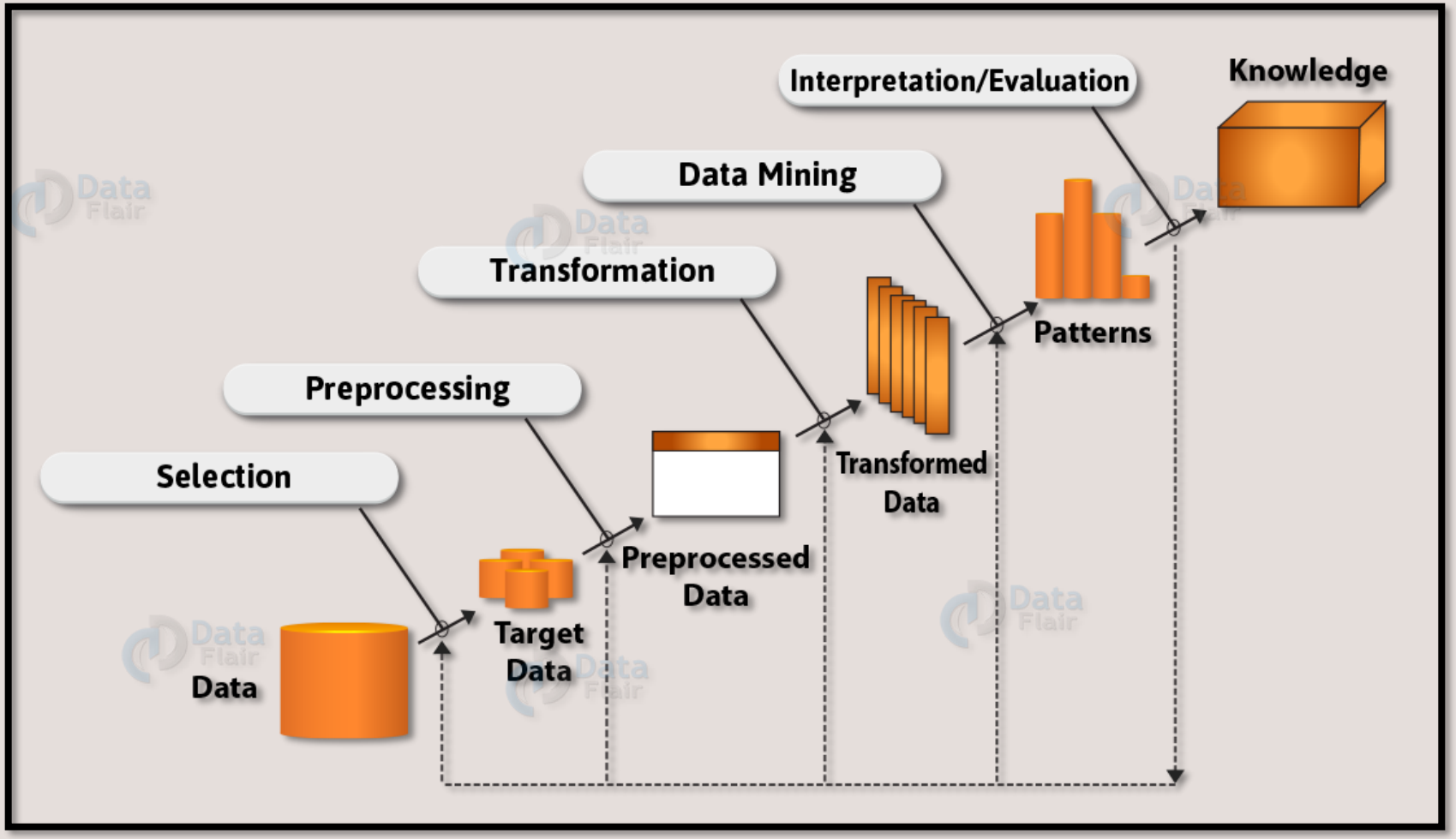
DATA MINING

Subject Incharge: Priya Sachdeva, Assistant Professor (CSE)

What is KDD

- **KDD** refers to the overall process of discovering useful knowledge from data.
- It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge.
- It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
- KDD in data mining is a programmed and analytical approach to model data from a database to extract useful and applicable ‘knowledge’.
- Data mining forms the backbone of KDD and hence is critical to the whole method.
- Data mining refers to the application of algorithms for extracting patterns from data.

