# Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, fraud detection, spam Email filtering, or even to discern the sentiment or opinion of users.

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and

determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table.

**Definition:** In simple words, data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

**Description:** Key features of data mining:

• Automatic pattern predictions based on trend and behaviour analysis.

• Prediction based on likely outcomes.

• Creation of decision-oriented information.

• Focus on large data sets and databases for analysis.

• Clustering based on finding and visually documented groups of facts not previously known.

The Data Mining Process:

1. Database Size: For creating a more powerful system more data is required to processed and maintained.

2. Query complexity: For querying or processing more complex queries and the greater the number of queries, the more powerful system is required.

Uses:

1. Data mining techniques are useful in many research projects, including mathematics, cybernetics, genetics and marketing.

2. With data mining, a retailer could manage and use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. The retailer could also develop products and promotions to appeal to specific customer segments based on mining demographic data from comment or warranty cards.

## Business Context:

Data Mining is largely used in several applications such as understanding consumer research marketing, product analysis, demand and supply analysis, e-commerce, investment trend in stocks & real estates, telecommunications and so on. Data Mining is based on mathematical algorithm and analytical skills to drive the desired results from the huge database collection.

Data Mining has great importance in today's highly competitive business environment. A new concept of Business Intelligence data mining has evolved now, which is widely used by leading corporate houses to stay ahead of their competitors. Business Intelligence (BI) can help in providing latest information and used for competition analysis, market research, economical trends, consume behavior, industry research, geographical information analysis and so on. Business Intelligence Data Mining helps in decision-making.

Data Mining applications are widely used in direct marketing, health industry, e-commerce, customer relationship management (CRM), FMCG industry, telecommunication industry and financial sector. Data mining is available in various forms like text mining, web mining, audio & video data mining, pictorial data mining, relational databases, and social networks data mining.

Data mining, however, is a crucial process and requires lots of time and patience in collecting desired data due to complexity and of the databases. This could also be possible that you need to look for help from outsourcing companies. These outsourcing companies are specialized in extracting or mining the data, filtering it and then keeping them in order for analysis. Data Mining has been used

in different context but is being commonly used for business and organizational needs for analytical purposes

Usually data mining requires lots of manual job such as collecting information, assessing data, using internet to look for more details etc. The second option is to make software that will scan the internet to find relevant details and information. Software option could be the best for data mining as this will save tremendous amount of time and labor. Some of the popular data mining software programs available are Connexor Machines, Free Text Software Technologies, Megaputer Text Analyst, SAS Text Miner, LexiQuest, WordStat, Lextek Profiling Engine.

## Market Analysis and Management

Listed below are the various fields of market where data mining is used −

- **Customer Profiling** − Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** − Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** − Data mining performs association/correlations between product sales.
- **Target Marketing** − Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** − Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** − Data mining provides us various multidimensional summary reports.

## Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector −

- **Finance Planning and Asset Evaluation** − It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** − It involves summarizing and comparing the resources and spending.
- **Competition** − It involves monitoring competitors and market directions.

# Technological Context:

### 1. Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining −

- Descriptive
- Classification and Prediction

### 2. Descriptive Function

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

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- 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**

CCluster 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.

# Social Context

Social media mining is the process of obtaining big data from user-generated content on social media sites and mobile apps in order to extract patterns, form conclusions about users, and act upon the information, often for the purpose of advertising to users or conducting research. The term is an analogy to the resource extraction process of mining for rare minerals. Resource extraction mining requires mining companies to sift through vast quantities of raw ore to find the precious minerals; likewise, social media mining requires human data analysts and automated software programs to sift through massive amounts of raw social media data in order to discern patterns and trends relating to social media usage, online behaviours, sharing of content, connections between individuals, online buying behaviour, and more. These patterns and trends are of interest to companies, governments and not-for-profit organizations, as these organizations can use these patterns and trends to design their strategies or introduce new programs, new products, processes or services.

