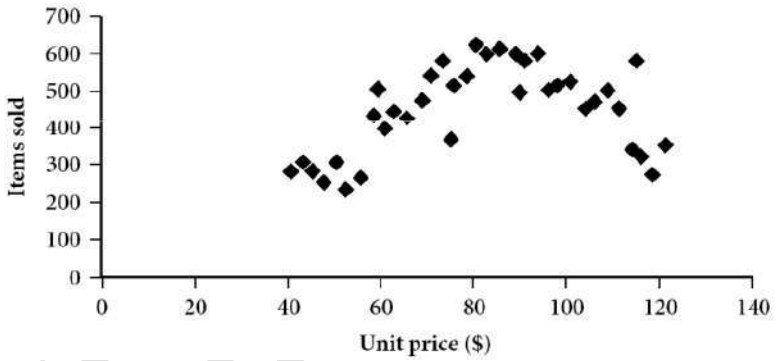
- **Example: A histogram**



Scatter Plot:

- Scatter plot
  - Is one of the most effective graphical methods for determining if there appears to be a relationship, clusters of points, or outliers between two numerical attributes.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

- A scatter plot for the data set of AllElectronics.



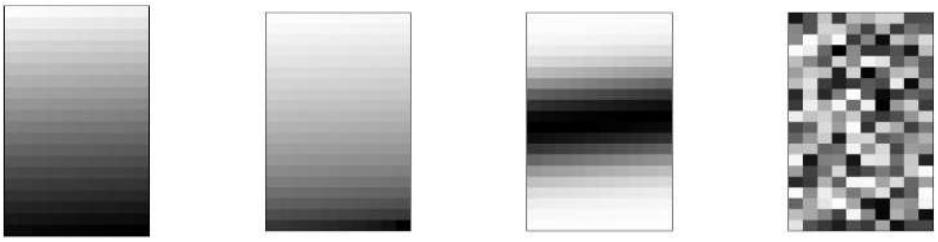
Data Visualization:

Visualization is the use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data.

- Categorization of visualization methods:
  a) Pixel-oriented visualization techniques
  b) Geometric projection visualization techniques
  c) Icon-based visualization techniques
  d) Hierarchical visualization techniques
  e) Visualizing complex data and relations

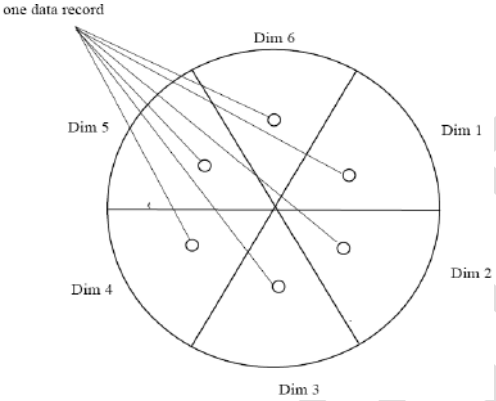
a) Pixel-oriented visualization techniques

- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



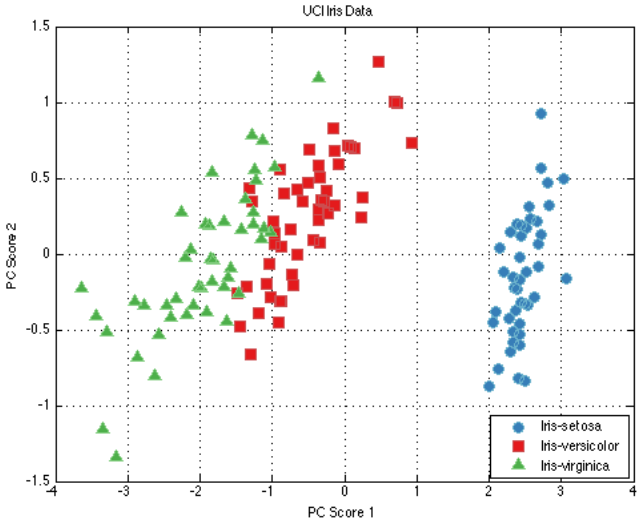
(a) Income (b) Credit Limit (c) transaction volume (d) age

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



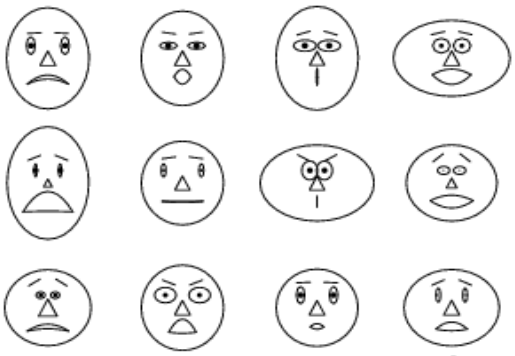
b) Geometric projection visualization techniques

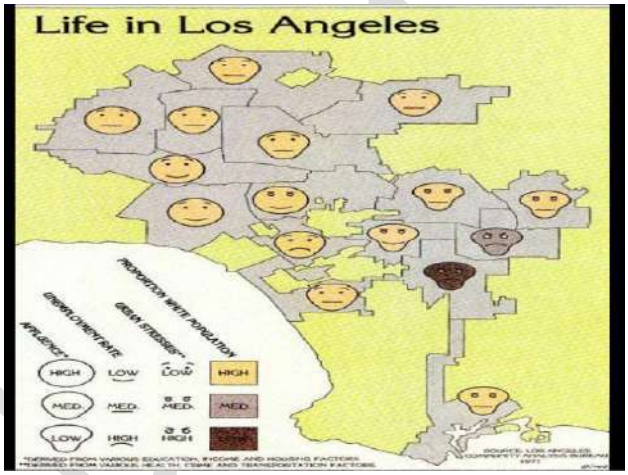
- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  - Parallel coordinates



c) Icon-based visualization techniques

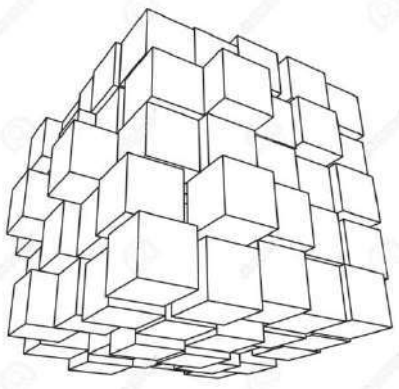
- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval



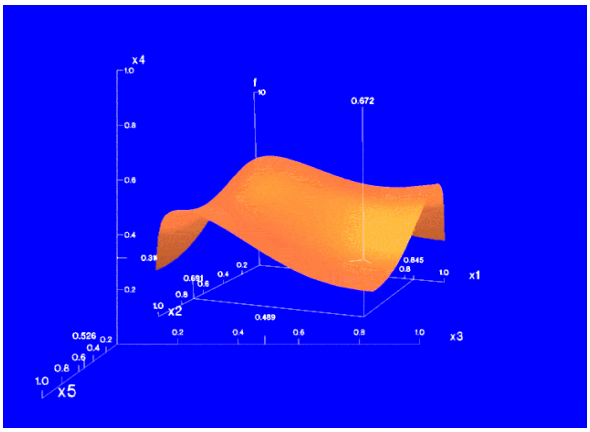


d) Hierarchical visualization techniques

- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube



Info Cube



Worlds-within-worlds

e) Visualizing complex data and relations

- Visualizing non-numerical data: text and social networks
- **Tag cloud:** visualizing user-generated tags
  - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing

Similarity and Dissimilarity

Distance or similarity measures are essential to solve many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in literature to compare two data distributions. As the names suggest, a similarity measures how close two distributions are. For multivariate data complex summary methods are developed to answer this question.

Similarity Measure

- Numerical measure of how alike two data objects are.
- Often falls between 0 (no similarity) and 1 (complete similarity).

Dissimilarity Measure

- Numerical measure of how different two data objects are.
- Range from 0 (objects are alike) to ∞ (objects are different).

Proximity refers to a similarity or dissimilarity.

Similarity/Dissimilarity for Simple Attributes

Here, *p* and *q* are the attribute values for two data objects.

Attribute Type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal	$s = 1 - \frac{\|p-q\|}{n-1}$ (values mapped to integer 0 to n-1, where n is the number of values)	$d = \frac{\|p-q\|}{n-1}$
Interval or Ratio	$s = 1 - \|p-q\|$, $s = \frac{1}{1+\|p-q\|}$	$d = \|p-q\|$

Common Properties of Dissimilarity Measures

Distance, such as the Euclidean distance, is a dissimilarity measure and has some well known properties:

1. $d(p, q) \geq 0$ for all p and q, and $d(p, q) = 0$ if and only if $p = q$,
2. $d(p, q) = d(q,p)$ for all p and q,
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all p, q, and r, where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

A distance that satisfies these properties are called a **metric**. Following is a list of several common distance measures to compare multivariate data. We will assume that the attributes are all continuous.

a) Euclidean Distance

Assume that we have measurements x_{ik}, $i = 1, \dots, N$, on variables $k = 1, \dots, p$ (also called attributes).

The Euclidean distance between the ith and jth objects is

$$d_E(i, j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

for every pair (i, j) of observations.

The weighted Euclidean distance is

$$d_{WE}(i, j) = \left(\sum_{k=1}^{p} W_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

If scales of the attributes differ substantially, standardization is necessary.

b) Minkowski Distance

The Minkowski distance is a generalization of the Euclidean distance.

With the measurement, x_{ik}, $i = 1, \dots, N$, $k = 1, \dots, p$, the Minkowski distance is

$$d_M(i, j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}},$$

where $\lambda \geq 1$. It is also called the L_λ metric.

- $\lambda = 1$: L_1 metric, Manhattan or City-block distance.
- $\lambda = 2$: L_2 metric, Euclidean distance.
- $\lambda \to \infty$: L_∞ metric, Supremum distance.

$$\lim \lambda \to \infty = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}} = \max \left(|x_{i1} - x_{j1}|, \dots, |x_{ip} - x_{jp}| \right)$$

Note that λ and p are two different parameters. Dimension of the data matrix remains finite.

c) Mahalanobis Distance

Let **X** be a N × p matrix. Then the ith row of **X** is

$$x_i^T = (x_{i1}, x_{i2}, \ldots\ldots, x_{ip})$$

The Mahalanobis distance is

$$d_{MH}(i,j) = \left((x_i - x_j)^T \Sigma^{-1} (x_i - x_j)\right)^{\frac{1}{2}}$$

where ∑ is the p×p sample covariance matrix.

Common Properties of Similarity Measures

Similarities have some well known properties:

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$,
2. $s(p, q) = s(q, p)$ for all p and q, where $s(p, q)$ is the similarity between data objects, p and q.

Similarity Between Two Binary Variables

The above similarity or distance measures are appropriate for continuous variables. However, for binary variables a different approach is necessary.

	q=1	q=0
p=1	$n_{1,1}$	$n_{1,0}$
p=0	$n_{0,1}$	$n_{0,0}$

Simple Matching and Jaccard Coefficients

- Simple matching coefficient = $(n_{1,1} + n_{0,0}) / (n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0})$.
- Jaccard coefficient = $n_{1,1} / (n_{1,1} + n_{1,0} + n_{0,1})$.