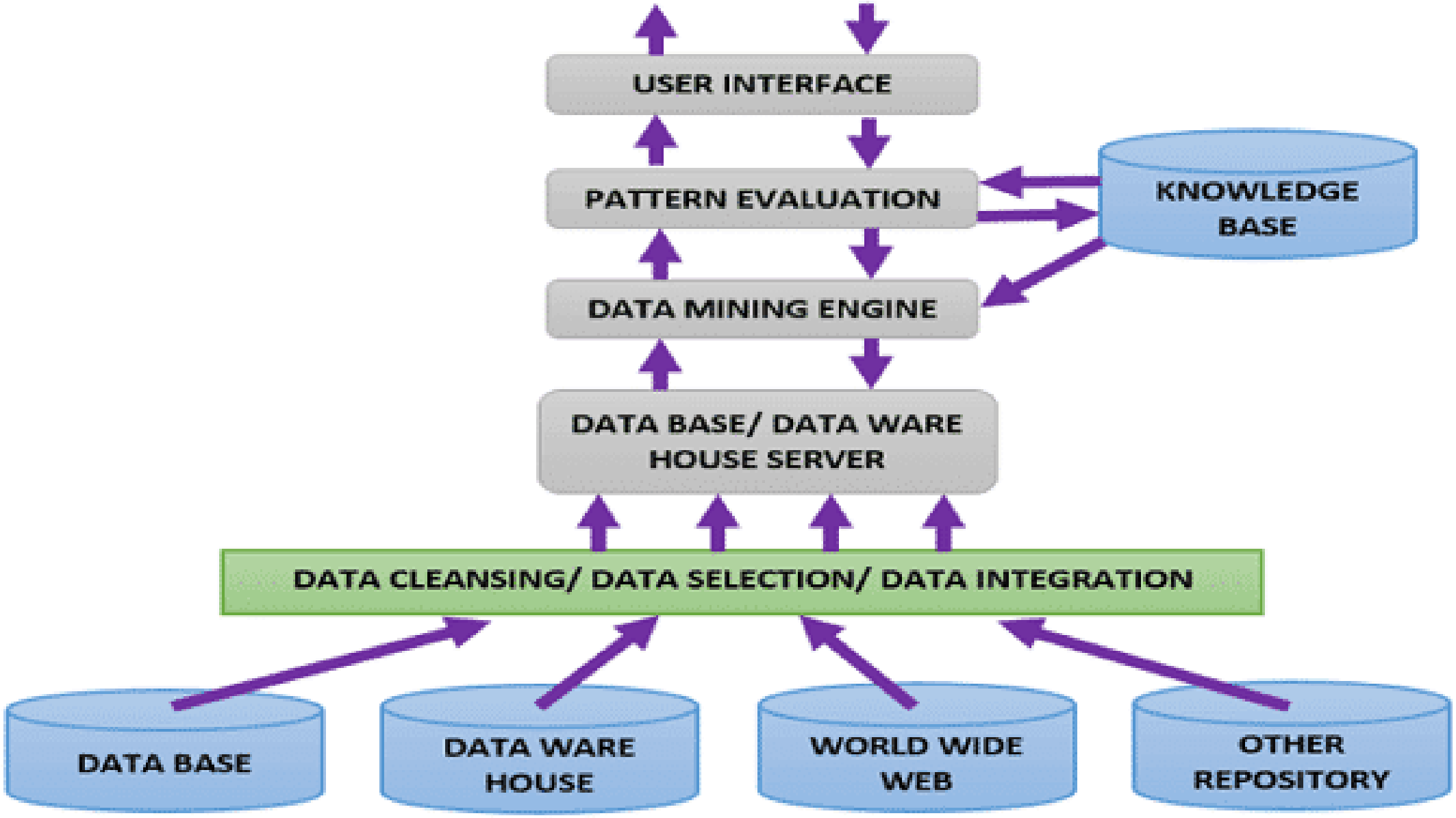
Introduction To
Data Mining

WHAT IS DATA MINING? DATA MINING IS

- The technique of collecting useful information from a huge amount of data is known as data mining. It assists in the exploration and identification of important trends and patterns in datasets.

- Data mining is an interdisciplinary field that employs statistics, database systems, artificial intelligence, and machine learning techniques. Algorithms are used in data mining to extract patterns from databases.