Social media mining uses a range of basic concepts from computer science, data mining, machine learning and statistics. Social media miners develop algorithms suitable for investigating massive files of social media data. Social media mining is based on theories and methodologies from social network analysis, network science, sociology, ethnography, optimization and mathematics. It encompasses the tools to formally represent, measure and model meaningful patterns from large-scale social media data. major corporations, governments and not-for-profit organizations engaged in social media mining to obtain data about customers, clients and citizens.

## Social Media Data Mining Methods

Applying data mining techniques to social media is relatively new as compared to other fields of research related to social network analytics. When we acknowledge the research in social media network analysis dates back to the 1930s. The application that uses data mining techniques developed by industry and academia are already being used commercially. For example, a "Social Media Analytics" organization offers services to us and track social media to provide customers data about how goods and services recognized and discussed through social media networks. Analysts in the organizations have applied text mining algorithms, and detect the propagation models to blogs to create techniques to understand better how data moves through the blogosphere.

Data mining techniques can be implemented to social media sites to comprehend information better and to make use of data for analytics, research, and business purposes. Representative Fields include a community or group detection, data diffusion, propagation of audiences, subject detection and tracking, individual behavior analysis, group behavior analysis, and market research for organizations.

## Representation of Data

Similar to other social media data, it is accepted to use a graph representation to study social media data sets. A graph comprises a set including vertexes (nodes) and edges (links). Users are usually shown as the nodes in the graph. Relationships or corporation between individuals (nodes) is shown as the links in the graph.

The graph depiction is common for information extracted from social networking sites where people interact with friends, family, and business associates. It helps to create a social network of friends, family, or business associates. Less apparent is how the graph structure is applied to blogs, wikis, opinion mining, and similar types of online social media platforms.

If we consider blogs, One graph representation blogged as nodes and can be regarded as "blog network," and another graph description has blog posts as the nodes, and can be regarded as "post-network." Edges are created in a blog post network when another blog post references another blog post. Other techniques used to represent blog networks concurrently account for individuals, relationships, content, and time simultaneously- called Internet Online Analytical Processing

(iOLAP). Wikis can be considered from the context of depicting authors as nodes, and edges are created when the authors contribute to an object.

The graphical representation allows the application of classic mathematical graph theory, traditional techniques of analyzing social media platforms and work on mining graph data. The probably big size of the graph used to depict social media platforms can present difficulties for automated processing as restricts on computer memory. The processing speeds are maximized and usually exceeded when trying to cope with huge social media data set. Other challenges to implementing automated procedures to allow social media data mining include identifying and dealing with spam, the variety of formats used in the same subcategory of social media, and continuously altering content and structure.

## Data Mining- Social Context

No matter what sort of social media is being studied, some fundamental things are essential to consider the most meaningful outcomes are feasible. Every kind of social media and every data mining purpose applied to social media may involve distinctive methods and algorithms to produce an advantage from data mining. Various data sets and data issues include different kinds of tools. If it is known how to organize the data, a classification tool might be appropriate. If we understand what the data is about, but cannot determine trends and patterns in the data, the use of a clustering tool may be the best.

The problem itself can conclude the best approach. There is no other option for understanding the data as much possible before applying data mining techniques as well as understanding the various data mining tools that are available. A subject analyst might be required to help better understand the data set. To better understand the various tools available for data mining, there are a host of data mining and machine learning text and different resources that are available to support more accurate information about a variety of particular data mining techniques and algorithms.

Once you understand the issues and select an appropriate data mining approach, consider any preprocessing that needs to be done. A systematic process may also be required to develop an adequate set of data to allow reasonable processing times. Pre-processing should include suitable privacy protection mechanisms. Although social media platforms incorporate huge amounts of

openly accessible data, it is important to guarantee individual rights, and social media platform copyrights are secured. The effect of spam should be considered along with the temporal representation.

