What is Data Mining?

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called **Data Mining**.

In other words, we can say that **Data Mining** is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful tools and techniques available to mine data and find better insight from it.

Types of Data

Data mining can be performed on following types of data

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

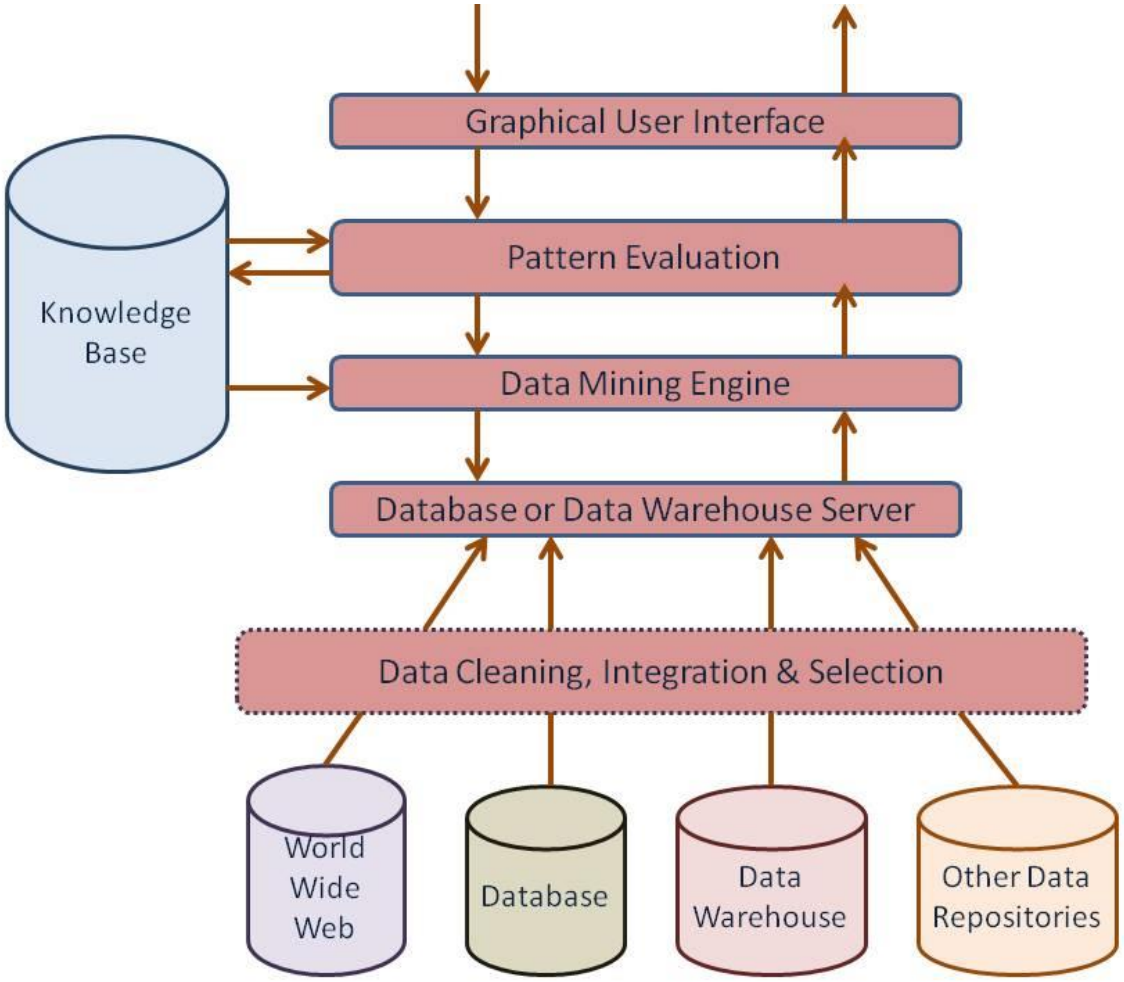
Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not

performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Data Mining Architecture

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.



a) Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spread sheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spread sheets. World Wide Web or the Internet is another big source of data.

Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

c) Data Mining Engine

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface

The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern

evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

Steps involved in KDD Process

1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
   - Cleaning in case of ***Missing values***.
   - Cleaning ***noisy*** data, where noise is a random or variance error.
   - Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.

2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
   - Data integration using ***Data Migration tools***.
   - Data integration using ***Data Synchronization tools***.
   - Data integration using ***ETL*** (Extract-Load-Transformation) process.

3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - Data selection using ***Neural Network***.
   - Data selection using ***Decision Trees***.
   - Data selection using ***Naive Bayes***.
   - Data selection using ***Clustering***, ***Regression***, etc.

4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
   Data Transformation is a two-step process:
   - ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
   - ***Code generation***: Creation of the actual transformation program.

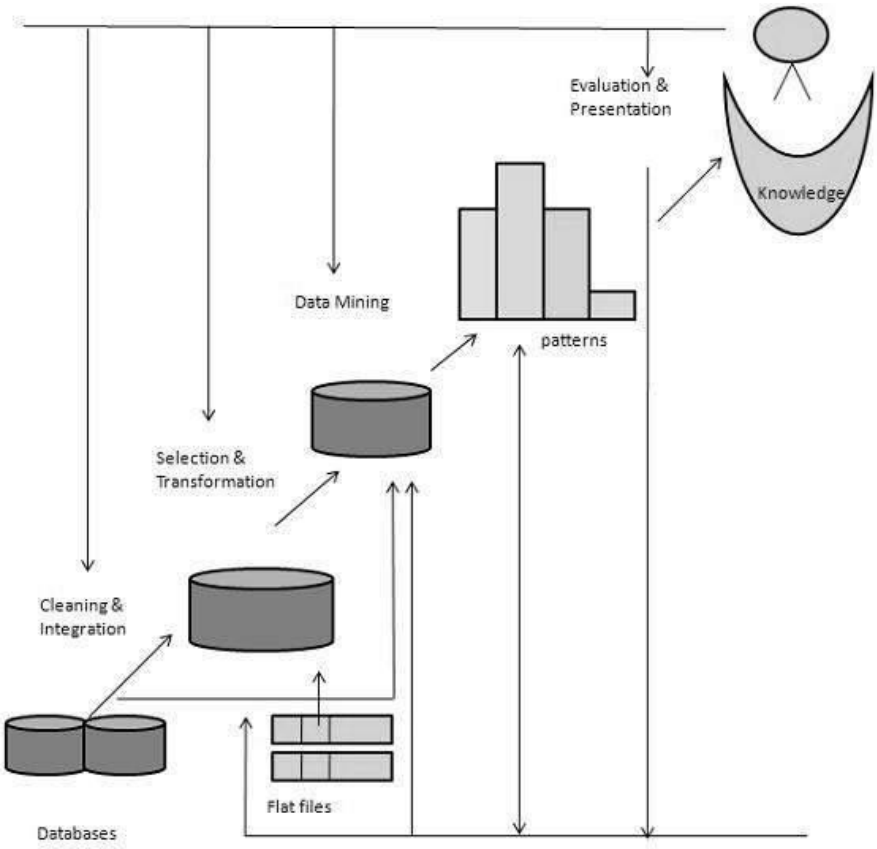
5. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   - Transforms task relevant data into ***patterns***.
   - Decides purpose of model
     using ***classification*** or ***characterization***.

6. ***Pattern Evaluation***: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
   - Find ***interestingness score*** of each pattern.
   - Uses ***summarization*** and ***Visualization*** to make data understandable by user.