Steps involved in the KDD process



The KDD process in data mining is a multi-step process that involves various stages to extract useful knowledge from large datasets. The following are the main steps involved in the KDD process –

1. **Data Selection -**

- The first step in the KDD process is identifying and selecting the relevant data for analysis.
- This involves choosing the relevant data sources, such as databases, data warehouses, and data streams, and determining which data is required for the analysis.

2. Data Preprocessing -

- After selecting the data, the next step is data preprocessing.
- This step involves cleaning the data, removing outliers, and removing missing, inconsistent, or irrelevant data.
- This step is critical, as the data quality can significantly impact the accuracy and effectiveness of the analysis.

3. **Data Transformation -**

- Once the data is preprocessed, the next step is to transform it into a format that data mining techniques can analyze.
- This step involves reducing the data dimensionality, aggregating the data, normalizing it, and discretizing it to prepare it for further analysis.

4. Data Mining -

- This is the heart of the KDD process and involves applying various data mining techniques to the transformed data to discover hidden patterns, trends, relationships, and insights.
- A few of the most common data mining techniques include clustering, classification, association rule mining, and anomaly detection.

5. **Pattern Evaluation** -

- After the data mining, the next step is to evaluate the discovered patterns to determine their usefulness and relevance.
- This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis.

6. Knowledge Representation -

- This step involves representing the knowledge extracted from the data in a way humans can easily understand and use.
- This can be done through visualizations, reports, or other forms of communication that provide meaningful insights into the data.