In addition to preprocessing, it is essential to think about the effect of time. Depending upon the inquiry and the research, we may get different outcomes at one time compared to another, although the time segment is an accessible consideration for specific areas. For example, subject detection, influence propagation, and network development, less evident is the effect of time on network identification, group behavior, and marketing. What defines a network at one point in time can be significantly different at another point in time. Group behavior and interests will change after some time, and what was offered to the individuals or groups today may not be trendy tomorrow.

With data depicted as a graph, the tasks start with a selected number of nodes, known as seeds. Graphs are traversed, starting with the arrangement of seeds, and as the link structure from the seed nodes is used, data is collected, and the structure itself is also reviewed. Utilizing the link structure to stretch out from the seed set and gather new information is known as crawling the network. The application and algorithms that are executed as a crawler should effectively manage the challenges present in powerful social media platforms such as restricted sites, format changes, and structure errors (invalid links). As the crawler finds the new data, it stores the new data in a repository for further analysis. As link data is found, the crawler updates the data about the network structure.

Some social media platforms such as Facebook, Twitter, and Technorati provide Application Programmer Interfaces (APIs) that allow crawler applications to interact with the data sources directly. However, these platforms usually restrict the number of API transactions per day, relying on the affiliation the API user has with the platform. For some platforms, it is possible to collect data (crawl) without utilizing APIs. Given the huge size of the social media platform data available, it might be necessary to restrict the amount of data that the crawler collects. When the crawler has collected the data, some postprocessing may be needed to validate and clean up the data. Traditional social media platforms analysis methods can be applied, for example, centrality measures and group structure studies. In many cases, additional data will be related to a node or a link opening opportunities for more complex methods to consider the more thoughtful semantics that can be exposed with text and data mining techniques.

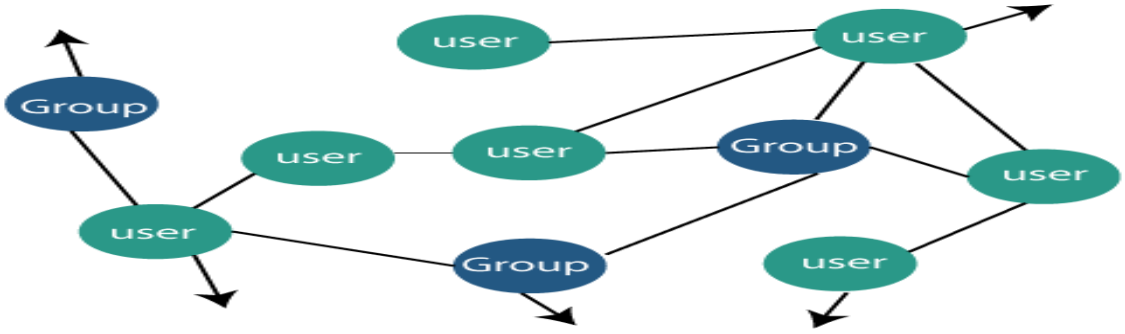
We now focus on two particular social media platform data to further represent how data mining techniques are applied to social media sites. The two major areas are social media platforms, and Blogs are powerful, and rich data sources portray both these areas. The two areas offer potential value to the more extensive scientific network as well as a business organization.

Social media platforms: Illustrative Examples

Social media platforms like Facebook or LinkedIn comprises of connected users with unique profiles. Users can interact with their friends and family and can share news, photos, story, videos, favorite links, etc. Users have an option to customize their profiles relying on individual preferences, but some common data may incorporate relationship status, birthday, an Email address, and hometown. Users have alternatives to choose how much data they include in their profile and who has access to it. The amount of data accessible via social media platforms have raised security concerns and is a related societal issue.

Here, the figure illustrates the hypothetical graph structure diagram for typical social media platforms, and Arrows indicate links to a larger part of the graph.

It is important to secure personal identity when working with social media platforms data. Recent reports highlight the need to secure privacy as it has been demonstrated that even anonymizing this sort of data can still reveal individual data when advanced data analysis strategies are utilized. Security settings also can restrict the ability of data mining applications to think about each data on social media platforms. However, some heinous techniques can be utilized to take over the security settings.