DATA MINING ARCHITECTURE



Applications of Data Mining

- 01 Research
- 02 Education sector
- 03 Transportation
- 04 Market Basket Analysis
- 05 Business Transactions
- 06 Intrusion Detection
- 07 Scientific Analysis
- 08 Finance and Banking sector
- 09 Insurance and Healthcare

- **Research Analysis**
- We have seen dramatic advances in research throughout history. Data mining is useful for data cleansing, data pre-processing, and database integration. The researchers can scan the database for any similar data that could affect the research.
- **Education**
- Educational Data Mining is a rapidly growing area that is concerned with creating ways for discovering information from data originating from educational environments. Predicting students' future learning behavior, researching the impacts of educational assistance, and improving scientific understanding about learning are all aims of EDM.
- An institution may utilize data mining to make correct judgments and anticipate student outcomes.

Market Basket Analysis

Market Basket Analysis is a method for analyzing the purchases made by a consumer in a supermarket. This notion identifies a customer's habit of regular purchases. This study may assist firms in to advertise bargains, offers, and sales, and data mining tools.

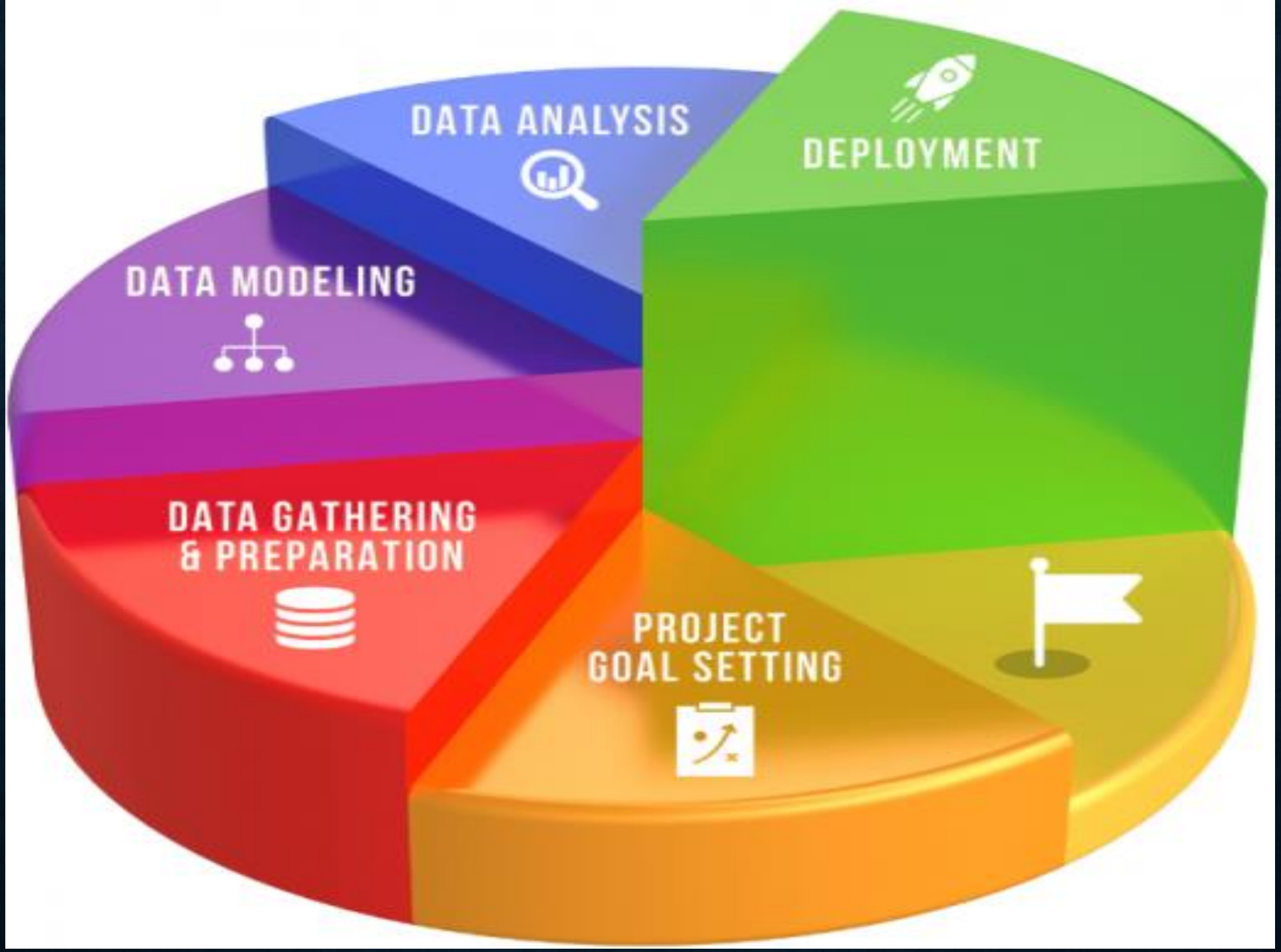
Banking and Finance

- The banking industry is now dealing with and managing massive volumes of data and transaction information as a result of digitalization.
- With its capacity to detect patterns, casualties, market risks, and other connections that are critical for managers to be aware of, data mining applications in banking can easily be the suitable answer.