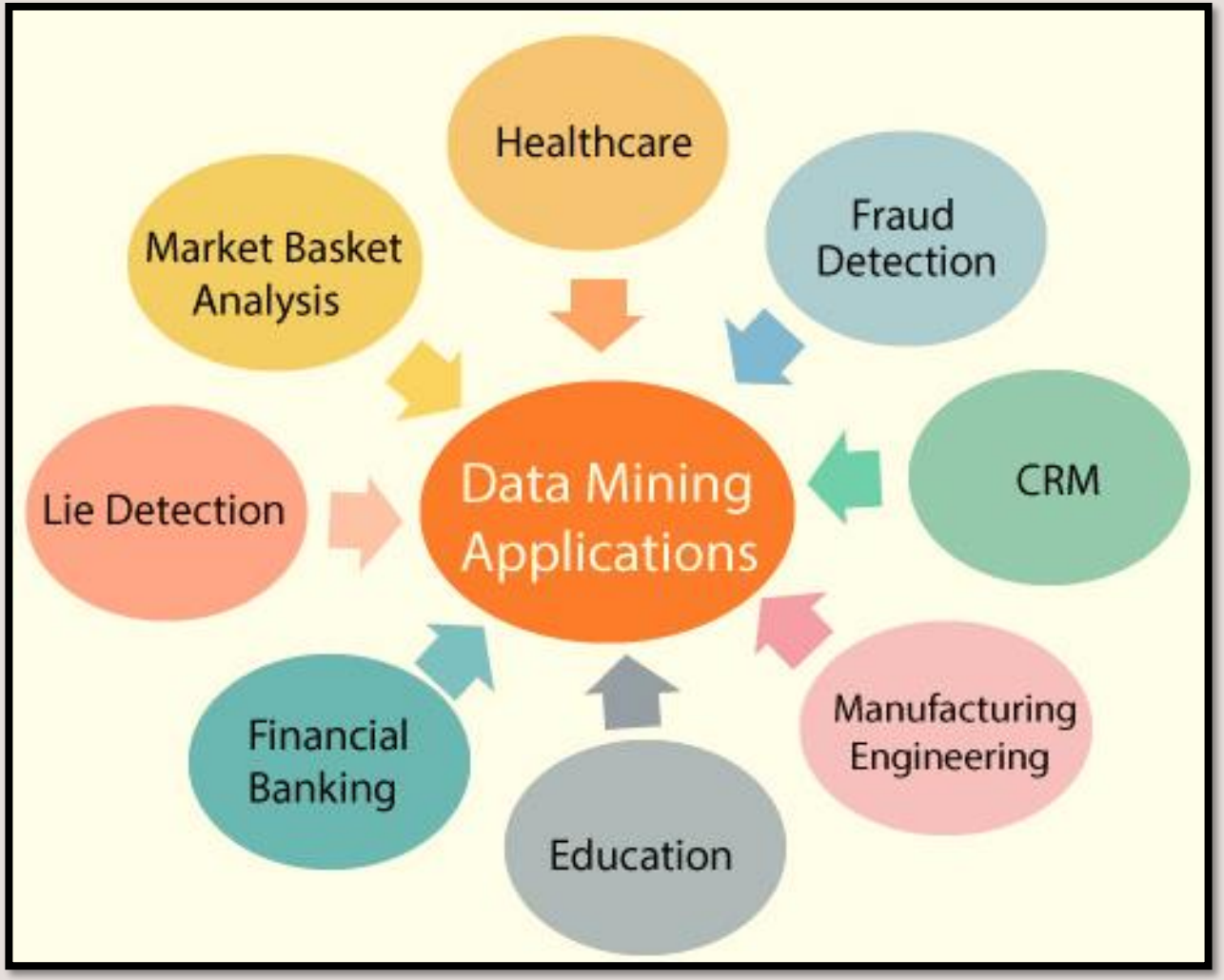
7. **Deployment -**

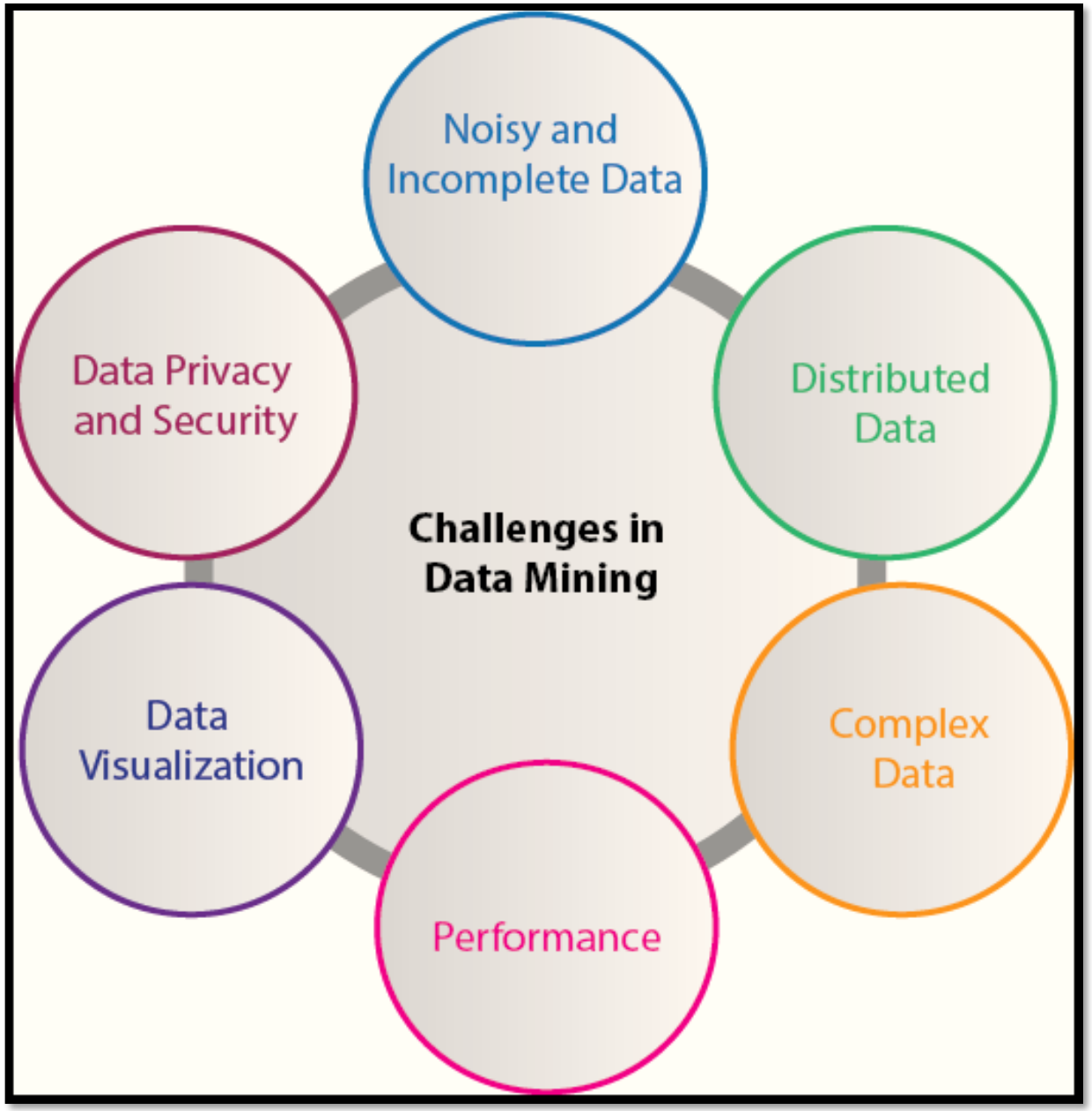
- The final step in the KDD process is to deploy the knowledge and insights gained from the data mining process to practical applications.
- This involves integrating the knowledge into decision-making processes or other applications to improve organizational efficiency and effectiveness.