# Data Mining Interface

The Data Mining Interface (DMI) is a Web-based, interactive, dynamic report-building module. DMI reports immediately access current data and refresh automatically when new data becomes available. Trending and baseline data is also available for customized reports. Trending data is transparently used when necessary, while baseline data is mixed with current data on the same screen.

DMI reports have variable time-range settings, variable resolution settings, and dynamic sorting and filtering mechanisms. Trending and baseline data is also available for DMI reports. Trending data is transparently used when necessary, while baseline data is mixed with current data on the same screen.

Use DMI to generate tabular reports and charts and mix multiple report sections on the same page. The reports can have a hierarchical structure with contextual drilldown, sibling, and parent reports.

Report definitions are saved in the database and reports are re-run when opened. The DMI is equipped with an integrated persistent report cache that optimizes report re-run requests in the context of real-time data changes in the database.

The DMI integrates with a Data Center Real User Monitoring database, providing access restrictions, based on the Data Center Real User Monitoring user identity. Predefined DMI reports are available for various types of users and include high-level scorecards for IT executives, and dedicated planning and monitoring reports for staff responsible for application service delivery.

The DMI can also be integrated with VantageView and used as the custom reporting engine.

DMI uses product-specific data views. Each data view supports its own set of dimensions and metrics.

## User interface

User interface is the module of data mining system that helps the communication between users and the data mining system. User Interface allows the following functionalities −

- Interact with the system by specifying a data mining query task.

- Providing information to help focus the search.
- Mining based on the intermediate data mining results.
- Browse database and data warehouse schemas or data structures.
- Evaluate mined patterns.
- Visualize the patterns in different forms.

# Data Mining Approaches:

## 1. Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

### Tree Pruning Approaches

Here is the Tree Pruning Approaches listed below −

- **Pre-pruning** − The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully grown tree.

### Cost Complexity

The cost complexity is measured by the following two parameters −

- Number of leaves in the tree, and
- Error rate of the tree.

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

## Rule Pruning

The rule is pruned is due to the following reason −

- The Assessment of quality is made on the original set of training data. The rule may perform well on training data but less well on subsequent data. That's why the rule pruning is required.
- The rule is pruned by removing conjunct. The rule R is pruned, if pruned version of R has greater quality than what was assessed on an independent set of tuples.

FOIL is one of the simple and effective method for rule pruning. For a given rule R,

FOIL_Prune = pos - neg / pos + neg

where pos and neg is the number of positive tuples covered by R, respectively.

**Note** − This value will increase with the accuracy of R on the pruning set. Hence, if the FOIL_Prune value is higher for the pruned version of R, then we prune R.

Here we will discuss other classification methods such as Genetic Algorithms, Rough Set Approach, and Fuzzy Set Approach.

## Genetic Algorithms

The idea of genetic algorithm is derived from natural evolution. In genetic algorithm, first of all, the initial population is created. This initial population consists of randomly generated rules. We can represent each rule by a string of bits.

For example, in a given training set, the samples are described by two Boolean attributes such as A1 and A2. And this given training set contains two classes such as C1 and C2.

We can encode the rule **IF A1 AND NOT A2 THEN C2** into a bit string **100**. In this bit representation, the two leftmost bits represent the attribute A1 and A2, respectively.

Likewise, the rule **IF NOT A1 AND NOT A2 THEN C1** can be encoded as **001**.

**Note** − If the attribute has K values where K>2, then we can use the K bits to encode the attribute values. The classes are also encoded in the same manner.

Points to remember −

- Based on the notion of the survival of the fittest, a new population is formed that consists of the fittest rules in the current population and offspring values of these rules as well.
- The fitness of a rule is assessed by its classification accuracy on a set of training samples.
- The genetic operators such as crossover and mutation are applied to create offspring.
- In crossover, the substring from pair of rules are swapped to form a new pair of rules.
- In mutation, randomly selected bits in a rule's string are inverted.

