# B.C.A. Semester – 4

## BCA-404

# Data Mining & Data Ware Housing

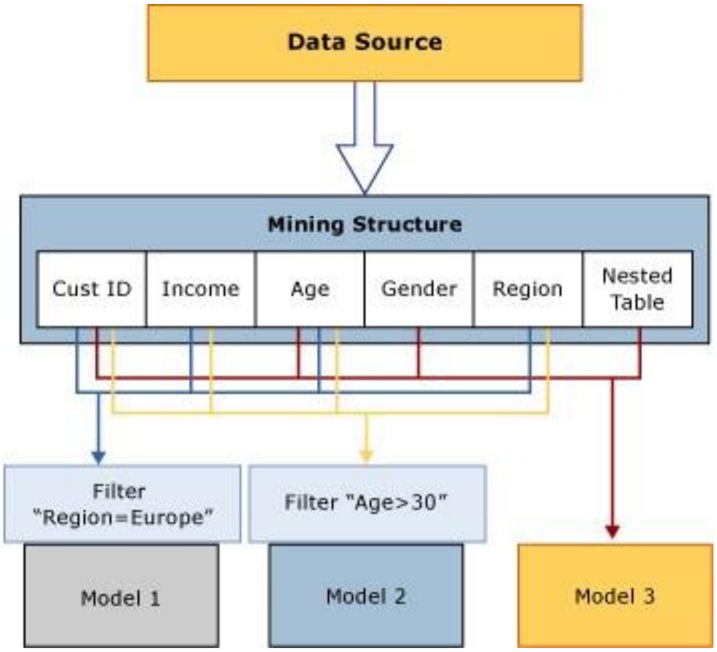
# UNIT - 4

## Introduction to DMDH

- Mining object :-

A data mining object is only an empty container until it has been processed. Processing a data mining model is also called training. Processing mining structures: A mining structure gets data from an external data source, as defined by the column bindings and usage metadata, and reads the data.

The following diagram illustrates the flow of data when a mining structure is processed, and when a mining model is processed.



- Processing mining structures:

A mining structure gets data from an external data source, as defined by the column bindings and usage metadata, and reads the data. This data is read in full and then analyzed to extract various statistics. Analysis Services stores a compact representation of the data, which is suitable for analysis by data mining algorithms, in a local cache. You can either keep this cache or delete it after your models have been processed. By default, the cache is stored.

- Processing mining models:

A mining model is empty, containing definitions only, until it is processed. To process a mining model, the mining structure that it is based on must have been processed. The mining model gets the data from the mining structure cache, applies any filters that may have been created on the model, and then passes the data set through the algorithm to detect patterns. After the model is processed, the model stores only the results of processing, not the data itself.

The following diagram illustrates the flow of data when a mining structure is processed, and when a mining model is processed.

# What is Spatial Data Mining?

A spatial database saves a huge amount of space-related data, including maps, preprocessed remote sensing or medical imaging records, and VLSI chip design data. Spatial databases have several features that distinguish them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands the unification of data mining with spatial database technologies.

It can be used for learning spatial records, discovering spatial relationships and relationships among spatial and  records, constructing spatial knowledge bases, reorganizing spatial databases, and optimizing spatial queries.

It is expected to have broad applications in geographic data systems, marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other areas where spatial data are used.

A central challenge to spatial data mining is the exploration of efficient spatial data mining techniques because of the large amount of spatial data and the difficulty of spatial data types and spatial access methods. Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information.

The term  is often associated with continuous geographic space, whereas the term spatial statistics is often associated with discrete space. In a statistical model that manages non-spatial records, one generally considers statistical independence among different areas of data.

There is no such separation among spatially distributed records because, actually spatial objects are interrelated, or more exactly spatially co-located, in the sense that the closer the two objects are placed, the more likely they send the same properties. For example, natural resources, climate, temperature, and economic situations are likely to be similar in geographically closely located regions.

Such a property of close interdependency across nearby space leads to the notion of spatial autocorrelation. Based on this notion, spatial statistical modeling methods have been developed with success. Spatial data mining will create spatial statistical analysis methods and extend them for large amounts of spatial data, with more emphasis on effectiveness, scalability, cooperation with database and data warehouse systems, enhanced user interaction, and the discovery of new kinds of knowledge.

# What is Multimedia Data Mining?

Multimedia mining is a subfield of data mining that is used to find interesting information of implicit knowledge from multimedia databases. Mining in multimedia is referred to as automatic annotation or annotation mining. Mining multimedia data requires two or more data types, such as text and video or text video and audio.

Multimedia data mining is an interdisciplinary field that integrates image processing and understanding, computer vision, data mining, and pattern recognition. Multimedia data mining discovers interesting patterns from multimedia databases that store and manage large collections of multimedia objects, including image data, video data, audio data, sequence data and hypertext data containing text, text markups, and linkages. Issues in multimedia data mining include *content-based retrieval and similarity search, generalization and multidimensional analysis*. Multimedia data cubes contain additional dimensions and measures for multimedia information.

The framework that manages different types of multimedia data stored, delivered, and utilized in different ways is known as a multimedia database management system. There are three classes of multimedia databases: static, dynamic, and dimensional media. The content of the Multimedia Database management system is as follows:

- **Media data:**The actual data representing an object.
- **Media format data:** Information such as sampling rate, resolution, encoding scheme etc., about the format of the media data after it goes through the acquisition, processing and encoding phase.
- **Media keyword data:**Keywords description relating to the generation of data. It is also known as content descriptive data. Example: date, time and place of recording.
- **Media feature data:** Content dependent data such as the distribution of colours, kinds of texture and different shapes present in data.

## Types of Multimedia Applications

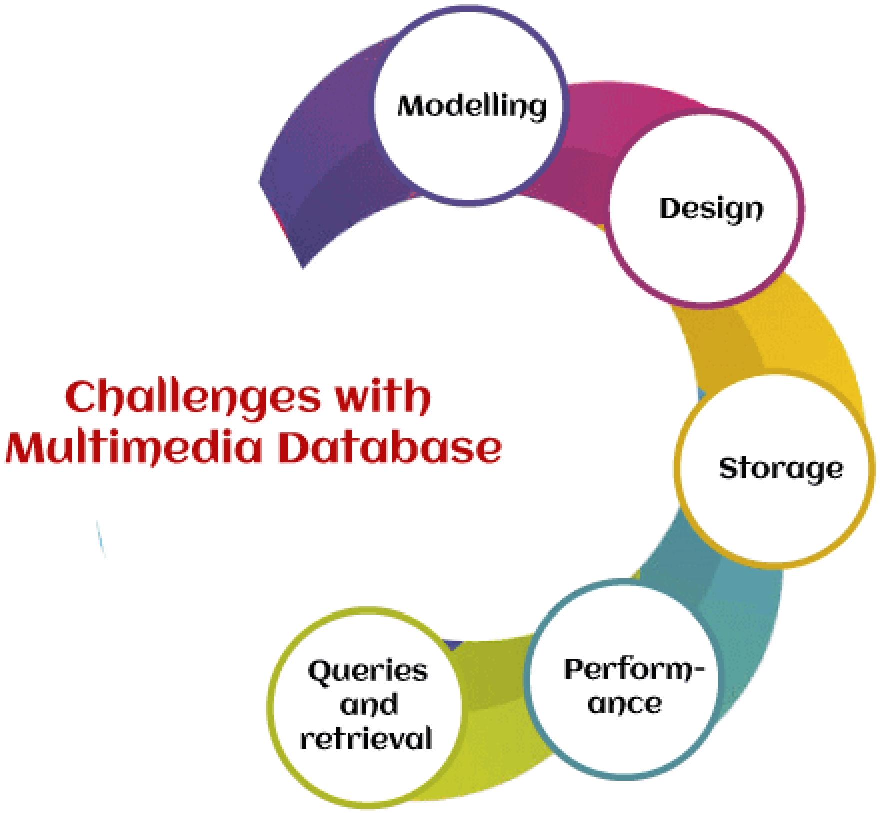
Types of multimedia applications based on data management characteristics are:

1. **Repository applications:** A Large amount of multimedia data and meta-data (Media format date, Media keyword data, Media feature data) that is stored for retrieval purposes, e.g., Repository of satellite images, engineering drawings, radiology scanned pictures.