Data Mining

- 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. i.e., 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.
- 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.
- 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 instruments and techniques available to mine data and find better insight from it.

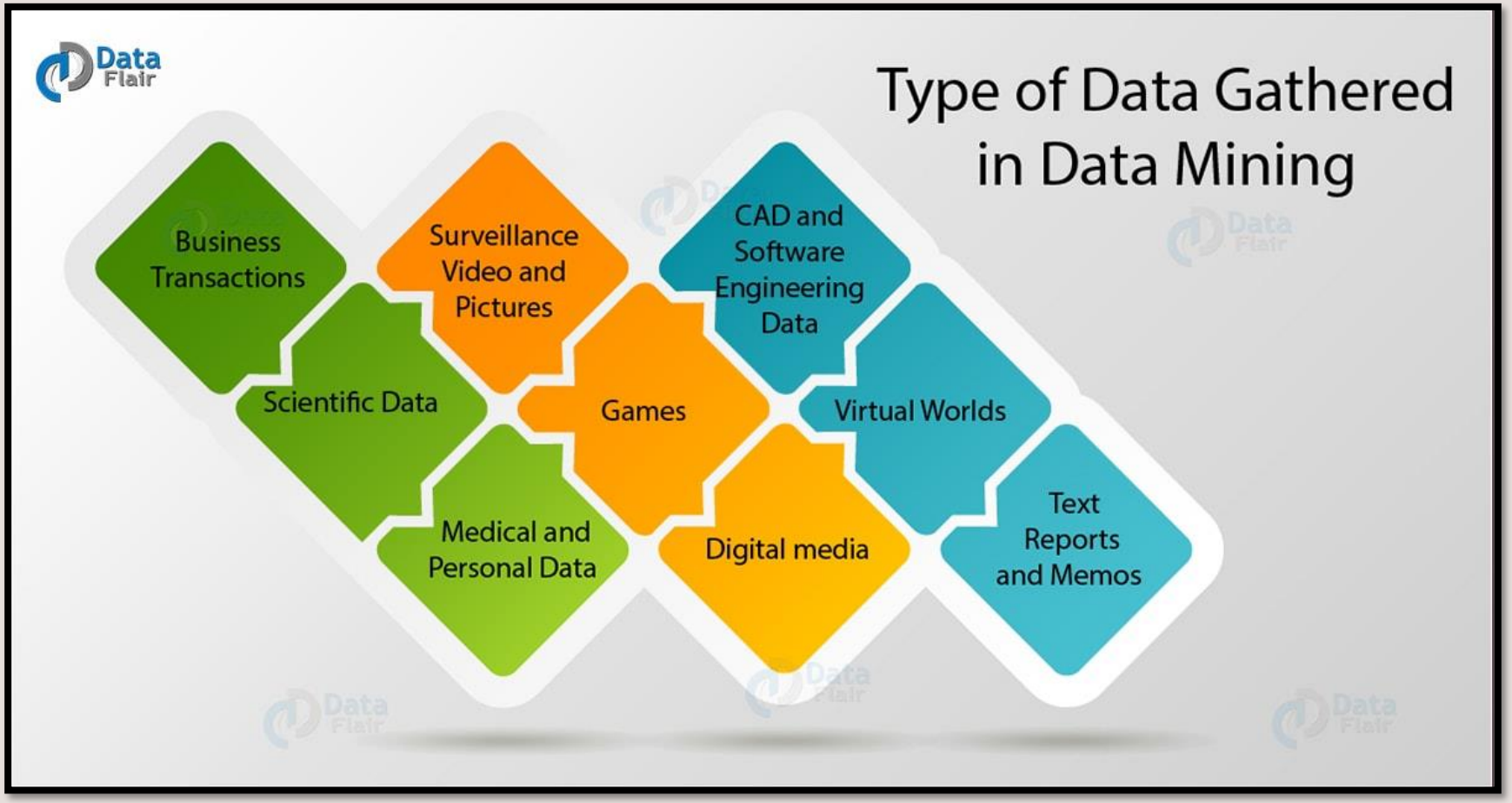
Applications of Data Mining

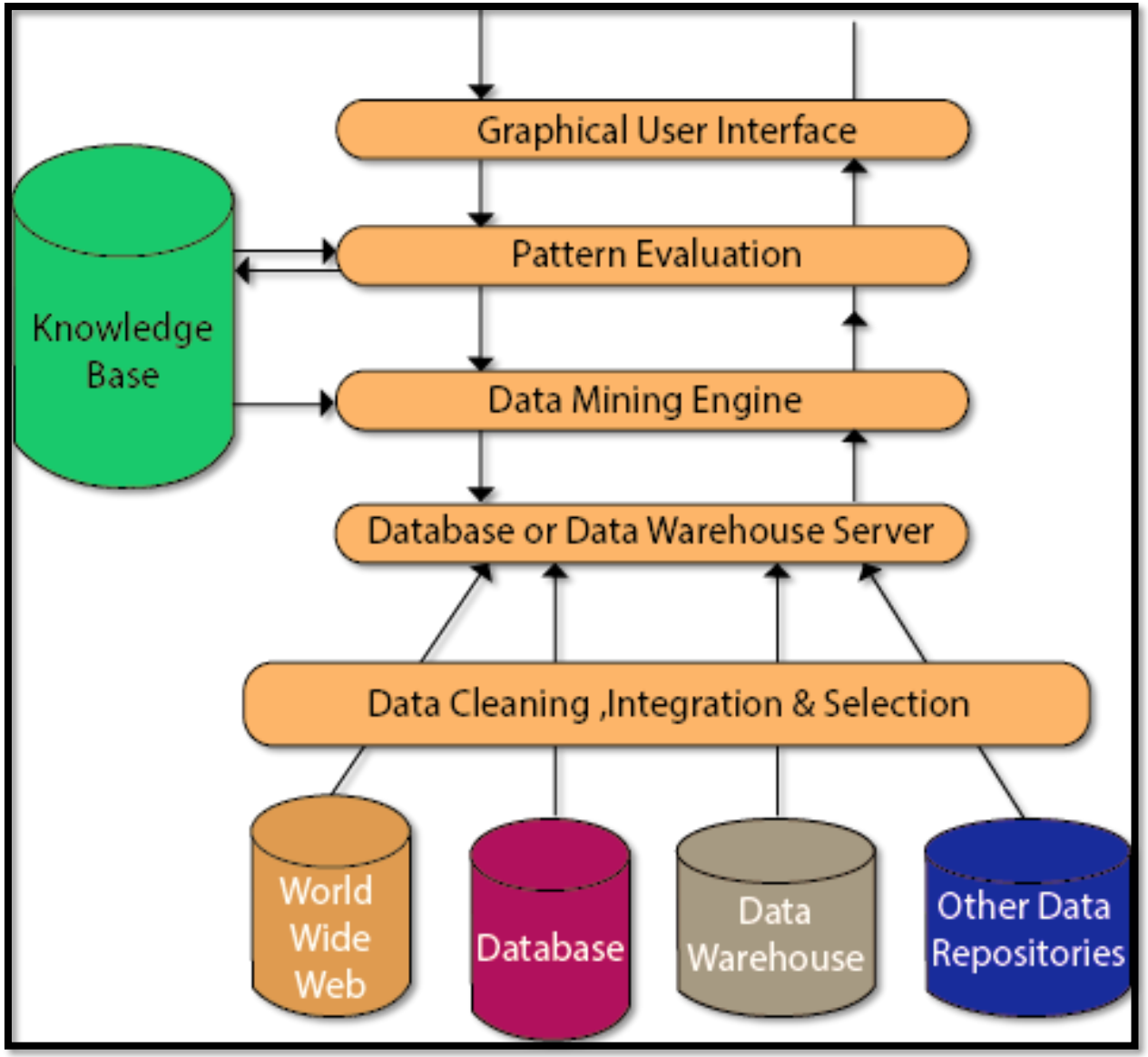