## 2. Rough Set Approach

**We can use the rough set approach to discover** structural relationship within imprecise and noisy data.

**Note** − This approach can only be applied on discrete-valued attributes. Therefore, continuous-valued attributes must be discretized before its use.

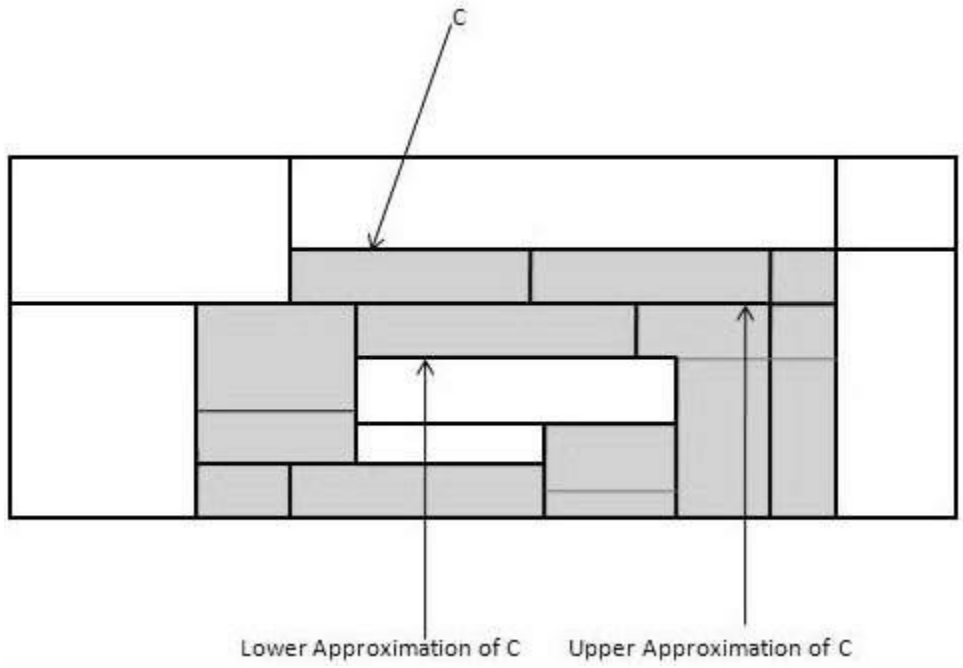
The Rough Set Theory is based on the establishment of equivalence classes within the given training data. The tuples that forms the equivalence class are indiscernible. It means the samples are identical with respect to the attributes describing the data.

There are some classes in the given real world data, which cannot be distinguished in terms of available attributes. We can use the rough sets to **roughly** define such classes.

For a given class C, the rough set definition is approximated by two sets as follows −

- **Lower Approximation of C** − The lower approximation of C consists of all the data tuples, that based on the knowledge of the attribute, are certain to belong to class C.
- **Upper Approximation of C** − The upper approximation of C consists of all the tuples, that based on the knowledge of attributes, cannot be described as not belonging to C.

The following diagram shows the Upper and Lower Approximation of class C:




## 3. Fuzzy Set Approaches

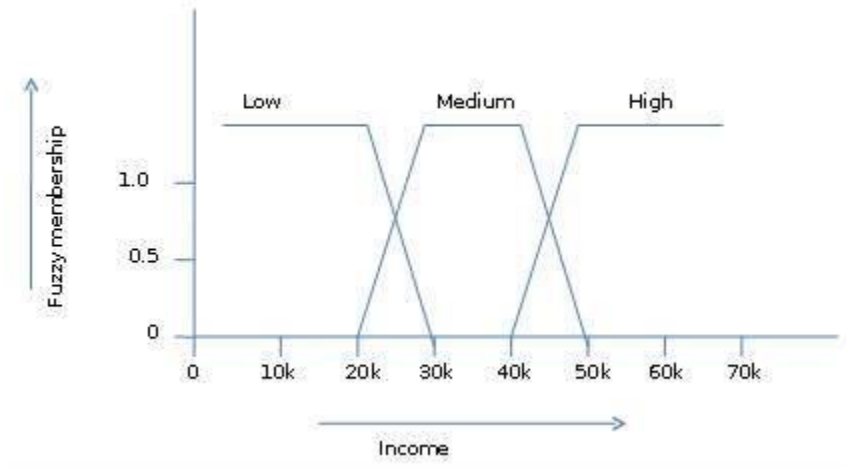
Fuzzy Set Theory is also called Possibility Theory. This theory was proposed by Lotfi Zadeh in 1965 as an alternative the **two-value logic** and **probability theory**. This theory allows us to work at a high level of abstraction. It also provides us the means for dealing with imprecise measurement of data.