2. **Presentation applications:** They involve delivering multimedia data subject to temporal constraints. Optimal viewing or listening requires DBMS to deliver data at a certain rate, offering the quality of service above a certain threshold. Here data is processed as it is delivered. Example: Annotating of video and audio data, real-time editing analysis.
3. **Collaborative work using multimedia information** involves executing a complex task by merging drawings and changing notifications. Example: Intelligent healthcare network.

## Challenges with Multimedia Database

There are still many challenges to multimedia databases, such as:



1. **Modelling:** Working in this area can improve database versus information retrieval techniques; thus, documents constitute a specialized area and deserve special consideration.
2. **Design:**The conceptual, logical and physical design of multimedia databases has not yet been addressed fully as performance and tuning issues at each level are far more complex as they consist of a variety of formats like JPEG, GIF, PNG, MPEG, which is not easy to convert from one form to another.

3. **Storage:**Storage of multimedia database on any standard disk presents the problem of representation, compression, mapping to device hierarchies, archiving and buffering during input-output operation. In DBMS, a BLOB (Binary Large Object) facility allows untyped bitmaps to be stored and retrieved.
4. **Performance:** Physical limitations dominate an application involving video playback or audio-video synchronization. The use of parallel processing may alleviate some problems, but such techniques are not yet fully developed. Apart from this, a multimedia database consumes a lot of processing time and bandwidth.
5. **Queries and retrieval:** For multimedia data like images, video, and audio accessing data through query open up many issues like efficient query formulation, query execution and optimization, which need to be worked upon.

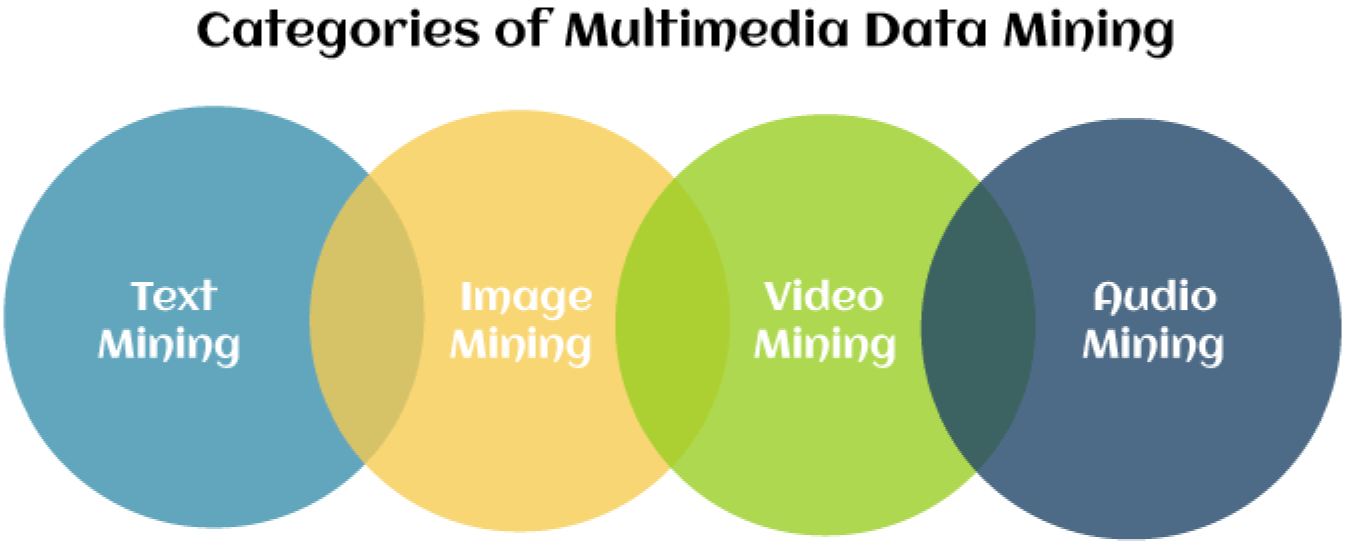
## Where is Multimedia Database Applied?

Below are the following areas where a multimedia database is applied, such as:

- **Documents and record management:** Industries and businesses keep detailed records and various documents. For example, insurance claim records.
- **Knowledge dissemination:**Multimedia database is a very effective tool for knowledge dissemination in terms of providing several resources. For example, electronic books.
- **Education and training:**Computer-aided learning materials can be designed using multimedia sources which are nowadays very popular sources of learning. Example: Digital libraries.
- **Travelling:** Marketing, advertising, retailing, entertainment and travel. For example, a virtual tour of cities.
- **Real-time control and monitoring:** With active database technology, multimedia presentation of information can effectively monitor and control complex tasks. For example, manufacturing operation control.

## Categories of Multimedia Data Mining

Multimedia mining refers to analyzing a large amount of multimedia information to extract patterns based on their statistical relationships. Multimedia data mining is classified into two broad categories: static and dynamic media. ***Static media*** contains text (digital library, creating SMS and MMS) and images (photos and medical images). ***Dynamic media*** contains Audio (music and MP3 sounds) and Video (movies). The below image shows the categories of multimedia data mining.



**1. Text Mining**

Text is the foremost general medium for the proper exchange of information. Text Mining evaluates a huge amount of usual language text and detects exact patterns to find useful information. Text Mining also referred to as text data mining, is used to find meaningful information from unstructured texts from various sources.

**2. Image Mining**

Image mining systems can discover meaningful information or image patterns from a huge collection of images. Image mining determines how low-level pixel representation consists of a raw image or image sequence that can be handled to recognize high-level spatial objects and relationships. It includes digital image processing, image understanding, database, AI, etc.

**3. Video Mining**

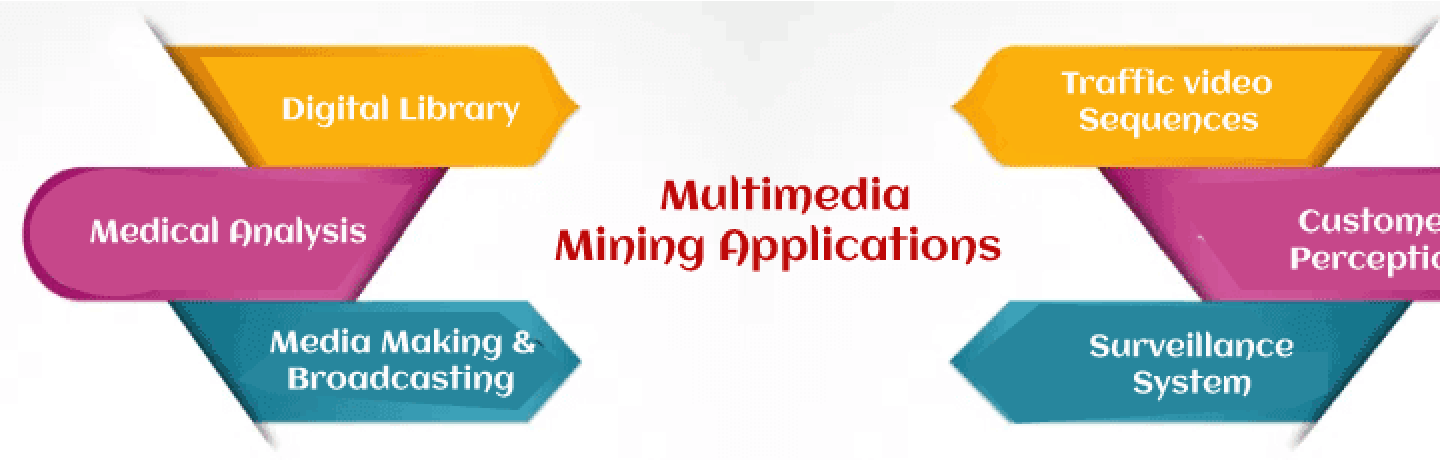
Video mining is unsubstantiated to find interesting patterns from many video data; multimedia data is video data such as text, image, metadata, visuals and audio. It is commonly used in security and surveillance, entertainment, medicine, sports and education programs. The processing is indexing, automatic segmentation, content-based retrieval, classification and detecting triggers.