Challenges in Data Mining

Type of Data gathered



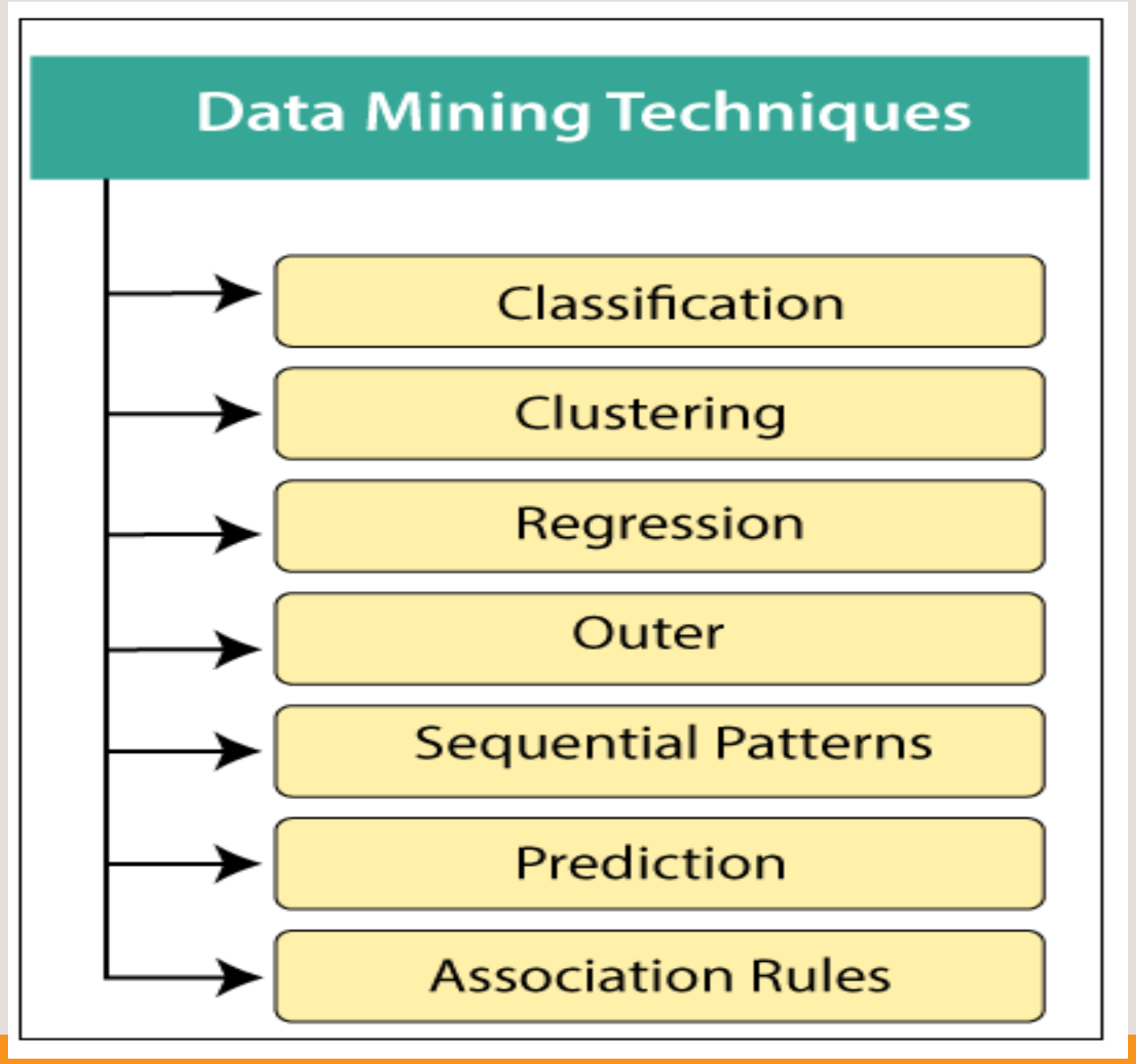


Data Mining Architecture

The components of Data Mining systems are:

1. Data source
2. Different Processes
3. Data warehouse Server
4. Data Mining Engine
5. Pattern Evaluation Module
6. Graphical User Interface
7. Knowledge base

Data Mining Techniques



1. Classification:

- This technique is used to obtain important and relevant information about data and metadata.

- This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

- **Classification of Data mining frameworks as per the type of data sources mined**
- **Classification of data mining frameworks as per the database involved**
- **Classification of data mining frameworks as per the kind of knowledge discovered**
- **Classification of data mining frameworks according to data mining techniques used**

2. Clustering

- Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data.
- Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

- Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor.
- It is used to define the probability of the specific variable.

4. Association Rules:

- This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

- Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases.

- Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

5. Outer detection:

- This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior.
- This technique may be used in various domains like intrusion, detection, fraud detection, etc.
- It is also known as Outlier Analysis or Outlier mining.

6. Sequential Patterns:

- The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns.
- In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

7. Prediction:

- Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc.

- It analyzes past events or instances in the right sequence to predict a future event.

Data Quality in Data Mining

- Data quality indicates how reliable a given dataset is. The data's quality will affect the user's ability to make accurate decisions regarding the subject of their study.
- Improved data quality leads to better decision-making across an organization. The more high-quality data you have, the more confidence you can have in your decisions. Good data decreases risk and can result in consistent improvements in results.
- Data quality meets six dimensions: accuracy, completeness, consistency, timeliness, validity, and uniqueness.

Measures to check Data Quality

Measurement Error: It refers to any problem resulting from the measurement process. In other words, it is the difference between measured and true value.

Data Collection Error: It refers to errors such as omitting data objects or attributes values, or including an unnecessary data object.

Noise: Noise is the random component of a measurement error. It involves either the distortion of a value or addition of objects that are not required. Techniques from signal and image processing are used to reduce noise. But, the removal of noise is a difficult task, hence much of the data mining work involves use of Robust Algorithms that can produce acceptable results even in the presence of noise.

Measures to check Data Quality

Precision: The closeness of repeated measurements (of the same quantity) to one another. It is often measured by the standard deviation of a set of values.

Bias: It is measured by taking the difference between the mean of the set of values and the known values of the quantity being measured. It can only be determined for those objects whose measured quantity is already known. For example, we have a standard laboratory weight with a mass of 1g and want to assess the precision and bias of our new laboratory scale. We weigh the mass five times, and obtain the following five values: {1.015, 0.990, 1.013, 1 .001, 0.986}.The mean of these values is 1.001, and hence, the bias is 0.001. The precision, as measured by the standard deviation, is 0.013

Accuracy: The closeness of measurements to the true value of the quantity being measured. Accuracy depends on precision and bias, but since it is a general concept, there is no specific formula for accuracy in terms of these two quantities.

Strategies for maintaining good Data Quality

- **Eliminate Data Objects or Attributes with missing values:** This is a simple and effective strategy. In this, if data set having a few objects that have missing values, we may omit them. A related strategy is to eliminate attributes that have missing values. This should be done with caution, however, since the eliminated attributes may be the ones that are critical to the analysis.
- **Estimate Missing Values:** Some missing data can be estimated reliably. If the attribute is continuous in nature, then the average of that attribute can be used in place of missing values. If the data is categorical, then the most occurring value can replace the missing values.
- **Ignore the Missing Values during Analysis:** Many data mining approaches can be modified to ignore missing values. For example, suppose that objects are being clustered and the similarity between pairs of data objects needs to be calculated. If one or both objects of a pair have missing values for some attributes, then the similarity can be calculated by using only the attributes that do not have missing values. It is true that the similarity will only be approximate, but unless the total number of attributes is small or the number of missing values is high, this degree of inaccuracy may not matter much.