Intrusion Detection

Any activity that jeopardizes the resource's integrity and confidentiality is considered an incursion. User authentication, avoiding programming mistakes, and information protection are some of the defensive measures taken to prevent infiltration.

- By adding a degree of emphasis to anomaly detection, data mining can assist enhance intrusion detection. It allows an analyst to discern between unusual network activity and normal network activity. Data mining also aids in the extraction of data that is more relevant to the issue.
- **Healthcare**
  Data mining has a lot of promise for improving healthcare systems. It identifies best practices for improving treatment and lowering costs using data and analytics.
- Patients receive appropriate care at the correct place and at the right time thanks to the development of processes. Healthcare insurers can employ data mining to detect fraud and misuse.


DATA ANALYSIS
DEPLOYMENT
DATA MODELING
DATA GATHERING
& PREPARATION
PROJECT
GOAL SETTING

THE DATA MINING PROCESS

- 1. Project Goal Setting
- 2. Data gathering and preparation
- 3. Data Modelling
- 4. Data Analysis
- 5. Deployment

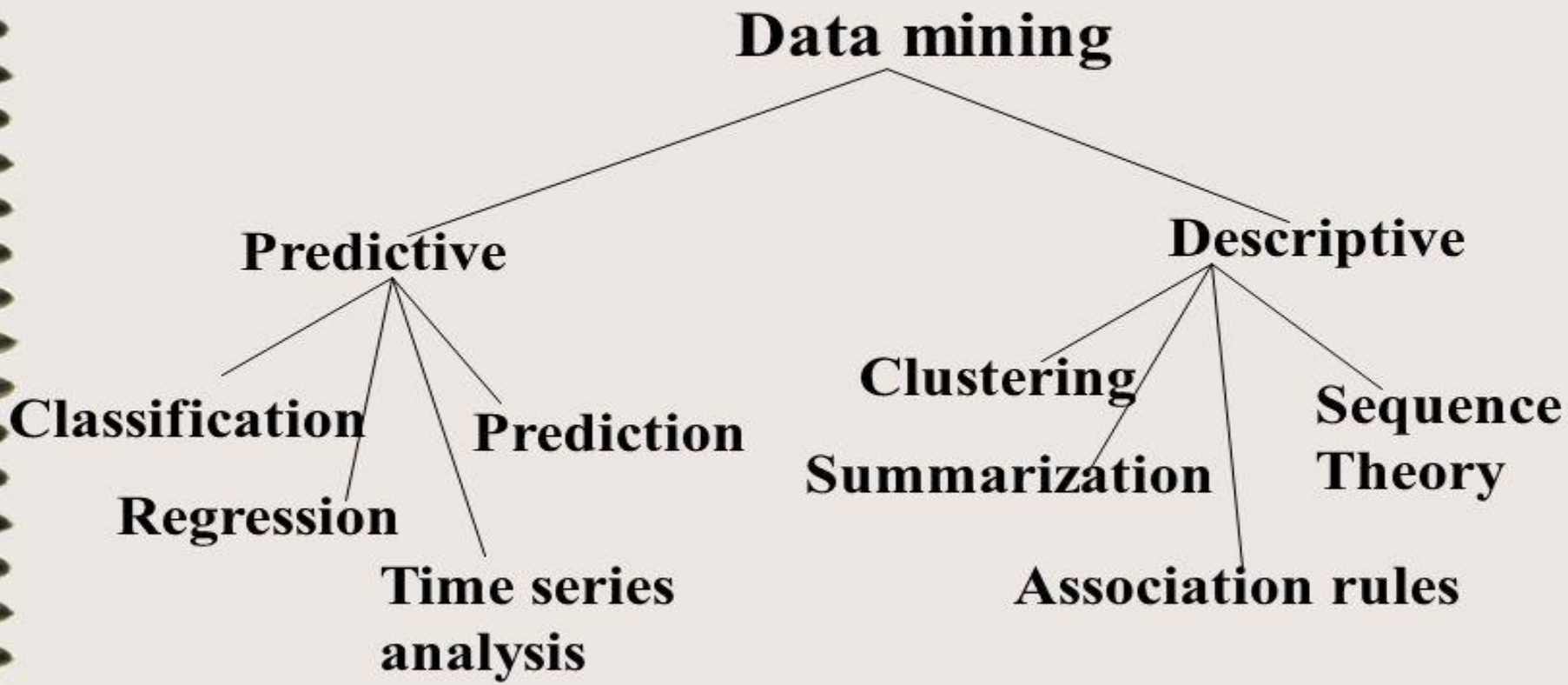
- **Project Goal Setting**
- Goal setting is the foundation of every successful data mining project. Through aligning on their project objectives and timelines, business and data mining teams can have a smoother working relationship throughout the experience.
- Goal setting allows teams to assign roles and make a clear plan to move forward. Expectation management is key to avoiding issues throughout the data mining process
- **Data Gathering & Preparation**
- The data gathering and preparation stage is all about making sure that the data is usable.
- For larger, more established clients, there must be mitigation of security risk. Trust is a necessary element when dealing with sensitive information. Data processing often uses modern database management systems (DBMS) to improve data mining speed. It is also a primary precaution when dealing with data that is confidential to an organization.

