What Is Data Mining?

The study of collecting, cleaning, processing, analyzing, and gaining useful insights from data

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation, such as predicting whether a newly arrived customer will spend more than $100 at a department store.

Not all information discovery tasks are considered to be data mining. For example, looking up individual records using a database management system or finding particular Web pages via a query to an Internet search engine are tasks related to the area of information retrieval

Data Mining and Knowledge Discovery

- Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information, as shown in Figure 1.1. This process consists of a series of transformation steps, from data preprocessing to postprocessing of data mining results.

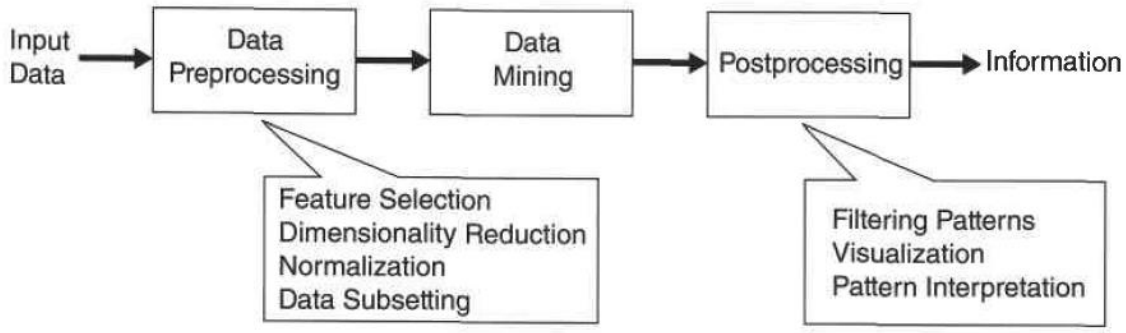
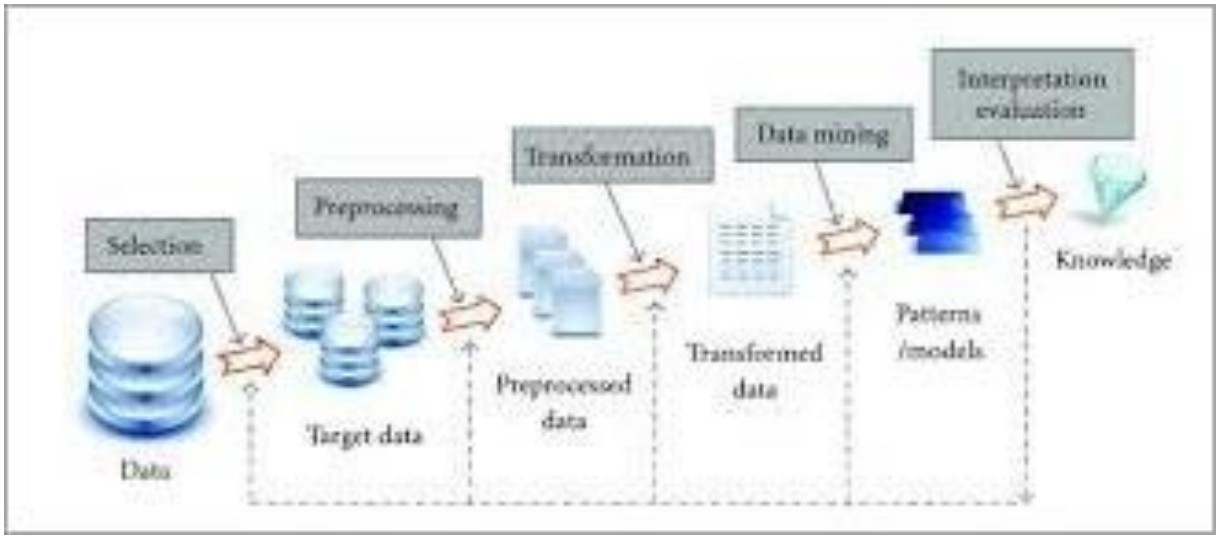


Figure 1.1. The process of knowledge discovery in databases (KDD).

- The input data can be stored in a variety of formats (flat files, spreadsheets, or relational tables) and may reside in a centralized data repository or be distributed across multiple sites. The purpose of preprocessing is to transform the raw input data into an appropriate format for subsequent analysis.
- The steps involved in data preprocessing include fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task at hand.
- Because of the many ways data can be collected and stored, data preprocessing is perhaps the most laborious and time-consuming step in the overall knowledge discovery process.

Stages of the Data Mining Process

The data mining process is divided into two parts i.e. Data Preprocessing and Data Mining. Data Preprocessing involves data cleaning, data integration, data reduction, and data transformation. The data mining part performs data mining, pattern evaluation and knowledge representation of data.



Why do we preprocess the data?

There are many factors that determine the usefulness of data such as accuracy, completeness, consistency, timeliness. The data has to quality if it satisfies the intended purpose. Thus preprocessing is crucial in the data mining process. The major steps involved in data preprocessing are explained below.

#1) Data Cleaning

Data cleaning is the first step in data mining. It holds importance as dirty data if used directly in mining can cause confusion in procedures and produce inaccurate results.

Basically, this step involves the removal of noisy or incomplete data from the collection. Many methods that generally clean data by itself are available but they are not robust.

This step carries out the routine cleaning work by:

(i) Fill The Missing Data:

Missing data can be filled by methods such as:

- Ignoring the tuple.
- Filling the missing value manually.
- Use the measure of central tendency, median or
- Filling in the most probable value.

(ii) Remove The Noisy Data: Random error is called noisy data.

Methods to remove noise are :

Binning: Binning methods are applied by sorting values into buckets or bins. Smoothening is performed by consulting the neighboring values.

Binning is done by smoothing by bin i.e. each bin is replaced by the mean of the bin. Smoothing by a median, where each bin value is replaced by a bin median. Smoothing by bin boundaries i.e. The minimum and maximum values in the bin are bin boundaries and each bin value is replaced by the closest boundary value.

- Identifying the Outliers
- Resolving Inconsistencies

#2) Data Integration

When multiple heterogeneous data sources such as databases, data cubes or files are combined for analysis, this process is called data integration. This can help in improving the accuracy and speed of the data mining process.

Different databases have different naming conventions of variables, by causing redundancies in the databases. Additional Data Cleaning can be performed to remove the redundancies and inconsistencies from the data integration without affecting the reliability of data.

Data Integration can be performed using Data Migration Tools such as Oracle Data Service Integrator and Microsoft SQL etc.

#3) Data Reduction

This technique is applied to obtain relevant data for analysis from the collection of data. The size of the representation is much smaller in volume while maintaining integrity. Data Reduction is performed using methods such as Naive Bayes, Decision Trees, Neural network, etc.

Some strategies of data reduction are:

- **Dimensionality Reduction:** Reducing the number of attributes in the dataset.
- **Numerosity Reduction:** Replacing the original data volume by smaller forms of data representation.
- **Data Compression:** Compressed representation of the original data.

#4) Data Transformation

In this process, data is transformed into a form suitable for the data mining process. Data is consolidated so that the mining process is more efficient and the patterns are easier to understand. Data Transformation involves Data Mapping and code generation process.

Strategies for data transformation are:

- **Smoothing:** Removing noise from data using clustering, regression techniques, etc.
- **Aggregation:** Summary operations are applied to data.
- **Normalization:** Scaling of data to fall within a smaller range.
- **Discretization:** Raw values of numeric data are replaced by intervals. **For Example,** Age.

#5) Data Mining

Data Mining is a process to identify interesting patterns and knowledge from a large amount of data. In these steps, intelligent patterns are applied to extract the data patterns. The data is represented in the form of patterns and models are structured using classification and clustering techniques.

#6) Pattern Evaluation

This step involves identifying interesting patterns representing the knowledge based on interestingness measures. Data summarization and visualization methods are used to make the data understandable by the user.

#7) Knowledge Representation

Knowledge representation is a step where data visualization and knowledge representation tools are used to represent the mined data. Data is visualized in the form of reports, tables, etc.

Data Mining Data Types (Types of Sources of Data)

The following are the data types (types of sources of data) in data mining:

1. Relational Databases

A relational database is a set of records which are linked between using some set of pre-defined constraints. These records are arranged with columns and rows in the form of tables. Tables are used to store data about the items that are to be described in the database.

A relational database is characterized as the set of data arranged in rows and columns in the database tables. In relational databases, the database structure can be defined using physical and logical schema. The physical schema is a schema which describes the database structure and the relationship between tables while logical schema is a schema which describes how tables are linked with one another. The relational database's standard API is SQL. Its applications are data processing, model ROLAP, etc.

2. Data Warehouses

The method of building a data pool using some set of rules is a data warehouse. Through combining data from several heterogeneous sources which enable a user for analytical reporting, standardized and/or ad hoc requests, and decision making. Data warehousing requires data cleaning, integration of data and storage of information. To help historical research, a data warehouse typically preserves several months or years of data. The data in a data warehouse is usually loaded from multiple data sources by an extraction, transformation, and loading process.

3. Transactional Databases

A transaction is, in technical words, a series of sequences of acts that are both independent and dependent at the same time. A transaction is said to be concluded only if all the activities that are part of the transaction are completed successfully. The transaction will be considered an error even if it fails, and all the actions need to be rolled back or undone.

There is a given starting point for any database transaction, followed by steps to change the data inside the database. In the end, before the transaction can be tried again, the database either commits the changes to make them permanent or rolls back the changes to the starting point.

Example - The case of a bank transaction. A bank transaction is said to be accurate only when the amount credited from one account is successfully debited to another account. If the amount is withdrawn but not received by a candidate then it is appropriate to roll back the whole transaction to the original point.

4. Database Management System

DBMS is an application for database development and management. It offers a structured way for users to create, retrieve, update, and manage the data. A person who uses DBMS to communicate with the database need not concern about how and where the data is processed. DBMS will take care of it.

DBMS is a collection of data in a structured manner. DBMS is a system for database management that records information that has some significance. As an example, if we have to create a student database, so we have to add certain attributes such as student ID, student name, student address, student mobile number, student email, etc., and all attributes have the same record type as a student have. The DBMS provides the final user with a reliable firm.

5. Advanced Database System

A new range of databases such as NoSQL/new SQL was targeted by specialized database management systems. New developments in data storage have risen by application demands, such as support for predictive analytics, research, and data processing, are also supported by advanced database management systems. The center of an effective database and information systems has always been advanced data management. It treats a wealth of different data models and surveys the foundations of structuring, sorting, storing, and querying data according to these models.

1.4 Data Mining Tasks

Data mining tasks are generally divided into two major categories:

Predictive tasks. The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the **target** or **dependent variable**, while the attributes used for making the prediction are known as the **explanatory** or **independent variables**.

Descriptive tasks. Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Figure 1.3 illustrates four of the core data mining tasks that are described in the remainder of this book.

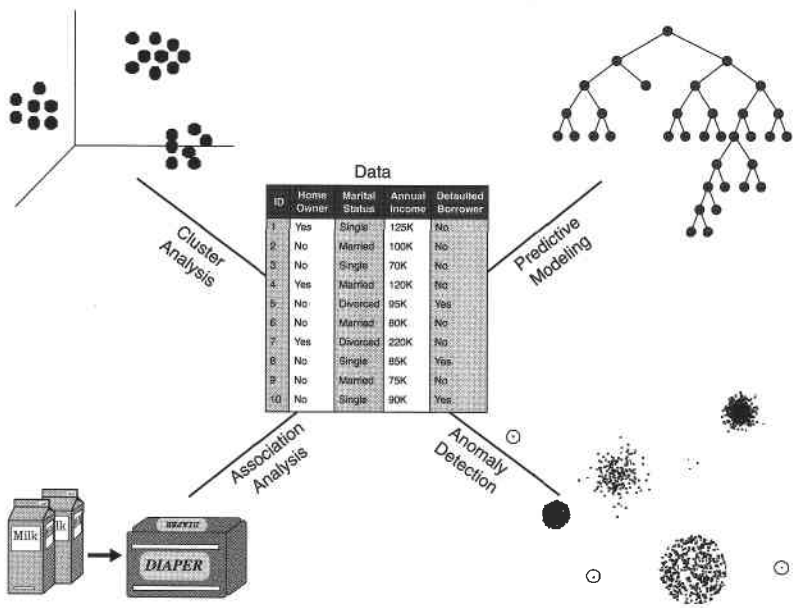


Figure 1.3. Four of the core data mining tasks.

Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks: **classification**, which is used for discrete target variables, and **regression**, which is used for continuous target variables. For example, predicting whether a Web user will make a purchase at an online bookstore is a classification task because the target variable is binary-valued. On the other hand, forecasting the future price of a stock is a regression task because price is a continuous-valued attribute. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable. Predictive modeling can be used to identify customers that will respond to a marketing campaign, predict disturbances in the Earth's ecosystem, or judge whether a patient has a particular disease based on the results of medical tests.

Example 1.1 (Predicting the Type of a Flower). Consider the task of predicting a species of flower based on the characteristics of the flower. In particular, consider classifying an Iris flower as to whether it belongs to one of the following three Iris species: Setosa, Versicolour, or Virginica. To perform this task, we need a data set containing the characteristics of various flowers of these three species. A data set with this type of information is the well-known Iris data set from the UCI Machine Learning Repository at `http://www.ics.uci.edu/~mlearn`. In addition to the species of a flower, this data set contains four other attributes: sepal width, sepal length, petal length, and petal width. (The Iris data set and its attributes are described further in Section 3.1.) Figure 1.4 shows a plot of petal width versus petal length for the 150 flowers in the Iris data set. Petal width is broken into the categories *low*, *medium*, and *high*, which correspond to the intervals [0, 0.75), [0.75, 1.75), [1.75, ∞), respectively. Also, petal length is broken into categories *low*, *medium*, and *high*, which correspond to the intervals [0, 2.5), [2.5, 5), [5, ∞), respectively. Based on these categories of petal width and length, the following rules can be derived:

Petal width low and petal length low implies Setosa.
Petal width medium and petal length medium implies Versicolour.
Petal width high and petal length high implies Virginica.

While these rules do not classify all the flowers, they do a good (but not perfect) job of classifying most of the flowers. Note that flowers from the Setosa species are well separated from the Versicolour and Virginica species with respect to petal width and length, but the latter two species overlap somewhat with respect to these attributes. ■

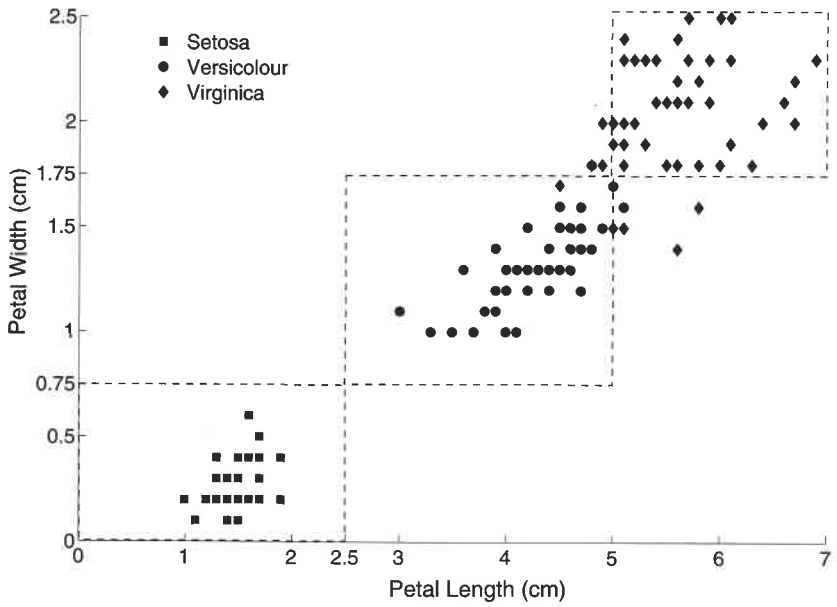


Figure 1.4. Petal width versus petal length for 150 Iris flowers.

Association analysis is used to discover patterns that describe strongly associated features in the data. The discovered patterns are typically represented in the form of implication rules or feature subsets. Because of the exponential size of its search space, the goal of association analysis is to extract the most interesting patterns in an efficient manner. Useful applications of association analysis include finding groups of genes that have related functionality, identifying Web pages that are accessed together, or understanding the relationships between different elements of Earth's climate system.

Example 1.2 (Market Basket Analysis). The transactions shown in Table 1.1 illustrate point-of-sale data collected at the checkout counters of a grocery store. Association analysis can be applied to find items that are frequently bought together by customers. For example, we may discover the rule $\{\texttt{Diapers}\} \longrightarrow \{\texttt{Milk}\}$, which suggests that customers who buy diapers also tend to buy milk. This type of rule can be used to identify potential cross-selling opportunities among related items. ■

Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other

Table 1.1. Market basket data.

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

than observations that belong to other clusters. Clustering has been used to group sets of related customers, find areas of the ocean that have a significant impact on the Earth's climate, and compress data.

Example 1.3 (Document Clustering). The collection of news articles shown in Table 1.2 can be grouped based on their respective topics. Each article is represented as a set of word-frequency pairs (w, c), where w is a word and c is the number of times the word appears in the article. There are two natural clusters in the data set. The first cluster consists of the first four articles, which correspond to news about the economy, while the second cluster contains the last four articles, which correspond to news about health care. A good clustering algorithm should be able to identify these two clusters based on the similarity between words that appear in the articles.

Table 1.2. Collection of news articles.

Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

■

Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as **anomalies** or **outliers**. The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous. In other words, a good anomaly detector must have a high detection rate and a low false alarm rate. Applications of anomaly detection include the detection of fraud, network intrusions, unusual patterns of disease, and ecosystem disturbances.

Example 1.4 (Credit Card Fraud Detection). A credit card company records the transactions made by every credit card holder, along with personal information such as credit limit, age, annual income, and address. Since the number of fraudulent cases is relatively small compared to the number of legitimate transactions, anomaly detection techniques can be applied to build a profile of legitimate transactions for the users. When a new transaction arrives, it is compared against the profile of the user. If the characteristics of the transaction are very different from the previously created profile, then the transaction is flagged as potentially fraudulent. ■

1.5 Scope and Organization of the Book

This book introduces the major principles and techniques used in data mining from an algorithmic perspective. A study of these principles and techniques is essential for developing a better understanding of how data mining technology can be applied to various kinds of data. This book also serves as a starting point for readers who are interested in doing research in this field.

We begin the technical discussion of this book with a chapter on data (Chapter 2), which discusses the basic types of data, data quality, preprocessing techniques, and measures of similarity and dissimilarity. Although this material can be covered quickly, it provides an essential foundation for data analysis. Chapter 3, on data exploration, discusses summary statistics, visualization techniques, and On-Line Analytical Processing (OLAP). These techniques provide the means for quickly gaining insight into a data set.

Chapters 4 and 5 cover classification. Chapter 4 provides a foundation by discussing decision tree classifiers and several issues that are important to all classification: overfitting, performance evaluation, and the comparison of different classification models. Using this foundation, Chapter 5 describes a number of other important classification techniques: rule-based systems, nearest-neighbor classifiers, Bayesian classifiers, artificial neural networks, support vector machines, and ensemble classifiers, which are collections of classi-

SCOPE OF DATA MINING

1. data mining technology can generate new business opportunities

2. Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

3. Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms

5. **Data mining can be implemented on new systems as existing platforms are upgraded and new products developed.**

When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

How Data Mining Works?

Data Mining is a process that consists of the following steps:-

- Select the required kind of data to build target datasets.
- Explore the data and preprocess it to bring it into acceptable formats.
- Data preparation by creating segmentation rules, cleaning noise, performing anomaly checks, filling in missing values, etc.
- In the last stage, use Machine Learning algorithms on the mined data to get the expected results.

Data Mining Architecture

Introduction

Data mining is a critical procedure for extracting potentially valuable and previously undiscovered information from massive amounts of data. The data mining process is made up of a number of components.

Data Mining Architecture

A data mining system's architecture is made up of these components. It is the process of analysing observational datasets to uncover previously unknown associations and summarise the data in unique ways that are both intelligible and helpful to the data owner.

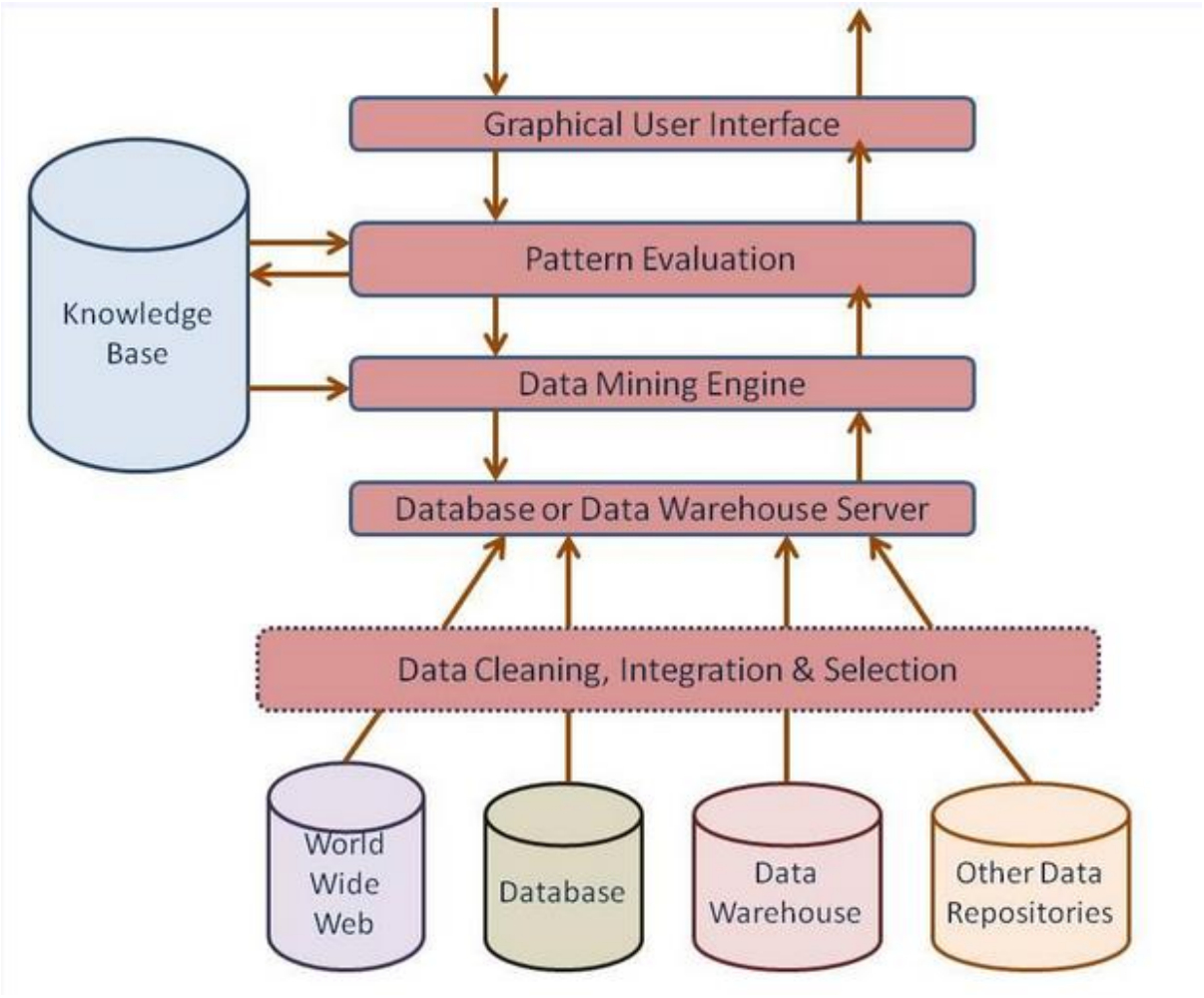
What is Data Mining Architecture

Data Mining Architecture is the process of selecting, exploring, and modelling large amounts of data to discover previously unknown regularities or relationships to generate clear and valuable findings for the database owner. Data mining is exploring and analysing large amounts of data using automated or semi-automated processes to identify practical designs and procedures.

The primary components of any data mining system are the Data source, data warehouse server, data mining engine, pattern assessment module, graphical user interface, and knowledge base.

Basic Working:

1. When a user requests data mining queries, these requests are sent to data mining engines for pattern analysis.
2. These software applications use the existing database to try to discover a solution to the query.
3. The retrieved metadata is then transmitted to the data mining engine for suitable processing, which may interact with pattern assessment modules to decide the outcome.
4. The result is finally delivered to the front end in a user-friendly format via an appropriate interface.



Components Of Data Mining Architecture

- Data Sources
- Database Server
- Data Mining Engine
- Pattern Evaluation Modules
- Graphic User Interface
- Knowledge Base

Data Sources

These sources provide the data in plain text, spreadsheets, or other media such as images or videos Data sources include databases, the World Wide Web (WWW), and data warehouses.

Database Server

The real data is stored on the database server and is ready to be processed. Its job is to handle data retrieval in response to the user's request.

Data Mining Engine:

It is one of the most important parts of the data mining architecture since it conducts many data mining techniques such as association, classification, characterisation, clustering, prediction, and so on.

Pattern Evaluation Modules:

They are responsible for identifying intriguing patterns in data and, on occasion, interacting with database servers to provide the results of user queries.

Graphic User Interface:

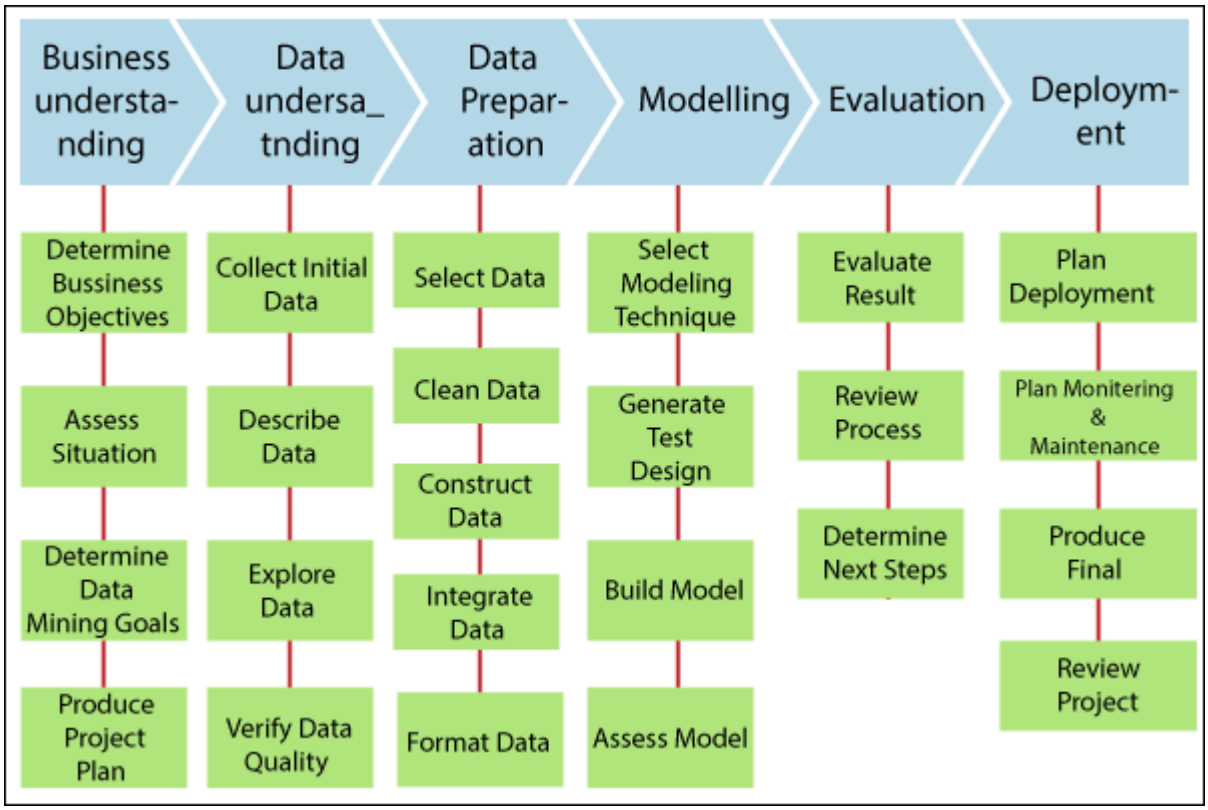
Because the user cannot completely comprehend the complexities of the data mining process, a graphical user interface assists the user in efficiently communicating with the data mining system.

Knowledge Base:

The Knowledge Base is an essential component of the data mining engine that aids in the search for outcome patterns. Occasionally, the knowledge base may also provide input to the data mining engine. This knowledge base might include information gleaned from user encounters. The knowledge base's goal is to improve the accuracy and reliability of the outcome. The Knowledge Base is a crucial component of the data mining engine that aids in the search for outcome patterns. Occasionally, the knowledge base may also provide input to the data mining engine. This knowledge base might include information gleaned from user encounters. The knowledge base's goal is to improve the accuracy and reliability of the outcome.

Data Mining Implementation Process

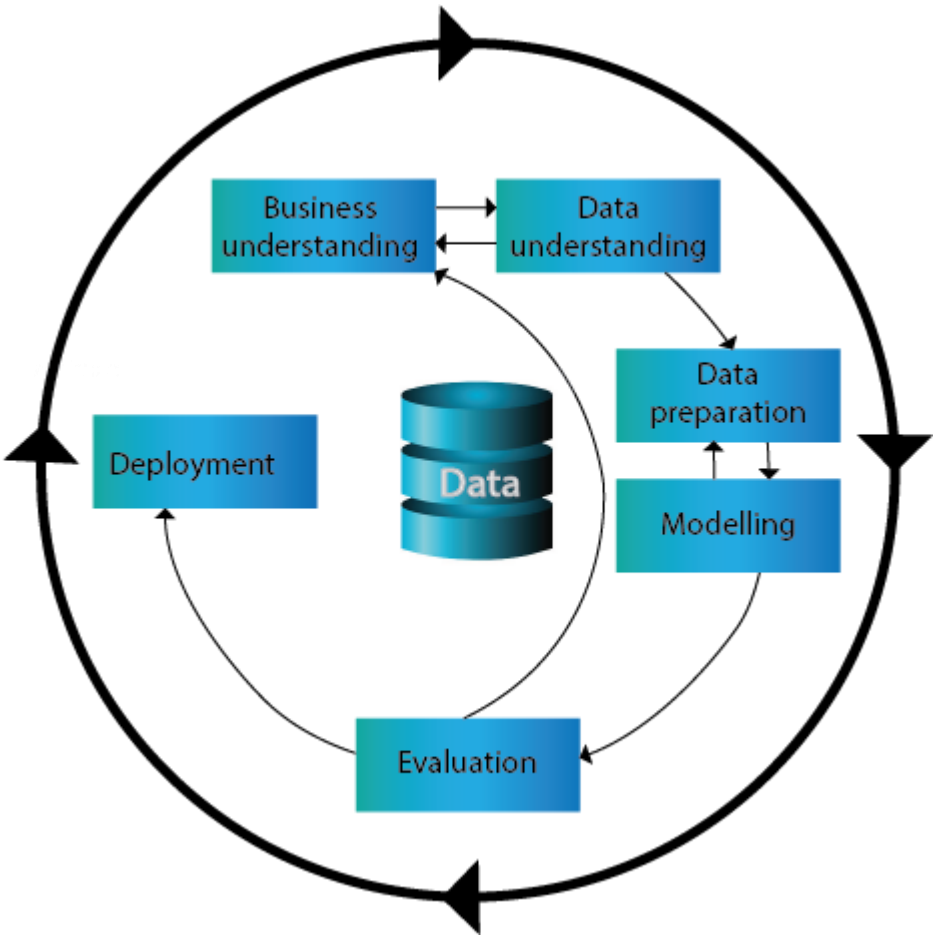
Data mining is described as a process of finding hidden precious data by evaluating the huge quantity of information stored in data warehouses, using multiple data mining techniques such as Artificial Intelligence (AI), Machine learning and statistics.



Let's examine the implementation process for data mining in details:

The Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-industry Standard Process of Data Mining (CRISP-DM) comprises of six phases designed as a cyclical method as the given figure:



1. Business understanding:

It focuses on understanding the project goals and requirements form a business point of view, then converting this information into a data mining problem afterward a preliminary plan designed to accomplish the target.

Tasks:

- Determine business objectives - It Understands the project targets and prerequisites from a business point of view andunderstand what the customer wants to achieve.
- Access situation - It requires a more detailed analysis of facts about all the resources, constraints and assumptions.
- Determine data mining goals - A business goal states the target of the business terminology. A data mining goal describes the project objectives.
- Produce a project plan

    - It states the targeted plan to accomplish the business and data mining plan.
    - The project plan should define the expected set of steps to be performed during the rest of the project, including the latest technique and better selection of tools.

2. Data Understanding:

Data understanding starts with an original data collection and proceeds with operations to get familiar with the data, to data quality issues, to find better insight in data, or to detect interesting subsets for concealed information hypothesis.

Tasks:

- Collects initial data
- Describe data
- Explore data
- Verify data quality

Collect initial data:

- It acquires the information mentioned in the project resources.
- It includes data loading if needed for data understanding.
- It may lead to original data preparation steps.
- If various information sources are acquired then integration is an extra issue, either here or at the subsequent stage of data preparation.

Describe data:

- It examines the "gross" or "surface" characteristics of the information obtained.
- It reports on the outcomes.

Explore data:

- Addressing data mining issues that can be resolved by **querying, visualizing,** and **reporting,** including:
    - Distribution of important characteristics, results of simple aggregation.
    - Establish the relationship between the small number of attributes.
    - Characteristics of important sub-populations, simple statical analysis.

Verify data quality:

- It examines the data quality and addressing questions.

3. Data Preparation:

- It usually takes more than 90 percent of the time.
- It covers all operations to build the final data set from the original raw information.
- Data preparation is probable to be done several times and not in any prescribed order.

Tasks:

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

Select data:

- It decides which information to be used for evaluation.
- In the data selection criteria include significance to data mining objectives, quality and technical limitations such as data volume boundaries or data types.
- It covers the selection of characteristics and the choice of the document in the table.

Clean data:

- It may involve the selection of clean subsets of data, inserting appropriate defaults or more ambitious methods, such as estimating missing information by modeling.

Construct data:

- It comprises of Constructive information preparation, such as generating derived characteristics, complete new documents, or transformed values of current characteristics.

Integrate data:

- Integrate data refers to the methods whereby data is combined from various tables, or documents to create new documents or values.

Format data:

- Formatting data refer mainly to linguistic changes produced to information that does not alter their significance but may require a modeling tool.

4. Modeling:

In modeling, various modeling methods are selected and applied, and their parameters are measured to optimum values. Some methods gave particular requirements on the form of data. Therefore, stepping back to the data preparation phase is necessary.

Tasks:

- Select modeling technique
- Generate test design
- Build model
- Access model

Select modeling technique:

- It selects the real modeling method that is to be used. For example, decision tree, neural network.
- If various methods are applied,then it performs this task individually for each method.

Generate test Design:

- Generate a procedure or mechanism for testing the validity and quality of the model before constructing a model. For example, in classification, error rates are commonly used as quality measures for data mining models. Therefore, typically separate the data set into train and test set, build the model on the train set and assess its quality on the separate test set.

Build model:

- To create one or more models, we need to run the modeling tool on the prepared data set.

Assess model:

- It interprets the models according to its domain expertise, the data mining success criteria, and the required design.
- It assesses the success of the application of modeling and discovers methods more technically.
- It Contacts business analytics and domain specialists later to discuss the outcomes of data mining in the business context.

5. Evaluation:

- It evaluates the model efficiently, and review the steps executed to build the model and to ensure that the business objectives are properly achieved.
- The main objective of the evaluation is to determine some significant business issue that has not been regarded adequately.
- It decides whether to complete the project and move on to deployment when necessary or whether to initiate further iterations or set up new data-mining initiatives.it includes resources analysis and budget that influence the decisions.

6. Deployment:

Determine:

- Deployment refers to how the outcomes need to be utilized.

Deploy data mining results by:

Tasks:

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

Plan deployment:

- To deploy the data mining outcomes into the business, takes the assessment results and concludes a strategy for deployment.
- It refers to documentation of the process for later deployment.

Plan monitoring and maintenance:

- It is important when the data mining results become part of the day-to-day business and its environment.
- It helps to avoid unnecessarily long periods of misuse of data mining results.
- It needs a detailed analysis of the monitoring process.

Produce final report:

- A final report can be drawn up by the project leader and his team.
- It may only be a summary of the project and its experience.
- It may be a final and comprehensive presentation of data mining.

Review project:

- Review projects evaluate what went right and what went wrong, what was done wrong, and what needs to be improved.

Data Mining Techniques

The most commonly used techniques in the field include:

- Detection of anomalies: Identifying unusual values in a dataset.
- Dependency modelling: Discovering existing relationships within a dataset. This frequently involves regression analysis.
- Clustering: Identifying structures (clusters) in unstructured data.
- Classification: Generalizing the known structure and applying it to the data.

Techniques in data mining are:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 31). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

Advantages of Data Mining

It helps gather reliable information – Data mining allows companies, organisations, and governments to gather reliable information.

Helps businesses make operational adjustments – Data mining helps businesses make profitable production and operational adjustments. Data mining can be used to find correlations between products, consumers, suppliers and other aspects of the business. This can help a company identify trends that might not have been identified before, or at least help them make more accurate predictions.

Helps to make informed decisions – It is often used for business purposes to improve decision making.

It helps detect risks and fraud – Data mining can help identify risks and fraud that may not be detectable through traditional means of data analysis.

Helps to analyse very large quantities of data quickly – Data mining can be used to analyse data that was previously too difficult to understand due to the sheer volume or type of information.

Helps to understand behaviours, trends and discover hidden patterns – Data mining can be used to find patterns and trends in user behaviour.

Disadvantages of Data Mining

- Data mining is not always unerring and in certain cases can lead to unwanted effects.
- A large database is required to go for mining thus making the process hard.
- Selection of the right tool for a certain business is a cumbersome task as each tool has a different algorithm.
- Data mining is hard and complex, thus a proper training about various tools is required.