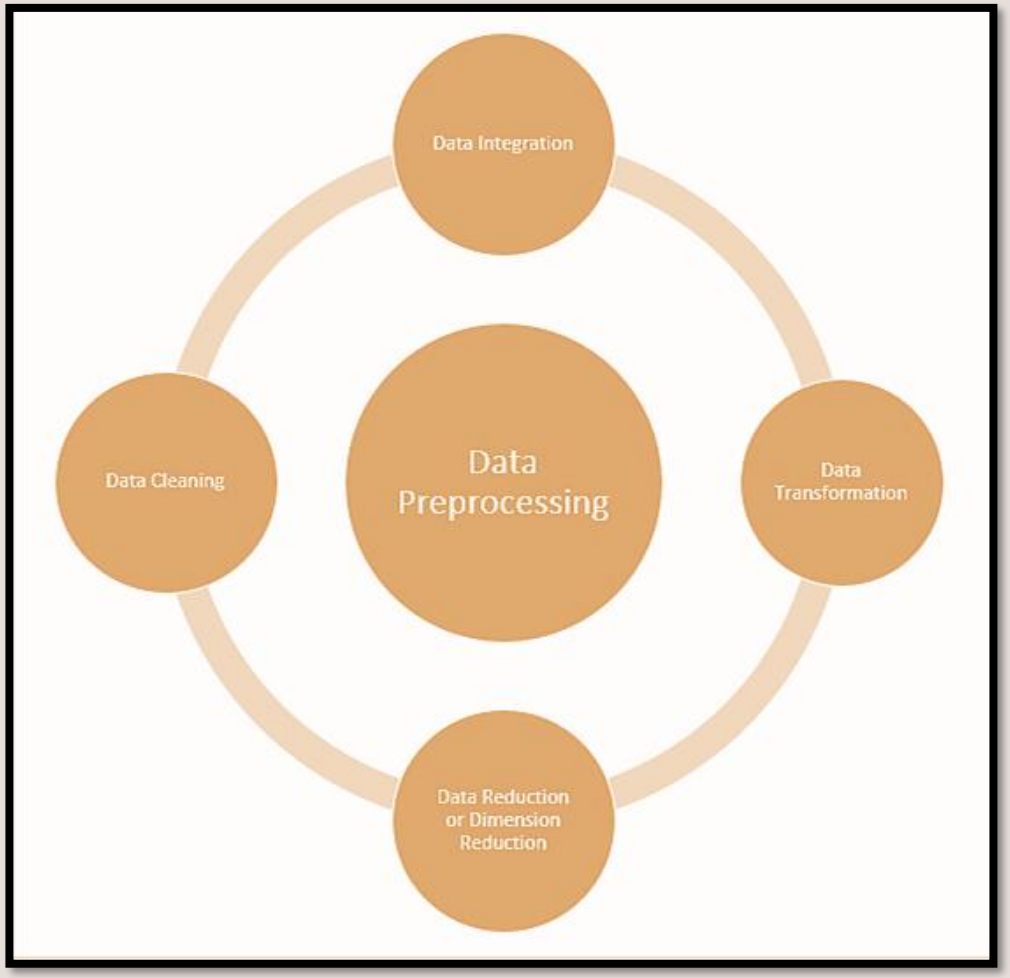
Strategies for maintaining good Data Quality

- **Inconsistent Values**: Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city. It may be that the individual entering this information transposed two digits, or perhaps a digit was misread when the information was scanned from a handwritten form. Some types of inconsistencies are easy to detect. For instance, a person's height should not be negative.

- **Duplicate Data:** A data set may include data objects that are duplicates. To detect and eliminate duplicates, two main issues must be addressed. First, if there are two objects that actually represent a single object, then the values of corresponding attributes may differ, and these inconsistent values must be resolved. Second, care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.

Data Preprocessing

- Raw, real-world data in the form of text, images, video etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design.
- Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.
- Major tasks in Data Preprocessing are:
  1. Data Cleaning
  2. Data Integration
  3. Data Reduction
  4. Data Transformation



Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It facilitates the automated discovery of hidden patterns as well as prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze huge amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money.
- Many data mining analytics software is difficult to operate and needs advance knowledge based training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to serious consequences in certain conditions.

Data Mining Tools

- Rapidminer
- Sas
- Rattle Data Mining
- Orange Data Mining
- DetaMelt Data Mining

Similarity Measure & Dissimilarity Measure

- **Similarity Measure:** Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1 (complete similarity). Similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative.
- **Dissimilarity Measure:** Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different). Dissimilarity is lower for more similar pairs of objects. Frequently, the term distance is used as a synonym for dissimilarity.
- **Proximity:** It refers to a similarity or dissimilarity.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Properties of Measures

Properties of Dissimilarity Measures/ Distance:

1. $d(x, y) \geq 0$ for all x and y, and $d(x, y) = 0$ if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all x and y. (Symmetry)
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, and z, where $d(x, y)$ is the distance (dissimilarity) between points (data objects), x and y. (Triangle Inequality)

❖A distance that satisfies these properties is called a Metric.

Properties of Similarity Measures:

1. $s(x, y) = 1$ (or maximum similarity) only if $x = y$
2. $s(x, y) = s(y, x)$ for all x and y, where $s(x, y)$ is the similarity between data objects, x and y.


Thank
you