- **Data Modeling**
- With the use of mathematical models and various data visualization tools, there are meaningful patterns discovered in the data. Through conceptual representations of how data objects and rules go hand in hand, they form a Database.
- A Database can be conceptual, physical, or logical, depending on the Data Model applied. With the right structure, it can help define relational tables, keys, and procedures. For Data Modeling to work, it needs to have quality data, security procedures, consistent semantics, default values, and naming conventions. There are two types of Data Modeling Techniques: Entity-Relationship (E-R) Model & Unified Modeling Language (UML).
- **Data Analysis**
- After the modeled data is analyzed, it is then extracted, transformed, and visualized. Data analysis helps bring together useful information to give insights or test hypotheses.
- With a combination of business intelligence and analytics models, Data Analysis orders raw data in a way that is relevant to the project goals. Armed with visual representations and insight on previously unrefined data, it is then ready for deployment towards relevant business units.

- **Deployment**
- In the last stage of Data Mining, relevant partners test the hypothesis. There are four different types of model deployment: data science tools, programming language, database, and SQL script or predictive model markup language.

- Mined data provides a single source of truth that can guide business decisions moving forward. With coordination between data scientists, IT teams, software developments, and business professionals work together to integrate the new models with the existing production system of an organization.

Data Mining Models and Tasks



DATA MINING TASKS

- 1. Classification: learning a function that maps an item into one of a set of predefined classes
- 2. Regression: learning a function that maps an item to a real value
- 3. Clustering: identify a set of groups of similar items
- **4.** Dependencies and associations:
  identify significant dependencies between data attributes
- 5. Summarization: find a compact description of the dataset or a subset of the dataset

Predictive model makes prediction based on the previous result sets ; it uses historical data.

For e.g a credit card use might be refused not because of the user's own credit history, but because of the current purchase is similar to earlier purchases that were subsequently found to be made stolen cards.

Here the predictive model is used to predict the ***credit risk.***

A descriptive model identifies patterns or relationship

Classification:

- Maps data into predefined groups or classes

- It is also referred as supervised learning because the classes are defined before examining the data.

-E.g whether to make a bank loan and identifying credit risks.

-Pattern recognition is a type of classification.

In pattern recognition an input pattern is classified into one of several classes based on its similarity to these predefined classes

Example:

An airport security screening station used to determine if passenger is terrorist or criminals

Regression:

It is used to map a data item to a real valued prediction variable.

In regression there is a learning of function that does mapping.

Regression assumes that the target data fit into some known type of function (e.g linear , logistic,etc);

For e.g A professor want to reach a certain level of savings

Time Series Analysis :

The value of an attribute is examined as it varies over time. The values are obtained as evenly spaced(daily,weekly,hourly etc.).

The time series plot is used to visualize the time series.

Prediction:

Prediction is a type of classification.

The only difference is that prediction is predicting a future state rather than current state.

e.g Predicting flooding ;

Clustering:

Clustering is alternatively referred to as unsupervised learning or segmentation.

The clustering is usually accomplished by determining the similarity among data on predefined attributes.

For e.g Catlogs of demographic groups;

Summarization :

It maps data into subsets with associated simple descriptions.

Summarization is also called characterization or generalization.

It extracts or derives representative information about the database.

For e.g One of many criteria used to compare universities by the U.S News and World Report is the average SAT or ACT score.

Association Rules:

An association rule is a model that identifies specific types of data associations.

Sequence Discovery:

Sequential analysis is used to determine sequential patterns in data.And these patterns are based on a time sequence of actions.

They are also similar to associations in that data are found to be related , but the relationship is based on time.