The fuzzy set theory also allows us to deal with vague or inexact facts. For example, being a member of a set of high incomes is in exact (e.g. if $50,000 is high then what about $49,000 and $48,000). Unlike the traditional CRISP set where the element either belong to S or its complement but in fuzzy set theory the element can belong to more than one fuzzy set.

For example, the income value $49,000 belongs to both the medium and high fuzzy sets but to differing degrees. Fuzzy set notation for this income value is as follows −

$m_{medium\_income}(\$49k)=0.15$ and $m_{high\_income}(\$49k)=0.96$

where 'm' is the membership function that operates on the fuzzy sets of medium_income and high_income respectively. This notation can be shown diagrammatically as follows −




---

### 4. Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

### 5. Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

## Data Mining Methodologies

Data mining is the process of looking at large banks of information to generate new information. Intuitively, you might think that data “mining” refers to the extraction of new data, but this isn’t the case; instead, data mining is about extrapolating patterns and new knowledge from the data you’ve already collected.

Relying on techniques and technologies from the intersection of database management, statistics, and machine learning, specialists in data mining have dedicated their careers to better understanding how to process and draw conclusions from vast amounts of information.

**1. Tracking patterns.** One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

**2. Classification.** Classification is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

**3. Association**. Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

**4. Outlier detection.** In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

**5. Clustering.** Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics

of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

**6. Regression.** Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

**7. Prediction.** Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future.

# Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

## Data Integration

Data Integration is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store. Data integration may involve inconsistent data and therefore needs data cleaning.

## Data Cleaning

Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse.

## Data Selection

Data Selection is the process where data relevant to the analysis task are retrieved from the database. Sometimes data transformation and consolidation are performed before the data selection process.

## 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.

## Data Transformation

In this step, data is transformed or consolidated into forms appropriate for mining, by performing summary or aggregation operations.

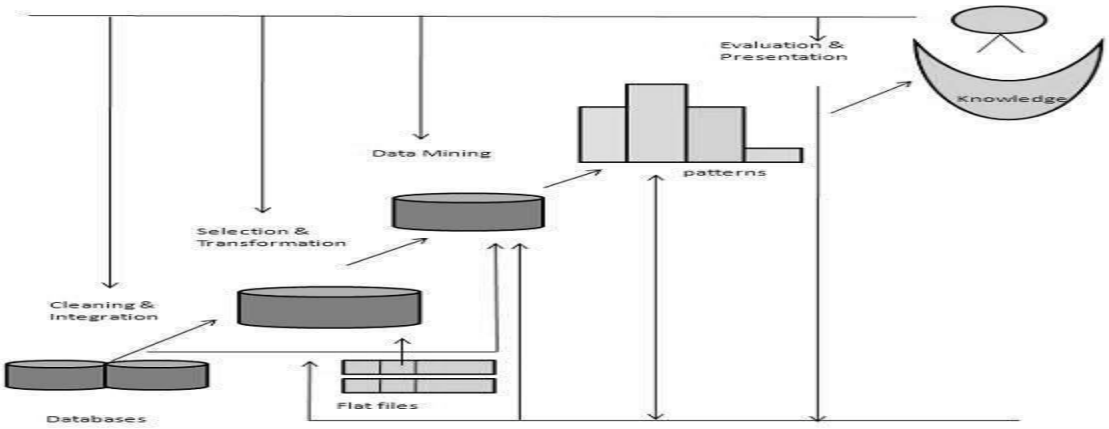
## Knowledge Discovery

Some people don't differentiate data mining from knowledge discovery 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.