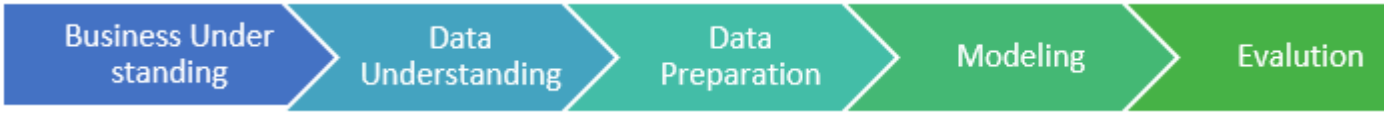
7. ***Knowledge Representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   - Generate ***reports***.
   - Generate ***tables***.
   - Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

Note:
- KDD is an ***iterative*** **process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
- ***Pre-processing*** **of databases** consists of **Data cleaning** and **Data Integration**.

The following diagram shows the process of knowledge discovery −



Data Mining Implementation Process



Data Mining Implementation Process

Let's study the Data Mining implementation process in detail

Business understanding:

In this phase, business and data-mining goals are established.

- First, you need to understand business and client objectives. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in resources, assumption, constraints, and other significant factors into your assessment.
- Using business objectives and current scenario, define your data mining goals.
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals.

Data understanding:

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

- First, data is collected from multiple data sources available in the organization.
- These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.
- Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here, Metadata should be used to reduce errors in the data integration process.
- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions

(decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

Data preparation:

In this phase, data is made production ready.

The data preparation process consumes about 90% of the time of the project.

The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).

Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.

For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 300. Data could be inconsistent. For instance, name of the customer is different in different tables.

Data transformation operations change the data to make it useful in data mining. Following transformation can be applied

Data transformation:

Data transformation operations would contribute toward the success of the mining process.

Smoothing: It helps to remove noise from the data.

Aggregation: Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.

Generalization: In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

Normalization: Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

Attribute construction: these attributes are constructed and included the given set of attributes helpful for data mining.

The result of this process is a final data set that can be used in modeling.

Modelling

In this phase, mathematical models are used to determine data patterns.

- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Evaluation:

In this phase, patterns identified are evaluated against the business objectives.

- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.

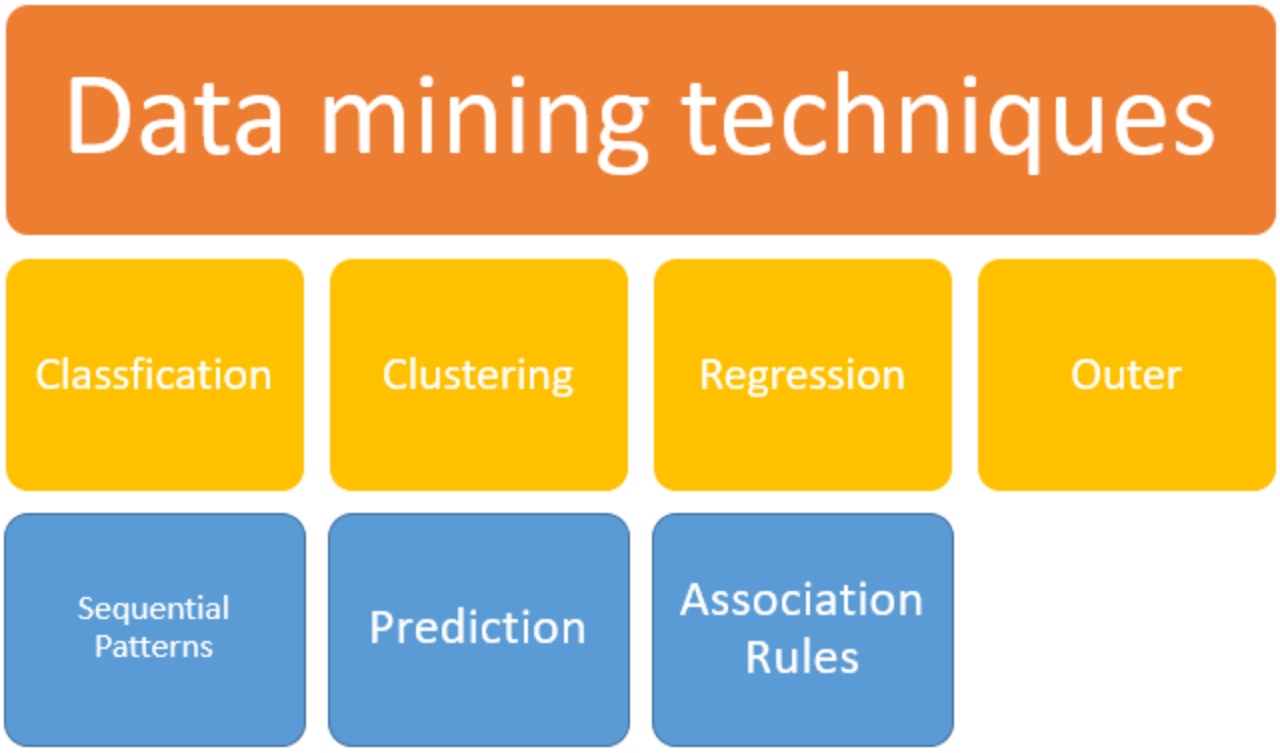
Deployment:

In the deployment phase, you ship your data mining discoveries to everyday business operations.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.

- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

Data Mining Techniques



Data Mining Techniques

1. Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outlier detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behaviour. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outlier detection is also called Outlier mining.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:

Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

Challenges of Implementation of Data Mining:

- Skilled Experts are needed to formulate the data mining queries.
- Overfitting: Due to small size training database, a model may not fit future states.
- Data mining needs large databases which sometimes are difficult to manage
- Business practices may need to be modified to determine to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- Integration information needed from heterogeneous databases and global information systems could be complex

Data mining Examples:

Now in this Data Mining course, let's learn about Data mining with examples:

Example 1:

Consider a marketing head of telecom service provides who wants to increase revenues of long distance services. For high ROI on his sales and marketing efforts customer profiling is important. He has a vast data pool of customer information like age, gender, income, credit history, etc. But it’s impossible to determine characteristics of people who prefer long distance calls with manual analysis. Using data mining techniques, he may uncover patterns between high long distance call users and their characteristics.

For example, he might learn that his best customers are married females between the age of 45 and 54 who make more than $80,000 per year. Marketing efforts can be targeted to such demographic.

Example 2:

A bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved.

Bank has multiple years of record on average credit card balances, payment amounts, credit limit usage, and other key parameters. They create a model to check the impact of the proposed new business policy. The data results show that cutting fees in half for a targeted customer base could increase revenues by $10 million.

Data Mining Tools

Following are 2 popular Data Mining Tools widely used in Industry

R-language:

R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handing and storage facility.

Oracle Data Mining:

Oracle Data Mining popularly known as ODM is a module of the Oracle Advanced Analytics Database. This Data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behaviour, develops customer profiles, identifies cross-selling opportunities.

Benefits of Data Mining:

- Data mining technique helps companies to get knowledge-based information.
- Data mining helps organizations to make the profitable adjustments in operation and production.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Data mining helps with the decision-making process.
- Facilitates automated prediction of trends and behaviours as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Disadvantages of Data Mining

- There are chances of companies may sell useful information of their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.
- Many data mining analytics software is difficult to operate and requires advance training to work on.

- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

Data Mining Tasks

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

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

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.

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 sub-sequences.

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.

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.

Mining of Clusters

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

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived

model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

The lists of functions involved in these processes are as follows −

- **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behaviour or model of the data available.
- **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behaviour changes over time.

Data Mining Task Primitives

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

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

- Set of task relevant data to be mined.
- Kind of knowledge to be mined.

- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following −

- Database Attributes
- Data Warehouse dimensions of interest

Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are −

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following. −

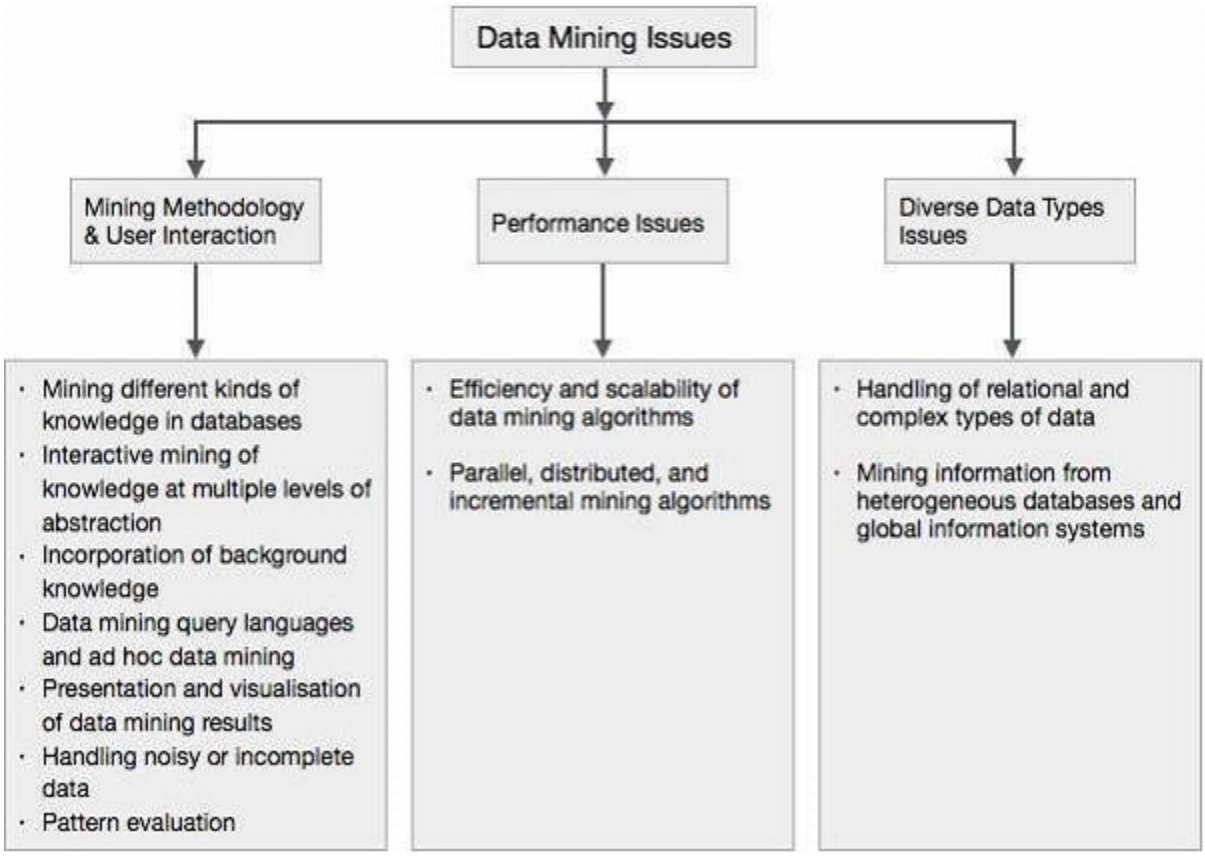
- Rules
- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

Data Mining Issues

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − Data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − Patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

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

Evaluation of Data Mining

Data Warehouse

A data warehouse exhibits the following characteristics to support the management's decision-making process −

- **Subject Oriented** − Data warehouse is subject oriented because it provides us the information around a subject rather than the organization's on-going operations. These subjects can be product, customers, suppliers, sales, revenue, etc. The data warehouse does not focus on the on-going operations, rather it focuses on modelling and analysis of data for decision-making.
- **Integrated** − Data warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhances the effective analysis of data.

- **Time Variant** − Data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from a historical point of view.
- **Non-volatile** – Non-volatile means the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database are not reflected in the data warehouse.

Data Warehousing

Data warehousing is the process of constructing and using the data warehouse. A data warehouse is constructed by integrating the data from multiple heterogeneous sources. It supports analytical reporting, structured and/or ad hoc queries, and decision making.

Data warehousing involves data cleaning, data integration, and data consolidations. To integrate heterogeneous databases, we have the following two approaches −

- Query Driven Approach
- Update Driven Approach

Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach is used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query Driven Approach

- When a query is issued to a client side, a metadata dictionary translates the query into the queries, appropriate for the individual heterogeneous site involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

This approach has the following disadvantages −

- The Query Driven Approach needs complex integration and filtering processes.
- It is very inefficient and very expensive for frequent queries.
- This approach is expensive for queries that require aggregations.

Update-Driven Approach

Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In the update-driven approach, the information from multiple heterogeneous sources is integrated in advance and stored in a warehouse. This information is available for direct querying and analysis.

Advantages

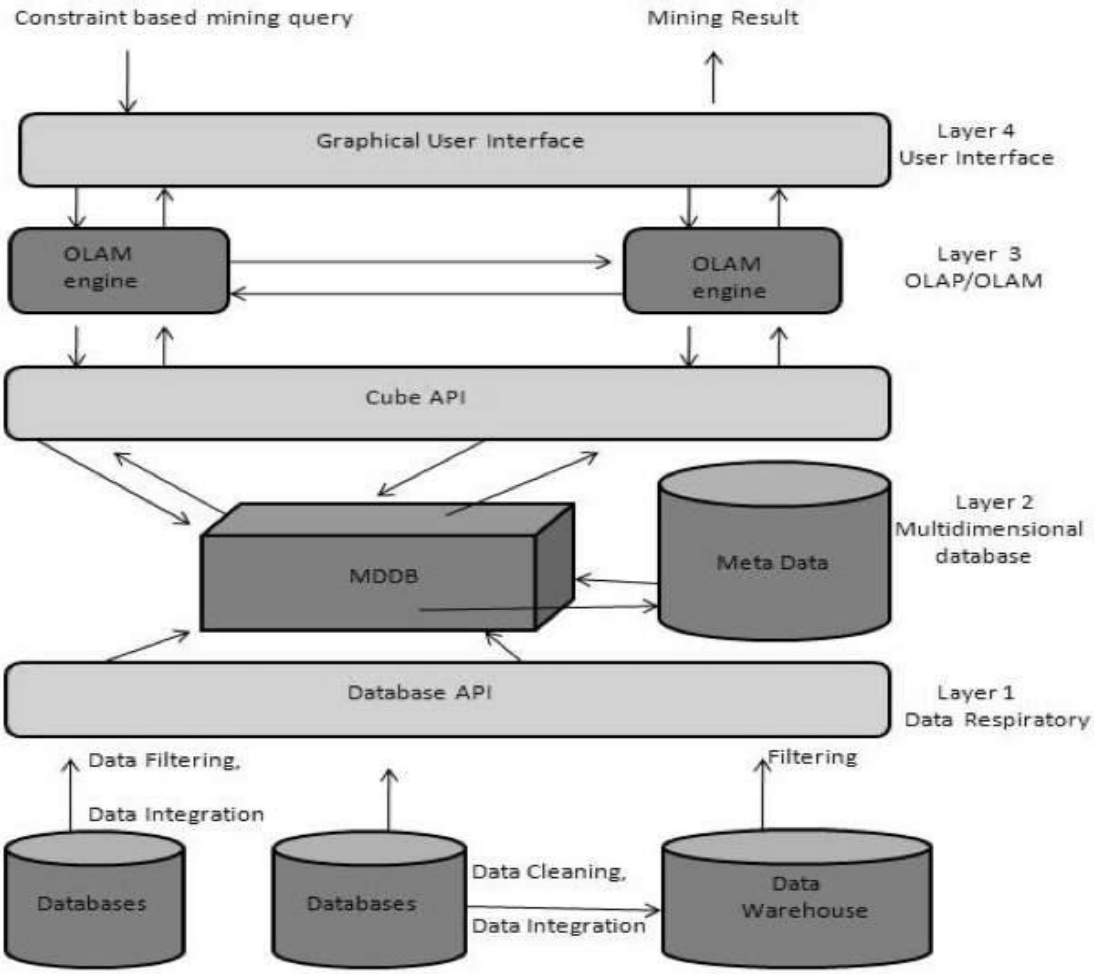
This approach has the following advantages −

- This approach provides high performance.
- The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.

Query processing does not require interface with the processing at local sources.

From Data Warehousing (OLAP) to Data Mining (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows the integration of both OLAP and OLAM −



Importance of OLAM

OLAM is important for the following reasons −

- **High quality of data in data warehouses** − The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the pre-processing of data. The data warehouses constructed by such pre-processing are valuable sources of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses** − Information processing infrastructure refers to accessing, integration, consolidation, and transformation of

multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.

- **OLAP−based exploratory data analysis** − Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subsets of data and at different levels of abstraction.
- **Online selection of data mining functions** − Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

Data Mining Applications

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

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

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural

that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

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

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and

biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modelling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data pre-processing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.

- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.

Choosing a Data Mining System

The selection of a data mining system depends on the following features −

- **Data Types** − Data mining system may handle formatted text, record-based data, and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore, we should check what exact format the data mining system can handle.
- **System Issues** − One must consider the compatibility of a data mining system with different operating systems. One data mining system may run on only one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
- **Data Sources** − Data sources refer to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while others on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.
- **Data Mining functions and methodologies** − There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search, etc.
- **Coupling data mining with databases or data warehouse systems** − Data mining systems need to be coupled with a

database or a data warehouse system. The coupled components are integrated into a uniform information processing environment. Here are the types of coupling listed below −

- No coupling
- Loose Coupling
- Semi tight Coupling
- Tight Coupling

- **Scalability** − There are two scalability issues in data mining −
  - **Row (Database size) Scalability** − A data mining system is considered as row scalable when the number or rows are enlarged 10 times. It takes no more than 10 times to execute a query.
  - **Column (Dimension) Scalability** − A data mining system is considered as column scalable if the mining query execution time increases linearly with the number of columns.
- **Visualization Tools** − Visualization in data mining can be categorized as follows −
  - Data Visualization
  - Mining Results Visualization
  - Mining process visualization
  - Visual data mining
- **Data Mining query language and graphical user interface** − An easy-to-use graphical user interface is important to promote user-guided, interactive data mining. Unlike relational database systems, data mining systems do not share underlying data mining query language.

Trends in Data Mining

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.

Technology Trends in Data Mining

- **Scalable and interactive data mining methods:** Added controls in the form of specifications and constraints can guide data mining systems in not only effectively handling huge volumes of data but also searching for interesting patterns.
- **Standardization of query languages:** Standard querying languages will improve interoperability between different data mining functions and promote systematic development of solutions.
- **Visual data mining:** Visual data mining has picked up pace as one of the top data mining trends, presenting innovative opportunities for knowledge discovery.
- **Research analysis:** Data mining applications are not limited to the tech world. Data cleaning, pre-processing, visualization, and integration of databases have transformed the broad field of research.
- **Web mining:** Web content mining, web log mining, and other mining services on the internet have secured a place among the flourishing subfields of data mining.
- **Multi-database and distributed data mining:** Multi-database data mining analyzes patterns across multiple databases. Whereas distributed data mining searches data from several network locations.
- **Real-time data mining:** Real-time data or ‘stream data’ is generated from web mining, mobile data mining, e-commerce, stock analysis, etc. This type of data requires dynamic data mining models.
- Privacy protection and information security have also come to light as a notable trend in the data mining space.