**4. Audio Mining**

Audio mining plays an important role in multimedia applications, is a technique by which the content of an audio signal can be automatically searched, analyzed and rotten with wavelet transformation. It is generally used in automatic speech recognition, where the analysis efforts to find any speech within the audio. Band energy, frequency centroid, zero-crossing rate, pitch period and bandwidth are often used for audio processing.

## Application of Multimedia Mining

There are different kinds of applications of multimedia data mining, some of which are as follows:

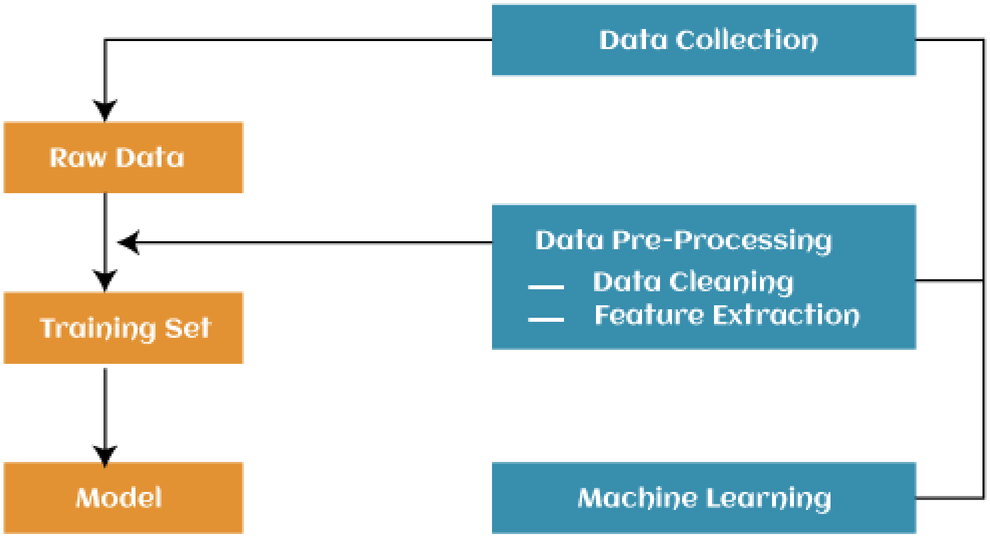


- **Digital Library:** The collection of digital data is stored and maintained in a digital library, which is essential to convert different digital data formats into text, images, video, audio, etc.
- **Traffic Video Sequences:** To determine important but previously unidentified knowledge from the traffic video sequences, detailed analysis and mining are to be performed based on vehicle identification, traffic flow, and queue temporal relations of the vehicle at an intersection. This provides an economic approach for regular traffic monitoring processes.
- **Medical Analysis:** Multimedia mining is primarily used in the medical field, particularly for analyzing medical images. Various data mining techniques are used for image classification. Examples, Automatic 3D delineation of highly aggressive brain tumours, Automatic localization and identification of vertebrae in 3D CT scans, MRI Scans, ECG and X-Ray.

- **Customer Perception:** It contains details about customers' opinions, products or services, customers complaints, customers preferences, and the level of customer satisfaction with products or services, which are collected together. The audio data serve as topic detection, resource assignment and evaluation of the quality of services. Many companies have call centres that receive telephone calls from customers.
- **Media Making and Broadcasting:** Radio stations and TV channels create broadcasting companies, and multimedia mining can be applied to monitor their content to search for more efficient approaches and improve their quality.
- **Surveillance system:** It consists of collecting, analyzing, summarizing audio, video or audiovisual information about specific areas like government organizations, multi-national companies, shopping malls, banks, forests, agricultural areas and, highways etc. The main use of this technology in the field of security; hence it can be utilized by military, police and private companies since they provide security services.

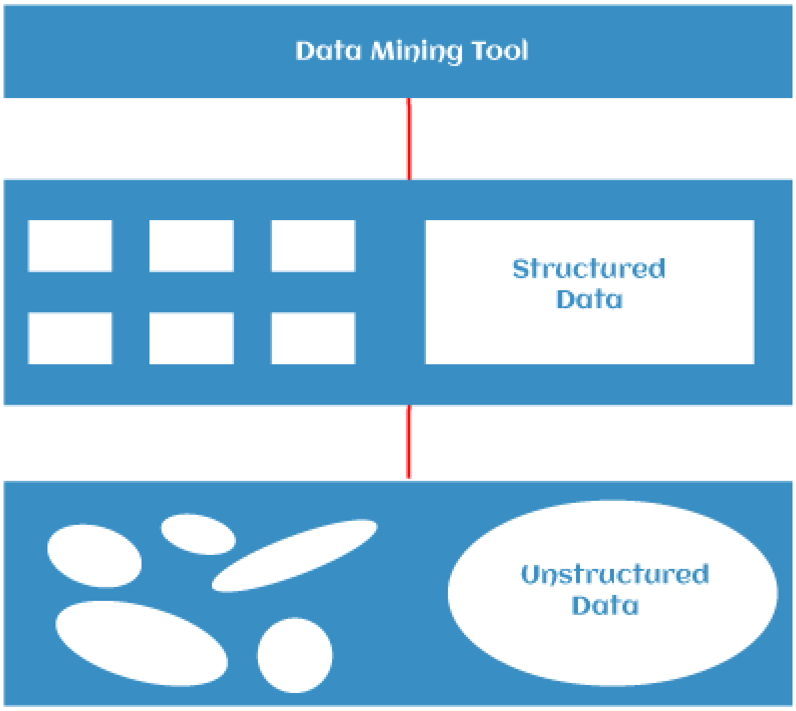
## Process of Multimedia Data Mining

The below image shows the present architecture, which includes the types of the multimedia mining process. Data Collection is the initial stage of the learning system; Pre-processing is to extract significant features from raw data. It includes data cleaning, transformation, normalization, feature extraction, etc. Learning can be direct if informative types can be recognized at preprocessing stage. The complete process depends extremely on the nature of raw data and the difficulty field. The product of preprocessing is the training set. A learning model must be selected for the specified training set to learn from it and make the multimedia model more constant.



Multimedia Mining Process

**Converting Un-structured data to structured data:** Data resides in a fixed field within a record or file is called structured data, and these data are stored in sequential form. Structured data has been easily entered, stored, queried and analyzed. Unstructured data is bitstream, for example, pixel representation for an image, audio, video and character representation for text. These files may have an internal structure, but they are still considered "unstructured" because their data does not fit neatly in a database. For example, images and videos of different objects have some similarities - each represents an interpretation of a building without a clear structure.

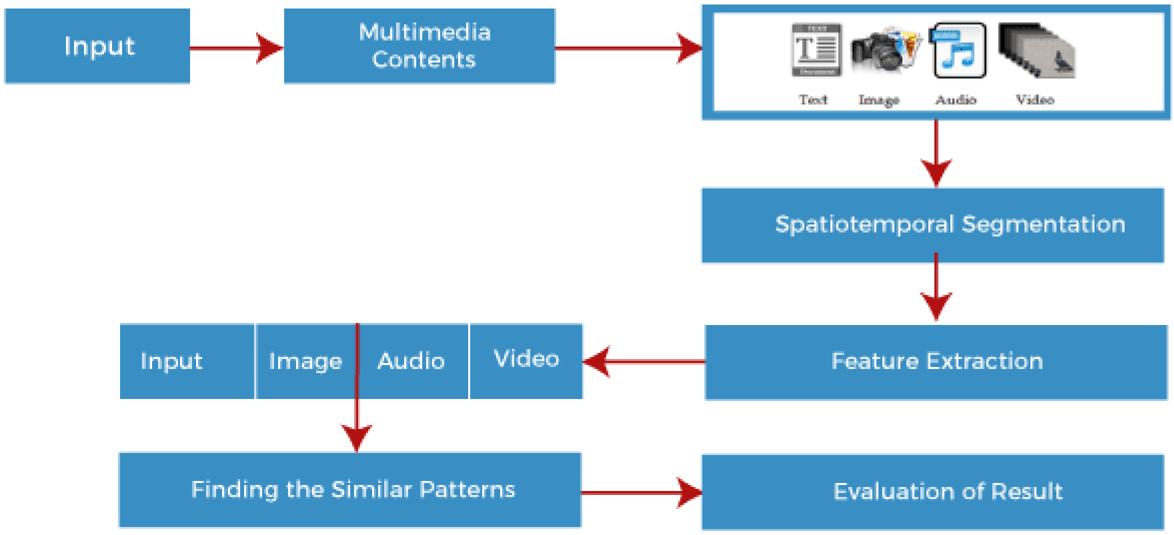


Unstruetured Data to Structured Data Conversion

Current data mining tools operate on structured data, which resides in a huge volume of the relational database, while data in multimedia databases are semi-structured or unstructured. Hence, the semi-structured or unstructured multimedia data is converted into structured one, and then the current data mining tools are used to extract the knowledge. The sequence or time element is different between unstructured and structured data mining. The architecture of converting unstructured data to structured data and which is used for extracting information from the unstructured database, is shown in the above image. Then data mining tools are applied to the stored structured databases.

## Architecture for Multimedia Data Mining

Multimedia mining architecture is given in the below image. The architecture has several components. Important components are Input, Multimedia Content, Spatiotemporal Segmentation, Feature Extraction, Finding similar Patterns, and Evaluation of Results.

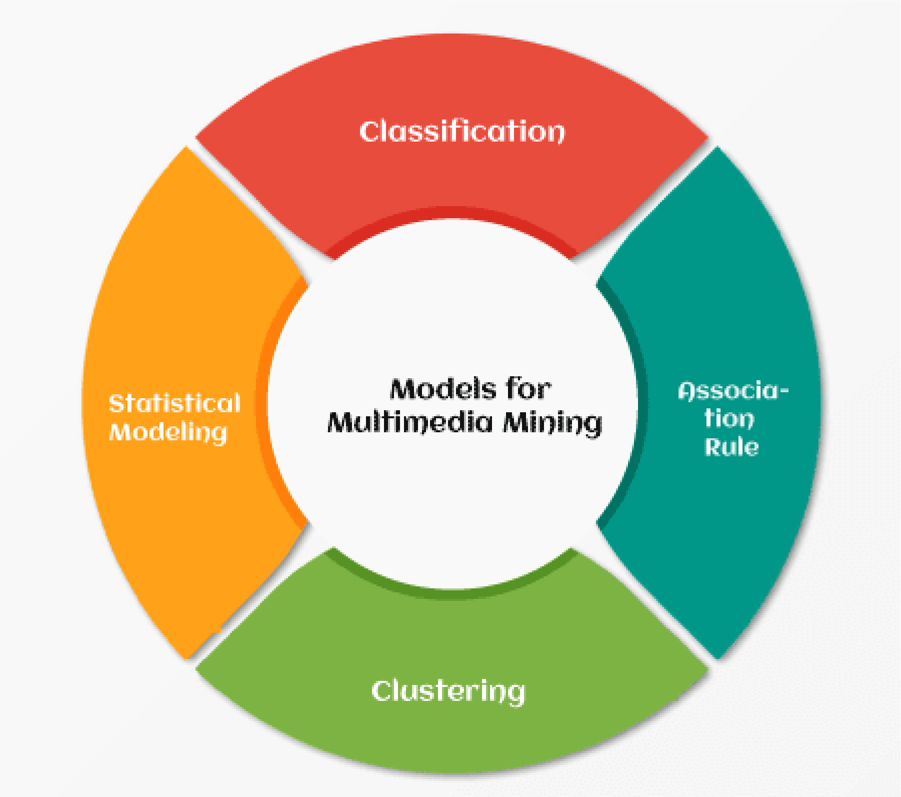


Multimedia Data mining Architecture

1. ***The input*** stage comprises a multimedia database used to find the patterns and perform the data mining.
2. ***Multimedia Content*** is the data selection stage that requires the user to select the databases, subset of fields, or data for data mining.
3. ***Spatio-temporal segmentation*** is nothing but moving objects in image sequences in the videos, and it is useful for object segmentation.
4. ***Feature extraction*** is the preprocessing step that involves integrating data from various sources and making choices regarding characterizing or coding certain data fields to serve when inputs to the pattern-finding stage. Such representation of choices is required because certain fields could include data at various levels and are not considered for finding a similar pattern stage. In MDM, the preprocessing stage is significant since the unstructured nature of multimedia records.
5. ***Finding a similar pattern*** stage is the heart of the whole data mining process. The hidden patterns and trends in the data are basically uncovered in this stage. Some approaches to finding similar pattern stages contain association, classification, clustering, regression, time-series analysis and visualization.
6. ***Evaluation of Results*** is a data mining process used to evaluate the results, and this is important to determine whether the prior stage must be revisited or not. This stage consists of reporting and using the extracted knowledge to produce new actions, products, services, or marketing strategies.

## Models for Multimedia Mining

The models which are used to perform multimedia data are very important in mining. Commonly four different multimedia mining models have been used. These are classification, association rule, clustering and statistical modelling.



1. **Classification:** Classification is a technique for multimedia data analysis that can learn from every property of a specified set of multimedia. It is divided into a predefined class label to achieve the purpose of classification. Classification is the process of constructing data into categories for its better effective and efficient use; it creates a function that well-planned data item into one of many predefined classes by inputting a training data set and building a model of the class attribute based on the rest of the attributes. Decision tree classification has a perceptive nature that the users conceptual model without loss of exactness. Hidden Markov Model is used to classify multimedia data such as images and videos as indoor-outdoor games.
2. **Association Rule:** Association Rule is one of the most important data mining techniques that help find relations between data items in huge databases. There are two types of associations in multimedia mining: image content and non-image content features.

Mining the frequently occurring patterns between different images becomes mining the repeated patterns in a set of transactions. Multi-relational association rule mining displays multiple reports for the same image. In image classification also, multiple-level association rule techniques are used.

3. **Clustering:** Cluster analysis divides the data objects into multiple groups or clusters. Cluster analysis combines all objects based on their groups. In multimedia mining, the clustering technique can be applied to group similar images, objects, sounds, videos and texts. Clustering algorithms can be divided into several methods: hierarchical methods, density-based methods, grid-based methods, model-based methods, k-means algorithms, and graph-based models.
4. **Statistical Modeling:** Statistical mining models regulate the statistical validity of test parameters and have been used to test hypotheses, undertake correlation studies, and transform and make data for further analysis. This is used to establish links between words and partitioned image regions to form a simple co-occurrence model.

## Issues in Multimedia Mining

Major Issues in multimedia data mining contains content-based retrieval, similarity search, dimensional analysis, classification, prediction analysis and mining associations in multimedia data.

### 1. Content-based retrieval and Similarity search

Content-based retrieval in multimedia is a stimulating problem since multimedia data is required for detailed analysis from pixel values. We considered two main families of multimedia retrieval systems, i.e. similarity search in multimedia data.

- ***Description-based retrieval system*** creates indices and object retrieval based on image descriptions, such as keywords, captions, size, and creation time.
- ***Content-based retrieval system*** supports image content retrieval, for example, colour histogram, texture, shape, objects, and wavelet transform.
- ***Use of content-based retrieval system:*** Visual features index images and promote object retrieval based on feature similarity; it is very desirable in various applications. These applications include diagnosis, weather prediction, TV production and internet search engines for pictures and e-commerce.

### 2. Multidimensional Analysis

To perform multidimensional analysis of large multimedia databases, multimedia data cubes may be designed and constructed similarly to traditional data cubes from relational data. A multimedia data cube has several dimensions. For example, the size of the image or video in bytes; the width and height of the frames, creating two dimensions, the date on which image or video was created or last modified, the format type of the image or video, frame sequence duration in seconds, Internet domain of pages referencing the image or video, the keywords like a colour dimension and edge orientation dimension. A multimedia data cube can have additional dimensions and measures for multimedia data, such as colour, texture, and shape.

The Multimedia data mining system prototype is MultiMediaMiner, the extension of the DBMiner system that handles multimedia data. The Image Excavator component of MultiMediaMiner uses image contextual information, like HTML tags on Web pages, to derive keywords. By navigating online directory structures, like Yahoo! directory, it is possible to build hierarchies of keywords mapped on the directories in which the image was found.

### 3. Classification and Prediction Analysis

Classification and predictive analysis has been used for mining multimedia data, particularly in scientific analysis like astronomy, seismology, and geoscientific analysis. Decision tree classification is an important method for reported image data mining applications. For example, consider the sky images, which astronomers have carefully classified as the training set. It can create models for recognizing galaxies, stars and further stellar objects based on properties like magnitudes, areas, intensity, image moments and orientation.

Image data mining classification and clustering are carefully connected to image analysis and scientific data mining. The image data are frequently in large volumes and need substantial processing power, such as parallel and distributed processing. Hence, many image analysis techniques and scientific data analysis methods could be applied to image data mining.

### 4. Mining Associations in Multimedia

Data Association rules involving multimedia objects have been mined in image and video databases. Three categories can be observed:

- Associations between image content and non-image content features

- Associations among image contents that are not related to spatial relationships
- Associations among image contents related to spatial relationships

First, an image contains multiple objects, each with various features such as colour, shape, texture, keyword, and spatial locations, so that many possible associations can be made. Second, a picture containing multiple repeated objects is essential in image analysis. The recurrence of similar objects should not be ignored in association analysis. Third, to find the associations between the spatial relationships and multimedia images can be used to discover object associations and correlations. With the associations between multimedia objects, we can treat every image as a transaction and find commonly occurring patterns among different images.

## What is Text Mining?

Text mining is a component of data mining that deals specifically with unstructured text data. It involves the use of natural language processing (NLP) techniques to extract useful information and insights from large amounts of unstructured text data. Text mining can be used as a preprocessing step for data mining or as a standalone process for specific tasks.

## Text Mining in Data Mining?

Text mining in data mining is mostly used for, the unstructured text data that can be transformed into structured data that can be used for data mining tasks such as classification, clustering, and association rule mining. This allows organizations to gain insights from a wide range of data sources, such as customer feedback, social media posts, and news articles.

## Why is Text Mining Important?

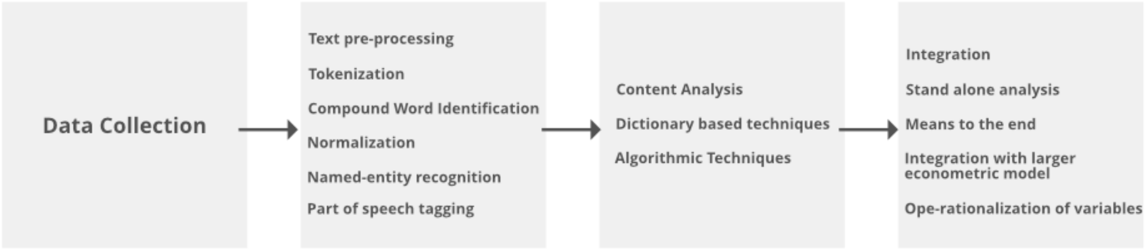
Text mining is widely used in various fields, such as natural language processing, information retrieval, and social media analysis. It has become an essential tool for organizations to extract insights from unstructured text data and make data-driven decisions.

“Extraction of interesting information or patterns from data in large databases is known as data mining.”

Text mining is a process of extracting useful information and nontrivial patterns from a large volume of text databases. There exist various strategies and devices to mine the text and find important data for the prediction and decision-making process. The selection of the right and accurate text mining procedure helps to enhance the speed and the time complexity also. This article briefly discusses and analyzes text mining and its applications in diverse fields.

As we discussed above, the size of information is expanding at exponential rates. Today all institutes, companies, different organizations, and business ventures are stored their information electronically. A huge collection of data is available on the internet and stored in digital libraries, database repositories, and other textual data like websites, blogs, social media networks, and e-mails. It is a difficult task to determine appropriate patterns and trends to extract knowledge from this large volume of data. Text mining is a part of Data mining to extract valuable text information from a text database repository. Text mining is a multi-disciplinary field based on data recovery, Data mining, AI,statistics, Machine learning, and computational linguistics.

# Text Mining Process



# Conventional Process of Text Mining

Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.

Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency in the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data.

Processing and controlling tasks are applied to review and further clean the data set.

Pattern analysis is implemented in Management Information System.

Information processed in the above steps is utilized to extract important and applicable data for a powerful and convenient decision-making process and trend analysis.

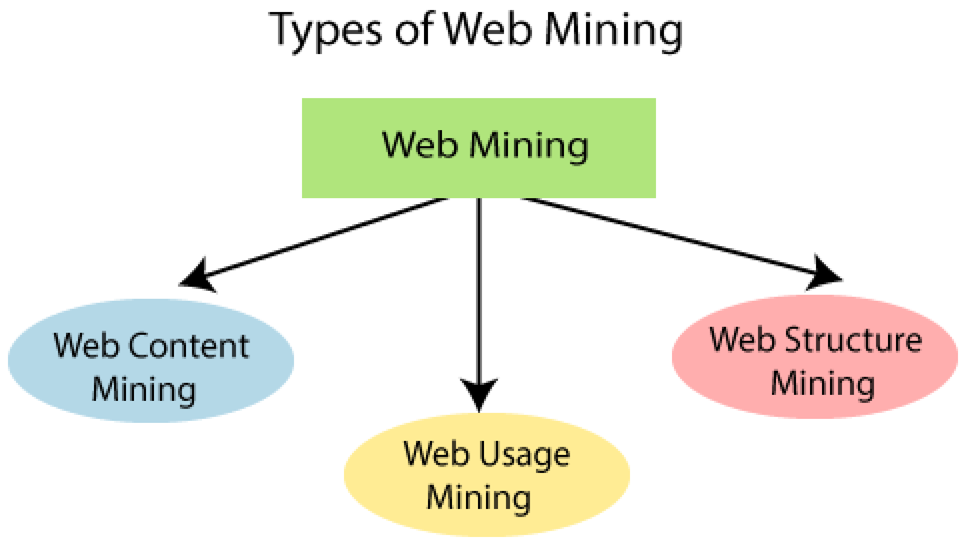
# Data Mining- World Wide Web

Over the last few years, the **World Wide Web** has become a significant source of information and simultaneously a popular platform for business. Web mining can define as the method of utilizing data mining techniques and algorithms to extract useful information directly from the web, such as Web documents and services, hyperlinks, Web content, and server logs. The World Wide Web contains a large amount of data that provides a rich source to data mining. The objective of Web mining is to look for patterns in Web data by collecting and examining data in order to gain insights.

## What is Web Mining?

Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a **knowledge discovery process**. Web mining has a distinctive property to provide a set of various data types. The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs. These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.

## There are three types of data mining:



### 1. Web Content Mining:

Web content mining can be used to extract useful data, information, knowledge from the web page content. In web content mining, each web page is considered as an individual document. The individual can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only the layout but also logical structure. The primary task of content mining is data extraction, where structured data is extracted from unstructured websites. The objective is to facilitate data aggregation over various web sites by using the extracted structured data. Web content mining can be utilized to distinguish topics on the web. For Example, if any user searches for a specific task on the search engine, then the user will get a list of suggestions.

### 2. Web Structured Mining:

The web structure mining can be used to find the link structure of hyperlink. It is used to identify that data either link the web pages or direct link network. In Web Structure Mining, an individual considers the web as a directed graph, with the web pages being the vertices that are associated with hyperlinks. The most important application in this regard is the Google search engine, which estimates the ranking of its outcomes primarily with the PageRank algorithm. It characterizes a page to be exceptionally relevant when frequently connected by other highly related pages. Structure and content mining methodologies are usually combined. For example, web structured mining can be beneficial to organizations to regulate the network between two commercial sites.

**3. Web Usage Mining:**

Web usage mining is used to extract useful data, information, knowledge from the weblog records, and assists in recognizing the user access patterns for web pages. In Mining, the usage of web resources, the individual is thinking about records of requests of visitors of a website, that are often collected as web server logs. While the content and structure of the collection of web pages follow the intentions of the authors of the pages, the individual requests demonstrate how the consumers see these pages. Web usage mining may disclose relationships that were not proposed by the creator of the pages.

Some of the methods to identify and analyze the web usage patterns are given below:

**I. Session and visitor analysis:**

The analysis of preprocessed data can be accomplished in session analysis, which incorporates the guest records, days, time, sessions, etc. This data can be utilized to analyze the visitor's behavior.

The document is created after this analysis, which contains the details of repeatedly visited web pages, common entry, and exit.

**II. OLAP (Online Analytical Processing):**

OLAP accomplishes a multidimensional analysis of advanced data.

OLAP can be accomplished on various parts of log related data in a specific period.

OLAP tools can be used to infer important business intelligence metrics

## Challenges in Web Mining:

The web pretends incredible challenges for resources, and knowledge discovery based on the following observations:

- **The complexity of web pages:**

The site pages don't have a unifying structure. They are extremely complicated as compared to traditional text documents. There are enormous amounts of documents in the digital library of the web. These libraries are not organized according to a specific order.

- **The web is a dynamic data source:**

The data on the internet is quickly updated. For example, news, climate, shopping, financial news, sports, and so on.

- **Diversity of client networks:**

The client network on the web is quickly expanding. These clients have different interests, backgrounds, and usage purposes. There are over a hundred million workstations that are associated with the internet and still increasing tremendously.

- **Relevancy of data:**

It is considered that a specific person is generally concerned about a small portion of the web, while the rest of the segment of the web contains the data that is not familiar to the user and may lead to unwanted results.

- **The web is too broad:**

The size of the web is tremendous and rapidly increasing. It appears that the web is too huge for data warehousing and data mining.

## Mining the Web's Link Structures to recognize Authoritative Web Pages:

The web comprises of pages as well as hyperlinks indicating from one to another page. When a creator of a Web page creates a hyperlink showing another Web page, this can be considered as the creator's authorization of the other page. The unified authorization of a given page by various creators on the web may indicate the significance of the page and may naturally prompt the discovery of authoritative web pages. The web linkage data provide rich data about the relevance, the quality, and structure of the web's content, and thus is a rich source of web mining.

## Application of Web Mining:

Web mining has an extensive application because of various uses of the web. The list of some applications of web mining is given below.

- Marketing and conversion tool
- Data analysis on website and application accomplishment.

- Audience behavior analysis
- Advertising and campaign accomplishment analysis.
- Testing and analysis of a site.

## Comparison Between Data mining and Web mining:

| Points | Data Mining | Web Mining |
|---|---|---|
| Definition | Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system. | Web Mining is the process of data mining techniques to automatically discover and extract information from web documents. |
| Application | Data Mining is very useful for web page analysis. | Web Mining is very useful for a particular website and e-service. |
| Target Users | Data scientist and data engineers. | Data scientists along with data analysts. |
| Access | Data Mining access data privately. | Web Mining access data publicly. |
| Structure | In Data Mining get the information from explicit structure. | In Web Mining get the information from structured, unstructured and semi-structured web pages. |
| Problem Type | Clustering, classification, regression, prediction, optimization and control. | Web content mining, Web structure mining. |
| Tools | It includes tools like machine learning algorithms. | Special tools for web mining are Scrapy, PageRank and Apache logs. |
| Skills | It includes approaches for data cleansing, machine learning algorithms. Statistics and probability. | It includes application level knowledge, data engineering with mathematical modules like statistics and probability. |

# • Application and trends in data mining

Data mining application

Data is a set of discrete objective facts about an event or a process that have little use by themselves unless converted into information. We have been collecting numerous data, from simple numerical measurements and text documents to more complex information such as spatial data, multimedia channels, and hypertext documents.

Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every year. An extracting data or seeking knowledge from this massive data, data mining techniques are used. Data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

Basically, the motive behind mining data, whether commercial or scientific, is the same – the need to find useful information in data to enable better decision-making or a better understanding of the world around us.

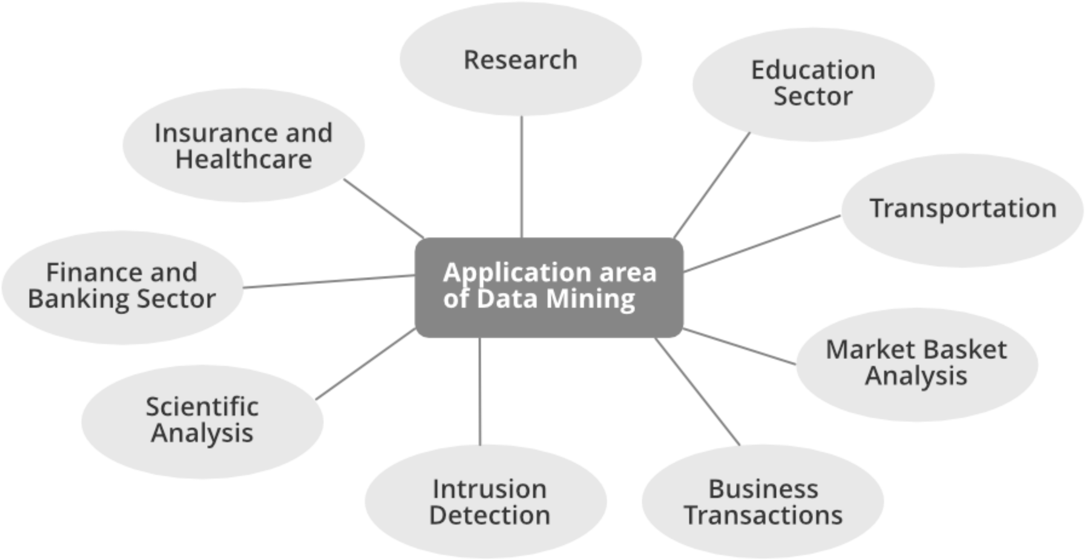
"Extraction of interesting information or patterns from data in large databases is known as data mining."

According to William J.Frawley "Data mining or KDD(Knowledge Discovery in Databases) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."

Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information. Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.

Data mining provides competitive advantages in the knowledge economy. It does this by providing the maximum knowledge needed to rapidly make valuable business decisions despite the enormous amounts of available data.

There are many measurable benefits that have been achieved in different application areas from data mining. So, let's discuss different applications of Data Mining:



Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

Sequence analysis in bioinformatics

Classification of astronomical objects

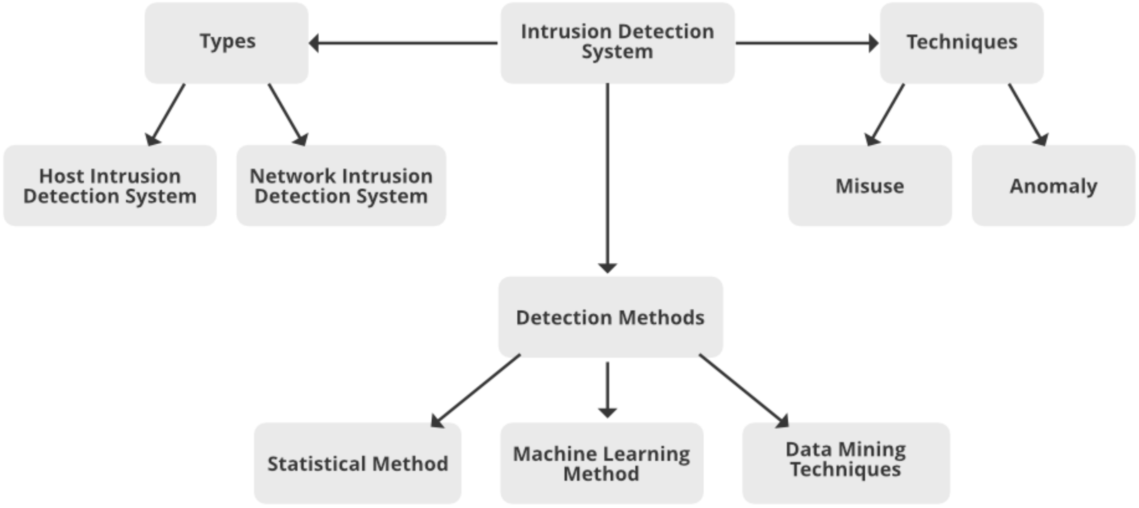
Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

Detect security violations

Misuse Detection

Anomaly Detection



Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

Direct mail targeting

Stock trading

Customer segmentation

Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.

Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.

Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

Predicting students admission in higher education

Predicting students profiling

Predicting student performance

Teachers teaching performance

Curriculum development

Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

Classification of uncertain data.

Information-based clustering.

Decision support system

Web Mining

Domain-driven data mining

IoT (Internet of Things)and Cybersecurity

Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

Claims analysis i.e which medical procedures are claimed together.

Identify successful medical therapies for different illnesses.

Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

Determine the distribution schedules among outlets.

Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

Credit card fraud detection.

Identify 'Loyal' customers.

Extraction of information related to customers.

Determine credit card spending by customer groups.

- ## What is Data Mining Trends and Research Frontiers?

Data mining is the process of analyzing a large size of information to find out the patterns, trends. It can be used for corporations to find out about customers' choices, make a good relationship with customers, increase the revenue, reduce risks. Data mining is based on complex algorithms that allow data segmentation to discover numerous trends and patterns, detect deviations, and estimate the likelihood of certain occurrences occurring. Raw data can be in both analog and digital formats, and it is essentially dependent on the data's source. Companies must keep up with the latest data mining trends and stay current in order to succeed in the industry and beat out the competition.

Types of Mining Sequence in Data Mining:

Mining time series

Mining symbolic sequence

Mining biological sequence

1. Mining Time Series

A specified number of data points are recorded at a specific time or events obtained over repeated measurements of time in a mining time series. The values or data are typically measured in equal time intervals like- hourly, weekly, daily. In time-series data is also recorded regular intervals or characteristic time-series components are trend, seasonal, cycle, irregular.

Application of Time Series:

Financial: Stock market analysis

Industry: Power consumption

Scientific: Experiment result

Meteorological: Precipitation

Time Series Analysis Methods:

Trend Analysis: Categories of Time Series movements:

Long-term or Trend Movements: General direction in which a time series is moving over a long interval of time.

Cyclic Movements: Long-term oscillation about a trend line or curve.

Seasonal Movements: A time series appears to follow substantially identical patterns during the corresponding months of subsequent years.

Irregular or Random Movements: It changes that occur randomly due to unplanned events.

Similarity Search:

Data Reduction

Indexing Methods

Similarity Search Methods

Query Languages

2. Mining Symbolic Sequence

A symbolic sequence is made up of an ordered list of elements that can be recorded with or without a sense of time. This sequence can be used in a variety of ways, including consumer shopping sequences, web clickstreams, software execution sequences, biological sequences, and so on.

Mining of sequential patterns entails identifying the subsequences that appear frequently in one or more sequences. As a result of substantial research in this area, a number of scalable algorithms have been developed. Alternatively, we can only mine the set of closed sequential patterns, where a sequential pattern s is closed if it is a correct subsequence of s’ and s’ has the same support as s.

For example:

if where a, b, c, d and e are items, then S is a subsequence of S’.

3. Mining Biological Sequence

Biological sequences are made up of nucleotide or amino acid sequences. In bioinformatics and modern biology, biological sequence analysis compares, aligns, indexes, and analyzes biological sequences. Biological sequences analysis plays a crucial role in bioinformatics and modern biology. Such analysis can be partitioned into two tasks- pairwise sequence alignment and multiple sequence alignment.

Biological Sequence Methods:

Alignment of Biological Sequences:

Pairwise Alignment

The BLAST Local Alignment Algorithm

Multiple Sequence Alignment Methods

Biological Sequence Analysis Using a Hidden Markov Model:

Markov Chain

Hidden Markov Model

Forward Algorithm

Viterbi Algorithm

Baum-Welch Algorithm

Application of Data Mining:

Financial Information Analysis:

Loan payment prediction/consumer credit policy analysis

Design and construction of information warehouse

Financial information collected in banks and money establishments area units are typically comparatively complete, reliable, and of top quality.

# Retail Industry:

Multidimensional analysis( sales, customers, products, time, etc.)

Sales campaign analysis

Customer retention

Product recommendation

Using visualization tools for data analysis

# Science and Engineering:

Data processing and data warehouse

Mining complex data types

Network-based mining

Graph-based mining

# Trends of Data Mining:

Exploration of applications: addressing application-specific issues

Data mining approaches that are scalable and interactive

Data mining integration with Web search engines, database systems, data warehouse systems, and cloud computing systems

Mining social and information networks

Mining spatiotemporal, moving objects, and cyber-physical systems

Mining multimedia, text, and web data

Mining biological and biomedical data

Visual and audio data mining

Distributed data mining and real-time data stream mining.

## social impacts of Data Mining

Data Mining is to intelligently discover useful information from large amounts of data to solve real-life problems. It is a combination of two words: data and mining. Data is a collection of instances, and mining is designed to filter useful information. Data mining, called knowledge discovery in databases (KDD), is responsible for analyzing data from different perspectives and classifying them. There are many data mining techniques used to transform raw data into useful data. It has various applications such as detecting anomalous behavior, detecting fraud and abuse, terrorist activities, and investigating crimes through lie detection. Data mining can offer many benefits by improving customer service and satisfaction, and lifestyle, in general. Data mining is present in many aspects of our daily lives, whether we realize it or not. It affects how we shop, work, what we search.

## Importance of data mining:

It helps in exploring the large increase in the database and gather only valid information by improving segmentation.

It’s an efficient, cost-effective solution by uncovering the risk and fraud that makes profitable production.

Sometimes customers having difficulty while purchasing helps in decision making and increases the sale.

Data mining techniques can help organizations in real-time plan and save time.

Also, saved money through fraud detection.

## Application area of data mining:

Future Healthcare

Market Basket Analysis

Manufacturing Engineering

Fraud Detection

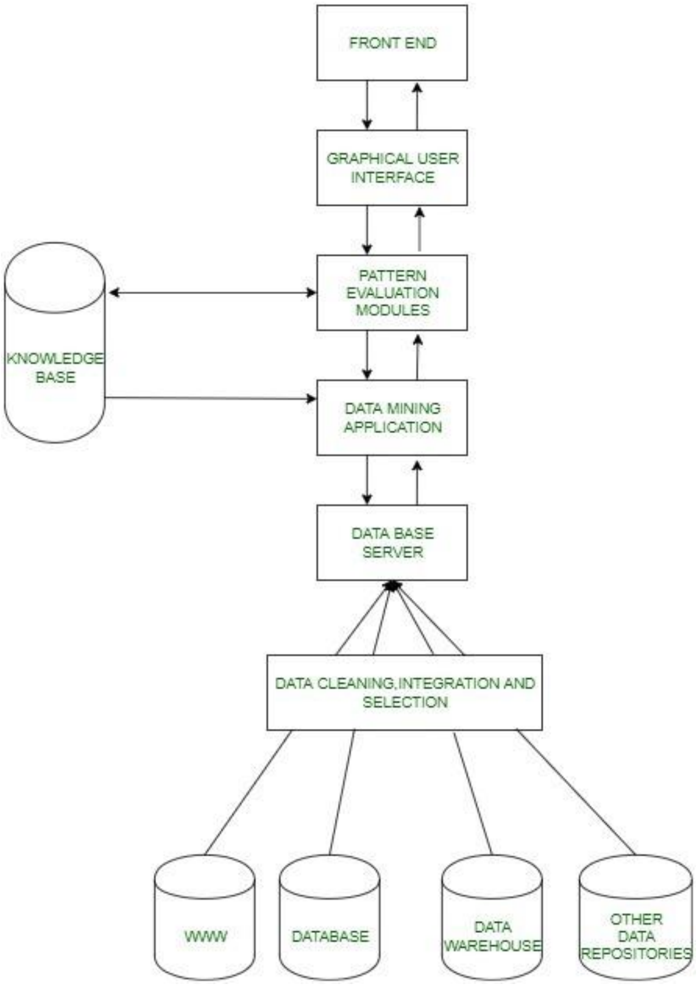
Intrusion Detection

Customer Segmentation

Financial Banking

# Data mining architecture:

Data mining architecture unveils how data extraction falls out. The architecture consists of various steps i.e Data source, data mining engine, data warehouse server, pattern evaluation, user interface, and knowledge base.



How data mining influences privacy, security, and socially:

Security and privacy have been an initial concern all the time.  It aimed at future predictions using previous data. Suppose we buy any product so based on past purchases they make predictions and which also target our personal information. The continuous development of data mining techniques brings serious threats to data security and privacy which is very important to protect. The real threat is that if information gets exposed to unauthorized parties, it will be impractical to stop misuse. Therefore, we must need a system that possesses to protect data and its resources concerning authenticity and integrity.

## How we can protect our data:

Due to minimal protection setup, we lose data so we need to  initiate a multilayer security system

Access Controls are only given to those who have been given the authorization  can access the data

Data must verify an individual user's identity

So, some privacy preservation methods protect sensitive or private data while allowing useful information to be extracted from the data set.

Privacy-Preserving Data Mining (PPDM): The main objective of the PPDM is to protect the privacy of the data and extract only relevant information. It ensures the protection of individual data to conserve privacy and provide accuracy by performing all the data mining operations.

Techniques of PPDM is further divided into various categories:

Data Hiding Technique: In, this technique the data is reform in such a way that the sensitive or private information will not be visible to other parties. Using various ways we can implement these techniques such as Cryptographic Technique, Data Perturbation, and Anonymization Technique.

Knowledge hiding Technique: In this technique, sensitive content is extracted from data using a data mining algorithm. There are different ways of implementing these techniques such as Association Rule Hiding, and Query Auditing.

Hybrid Technique: It is a combination of the two techniques which infuse the limitations of the above two techniques.

## Social Impacts of Data Mining:

Data mining has innovatively influenced our daily lifestyle like how we work, shop, what we buy, search for any information, importantly saves our precious time and offers personalized product recommendations based on our previous history like amazon, Flipkart, etc.

Data mining emerging in all fields like Healthcare, Finance, Marketing, and social media. But there is a higher contribution towards healthcare and well-being by using data mining software to analyze data when developing drugs and to find associations between patients, drugs, and outcomes. And improving patient satisfaction, providing more patient-centered care, and decreasing costs, and increase operating efficiency and Insurance organizations can detect medical insurance fraud and abuse through data mining and reduce their losses.

An old payment system has now taken different forms of transactions depending on usage, acceptability, methods, technology, and availability. It changes the physical financial transactions to virtual payment transactions. So, data mining focuses on successful transactions and keeps track of fake transactions.

It is also used in Web-wide tracking technology that tracks user’s interests while visiting any site. So, information about every site is been recorded, which can be used further to provide marketers with information reflecting your interests.

It is also used for customer relationship management which helps in providing more customized, personal service to individual customers. By studying browsing and purchasing history on Web stores, companies can tailor advertisements and promotions to customer profiles,  only those who are interested and less likely to be annoyed with unwanted mailings. This helps in reducing costs, the waste of time, and improving work productivity.

## Advantages:

Improved security: Data mining can help identify patterns and anomalies that could indicate security breaches, enabling organizations to take action to prevent future attacks.

Personalized services: Data mining can enable personalized services, such as customized product recommendations or personalized healthcare treatments, that can improve the overall customer experience.

Improved decision making: Data mining can provide insights into customer behavior, market trends, and other important factors, enabling organizations to make more informed decisions.

Improved efficiency: Data mining can help streamline business processes, enabling organizations to operate more efficiently.

## Disadvantages:

Privacy concerns: Data mining can raise privacy concerns, as it involves collecting and analyzing data about individuals without their explicit consent. This can result in the disclosure of sensitive information, which can have negative consequences for individuals.

Security risks: The data used in data mining can be subject to security risks, such as unauthorized access or hacking, which can result in the exposure of sensitive information.

Bias: Data mining algorithms can be biased, which can lead to discriminatory outcomes or reinforce existing biases.

Social impacts: Data mining can have social impacts, such as increasing surveillance and reducing personal autonomy, which can have negative consequences for society as a whole.

# Trends in Data Mining

Data mining is one of the most widely used methods to extract data from different sources and organize them for better usage. Despite having different commercial systems for data mining, many challenges come up when they are actually implemented. With the rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments.

Complex algorithms form the basis for data mining as they allow data segmentation to identify trends and patterns, detect variations, and predict the probabilities of various events. The raw data may come in both analog and digital formats and is inherently based on the source of the data. Companies need to keep track of the latest data mining trends and stay updated to do well in the industry and overcome challenging competition.

Corporations can use data mining to discover customers' choices, make a good relationship with customers, increase revenue, and reduce risks. Data mining is based on complex algorithms that allow data segmentation to discover numerous trends and patterns, detect deviations, and estimate the likelihood of certain occurrences occurring. Raw data can be in both analog and digital formats, and it is essentially dependent on the data's source. Companies must keep up with the latest data mining trends and stay current to succeed in the industry and beat out the competition.