The following diagram shows the process of knowledge discovery –




1. Real world data are generally
    - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - Noisy: containing errors or outliers
    - Inconsistent: containing discrepancies in codes or names
2. Tasks in data preprocessing
    - Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
    - Data integration: using multiple databases, data cubes, or files.
    - Data transformation: normalization and aggregation.
    - Data reduction: reducing the volume but producing the same or similar analytical results.
    - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

## Data cleaning

1. Fill in missing values (attribute or class value):
    - Ignore the tuple: usually done when class label is missing.

- Use the attribute mean (or majority nominal value) to fill in the missing value.
- Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
- Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:
    - Binning
        - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
        - Then smooth by bin means, bin median, or bin boundaries.
    - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
    - Regression: smooth by fitting the data into regression functions.
3. Correct inconsistent data: use domain knowledge or expert decision.

## Data transformation

1. Normalization:
    - Scaling attribute values to fall within a specified range.
        - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
    - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev
2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

## Data reduction

1. Reducing the number of attributes
    - Data cube aggregation: applying roll-up, slice or dice operations.

- Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).
- Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..

2. Reducing the number of attribute values
   - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
   - Clustering: grouping values in clusters.
   - Aggregation or generalization
3. Reducing the number of tuples
   - Sampling

## Discretization and generating concept hierarchies

1. **Unsupervised discretization** - class variable is not used.
   - Equal-interval (equiwidth) binning: split the whole range of numbers in intervals with equal size.
   - Equal-frequency (equidepth) binning: use intervals containing equal number of values.
2. **Supervised discretization** - uses the values of the class variable.
   - Using class boundaries. Three steps:
     - Sort values.
     - Place breakpoints between values belonging to different classes.
     - If too many intervals, merge intervals with equal or similar class distributions.
   - **Entropy (information)-based discretization.** Example:
     - Information in a class distribution:
       - Denote a set of five values occurring in tuples belonging to two classes (+ and -) as [+,+,+,-,-]
       - That is, the first 3 belong to "+" tuples and the last 2 - to "-" tuples
       - Then, Info([+,+,+,-,-]) = -(3/5)*log(3/5)-(2/5)*log(2/5) (logs are base 2)
       - 3/5 and 2/5 are relative frequencies (probabilities)

        - Ignoring the order of the values, we can use the following notation: [3,2] meaning 3 values from one class and 2 - from the other.
        - Then, Info([3,2]) = -(3/5)*log(3/5)-(2/5)*log(2/5)
    - Information in a split (2/5 and 3/5 are weight coefficients):
        - Info([+,+],[+,-,-]) = (2/5)*Info([+,+]) + (3/5)*Info([+,-,-])
        - Or, Info([2,0],[1,2]) = (2/5)*Info([2,0]) + (3/5)*Info([1,2])
    - **Method:**
        - Sort the values;
        - Calculate information in all possible splits;
        - Choose the split that minimizes information;
        - Do not include breakpoints between values belonging to the same class (this will increase information);
        - Apply the same to the resulting intervals until some stopping criterion is satisfied.
3. **Generating concept hierarchies:** recursively applying partitioning or discretization methods.

# Data Mining Technologies

1. **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

2. **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

3. **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include

Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

4. **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

5. **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

6. **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

# Clustering Analysis

Clustering is the process of making a group of abstract objects into classes of similar objects.

**Points to Remember**

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

## Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

## Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining −

- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

## Clustering Methods

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

## Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

**Points to remember −**

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

## Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

## Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

## Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

## Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

## Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

## Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

**Advantage**

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

## Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

## Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured.

For example, a document may contain a few structured fields, such as title, author, publishing_date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining.