DATA MINING

Department of Computer Applications

Course: 23CAT705- RESEARCH METHODOLOGY

UNIT I : RESEARCH DESIGN

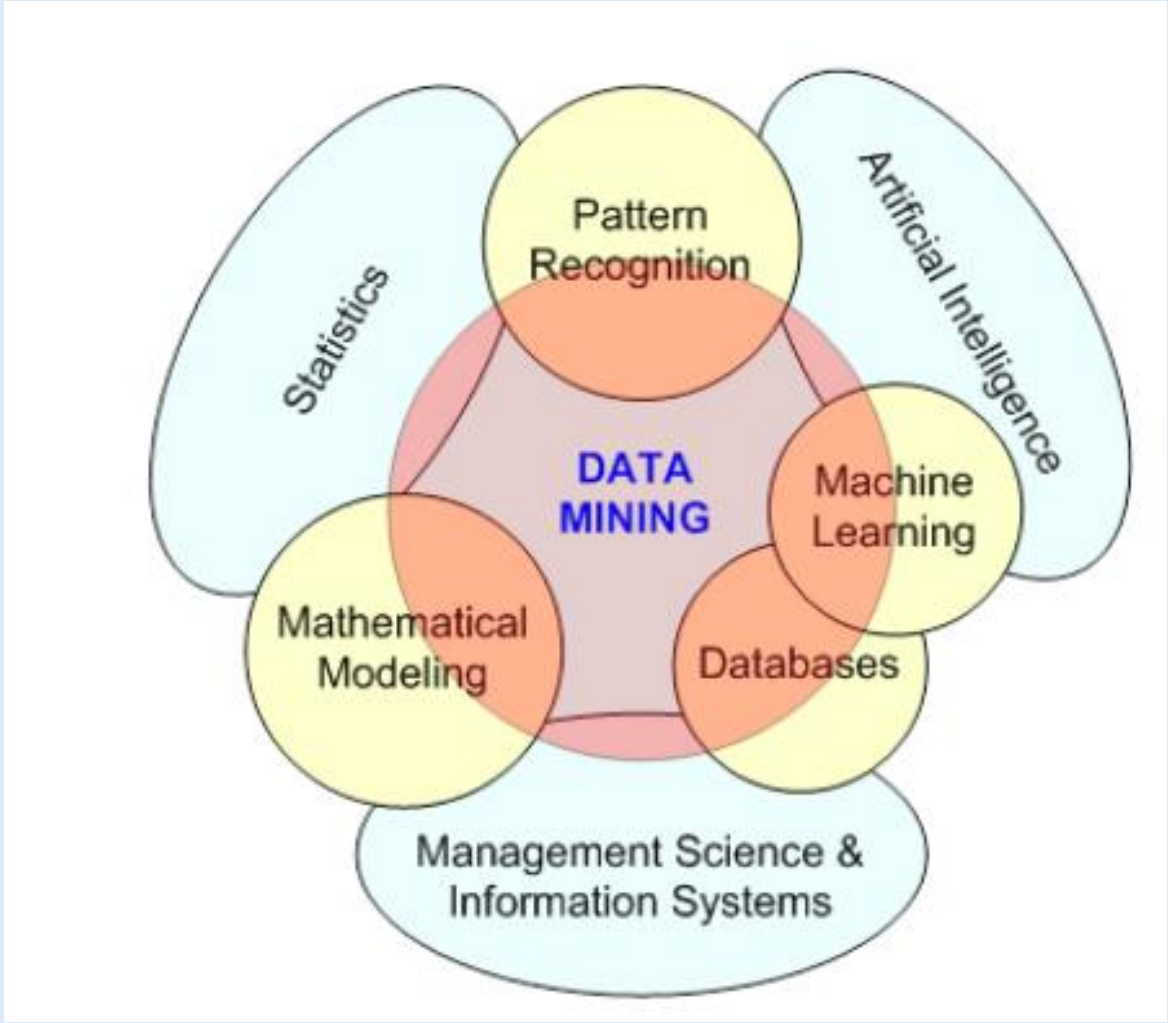
Class: II MCA / III SEMESTER

Data mining

- Extraction of useful patterns/hidden knowledge/insight from data sources, e.g., databases, texts, web, image
- Patterns must be valid, novel, potentially useful, understandable
- It is also called **knowledge discovery and data mining (KDD)**
- Patterns might be different data mining tasks

Data Mining

An emerging multi-disciplinary field connects



Quantitative Methods

Data Come from Everywhere

But, they have different form

Hospital

Weather Station

Social Media

Data Types

Record Data
- Transactional Data

Temporal Data
- Time Series Data
- Sequence Data

Spatial & Spatial-Temporal Data
- Spatial Data
- Spatial-Temporal Data

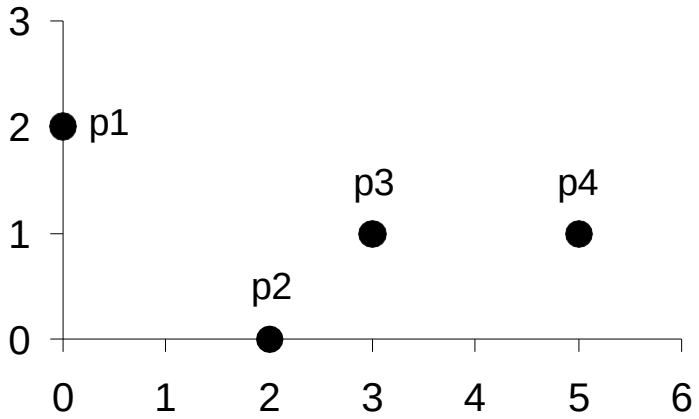
Graph Data
- Transactional Data

UnStructured Data
- Twitter Status Message
- Review, news article

Semi-Structured Data
- Paper Publications Data
- XML format

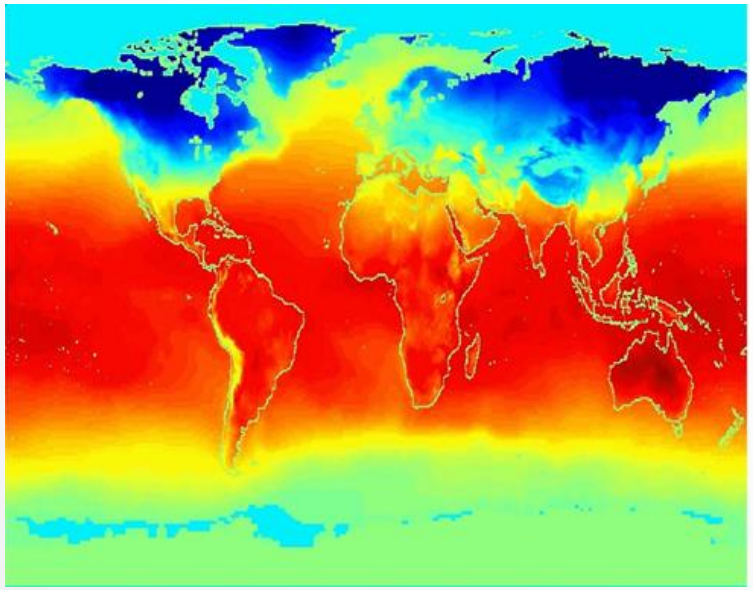
Time Series

Time Series Data

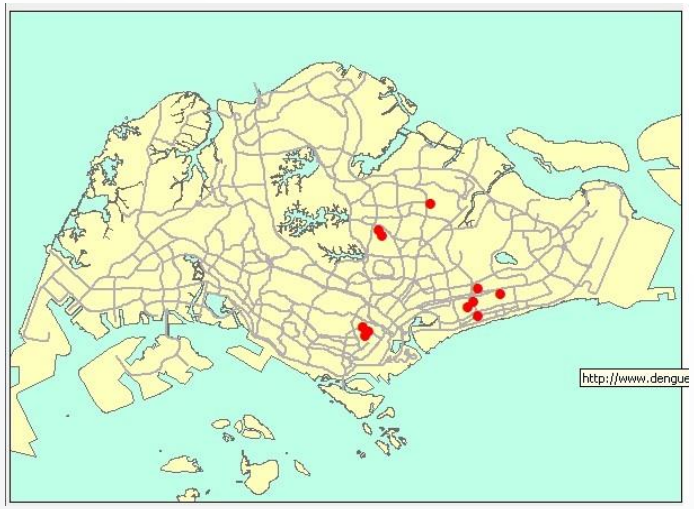


Distance Matrix

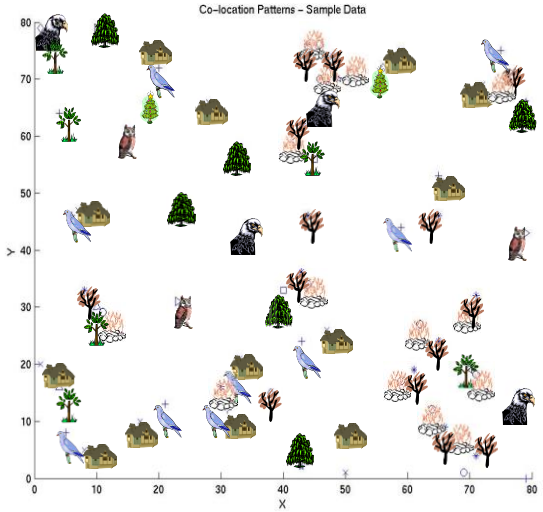
Spatial Data



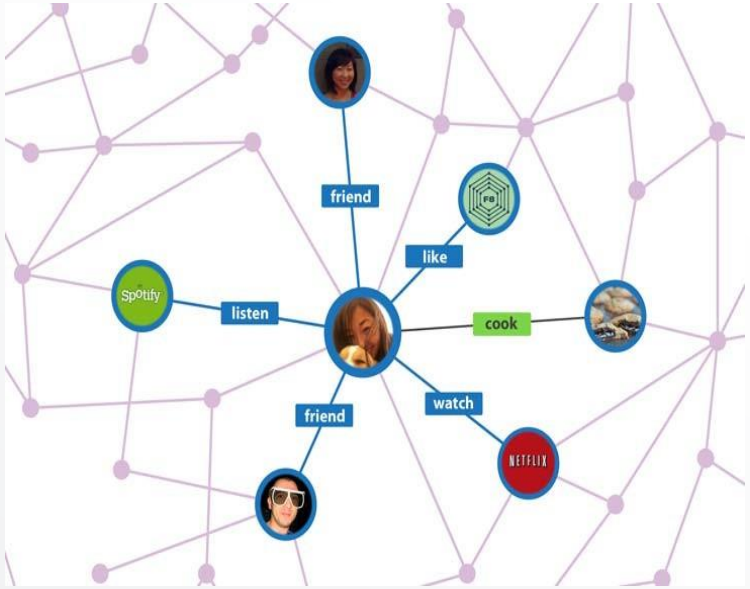
Average Monthly Temperature of land and ocean



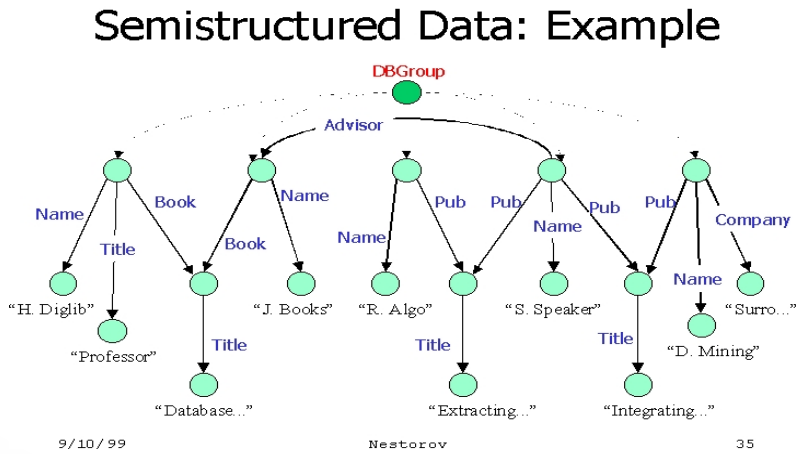
Dengue Disease Dataset (Singapore)



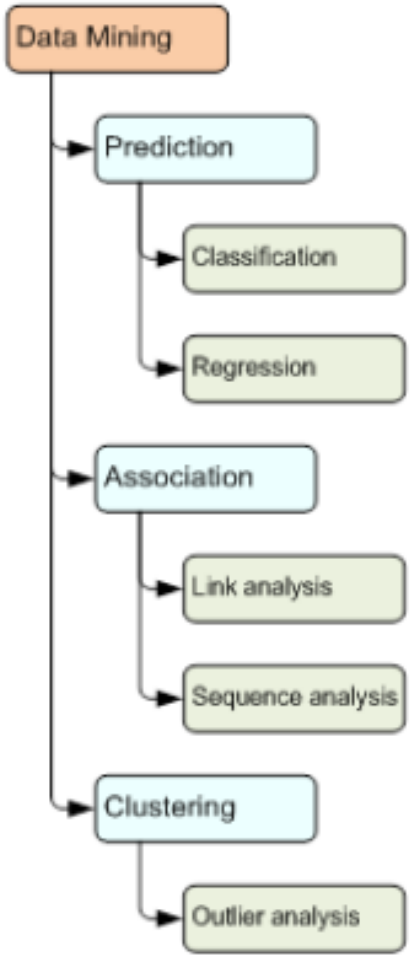
Data



Graph Data



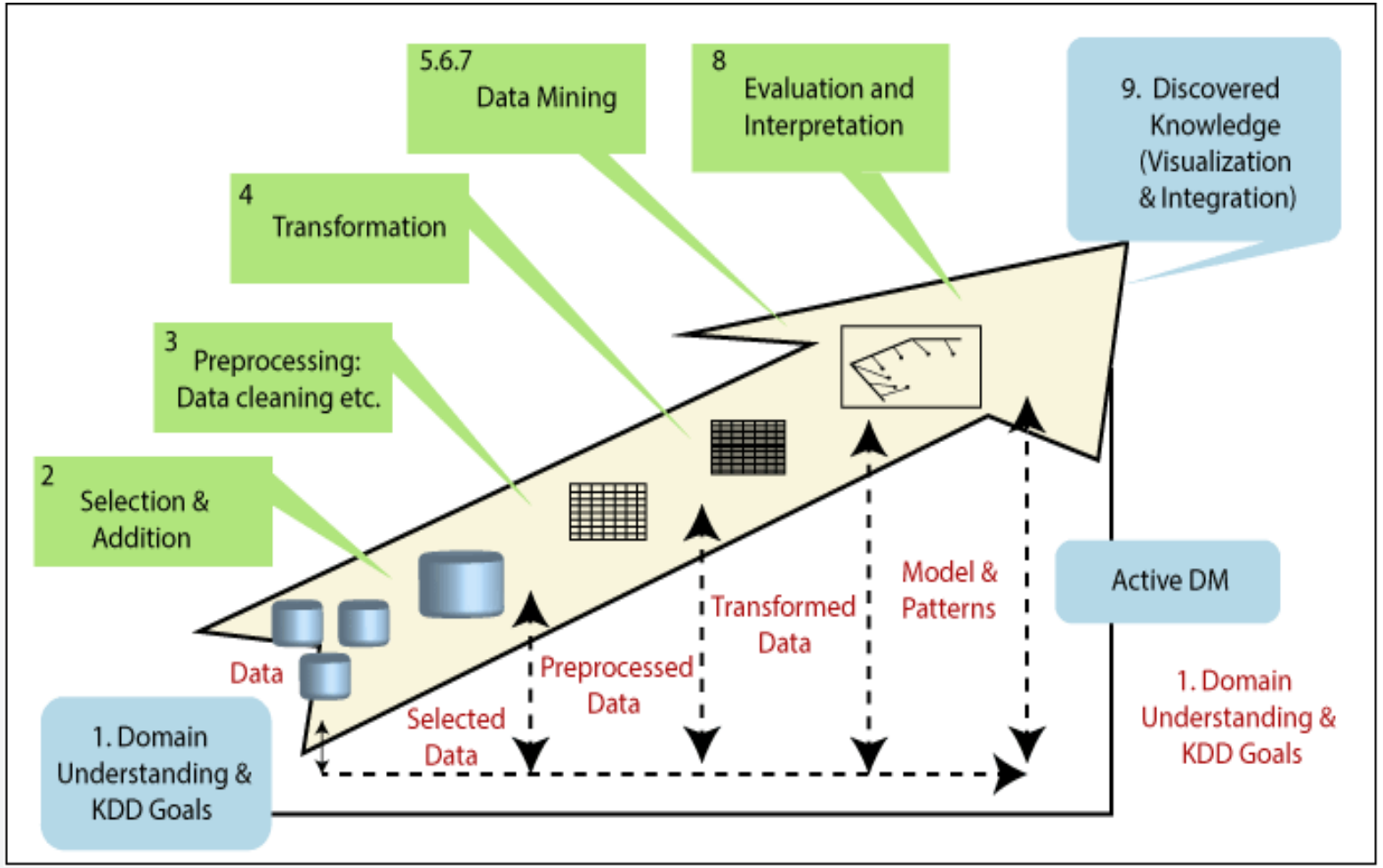
Data Mining Tasks



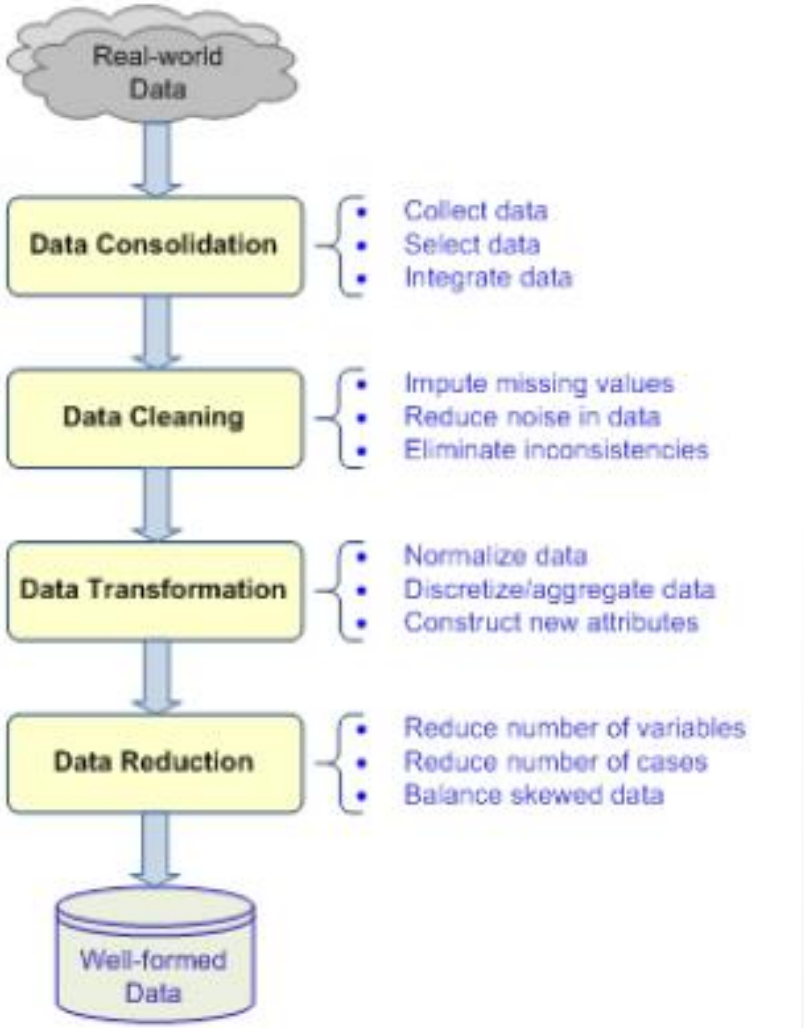
Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

KDD Process

Knowledge Discovery in Databases (KDD) refers to the **process** of extracting useful insights, patterns, and knowledge from large volumes of **data**



Data Preparation



Data mining Techniques

Classification

- Categorizing data into predefined classes (e.g., spam vs. non-spam)

Clustering

- Grouping similar data points together without predefined labels (e.g., customer segmentation)

Association rule mining

- mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items

"80% of customers who buy *cheese* and *milk* also buy *bread*, and 5% of customers buy all of them together"

Cheese, Milk→ Bread [sup =5%, confid=80%]

Sequential pattern mining

- identifying a set of similarity groups in the data

Primary Data Collection

Deviation detection

- discovering the most significant changes in data

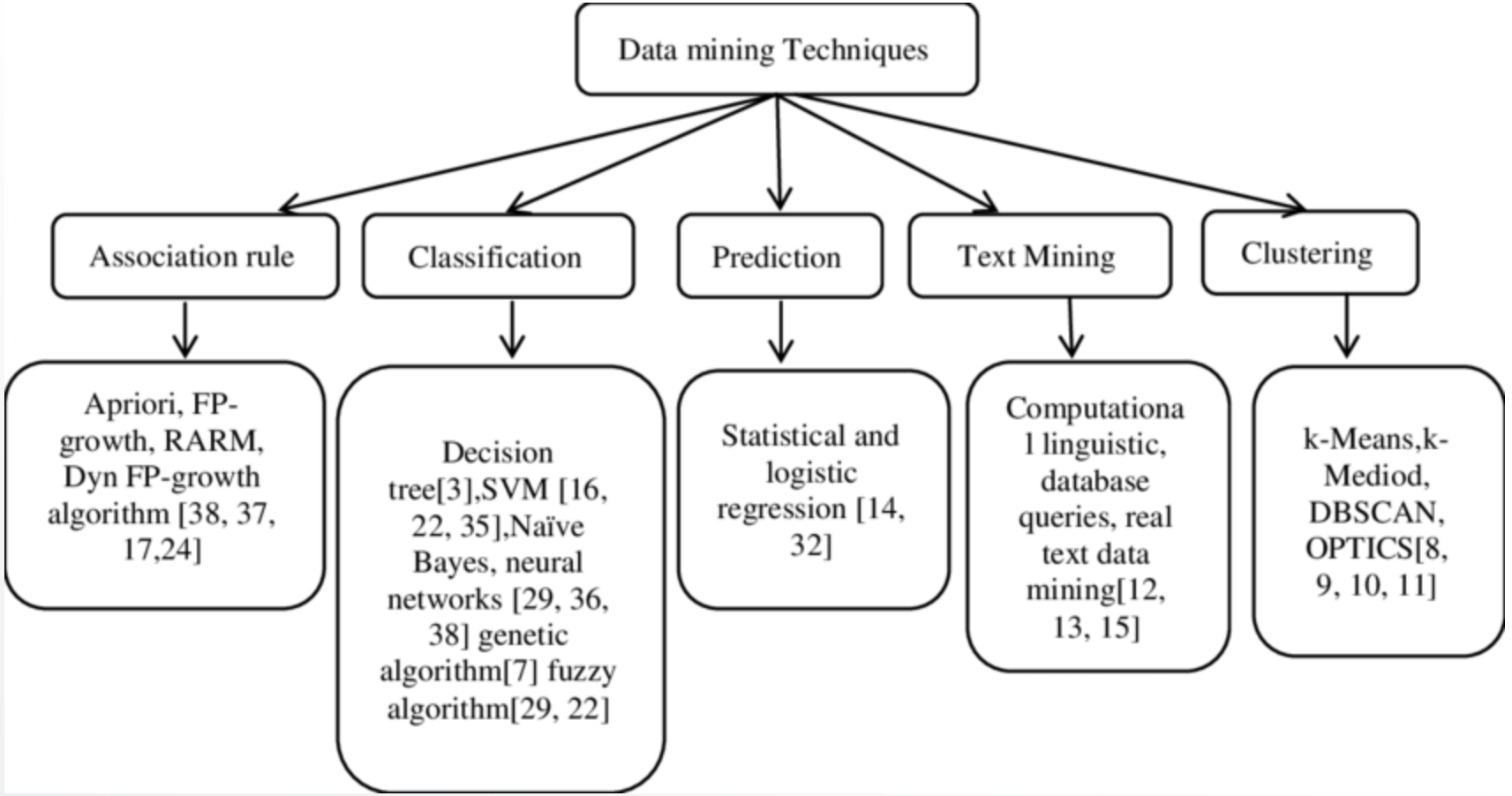
Data visualization

- using graphical methods to show patterns in data

Anomaly Detection

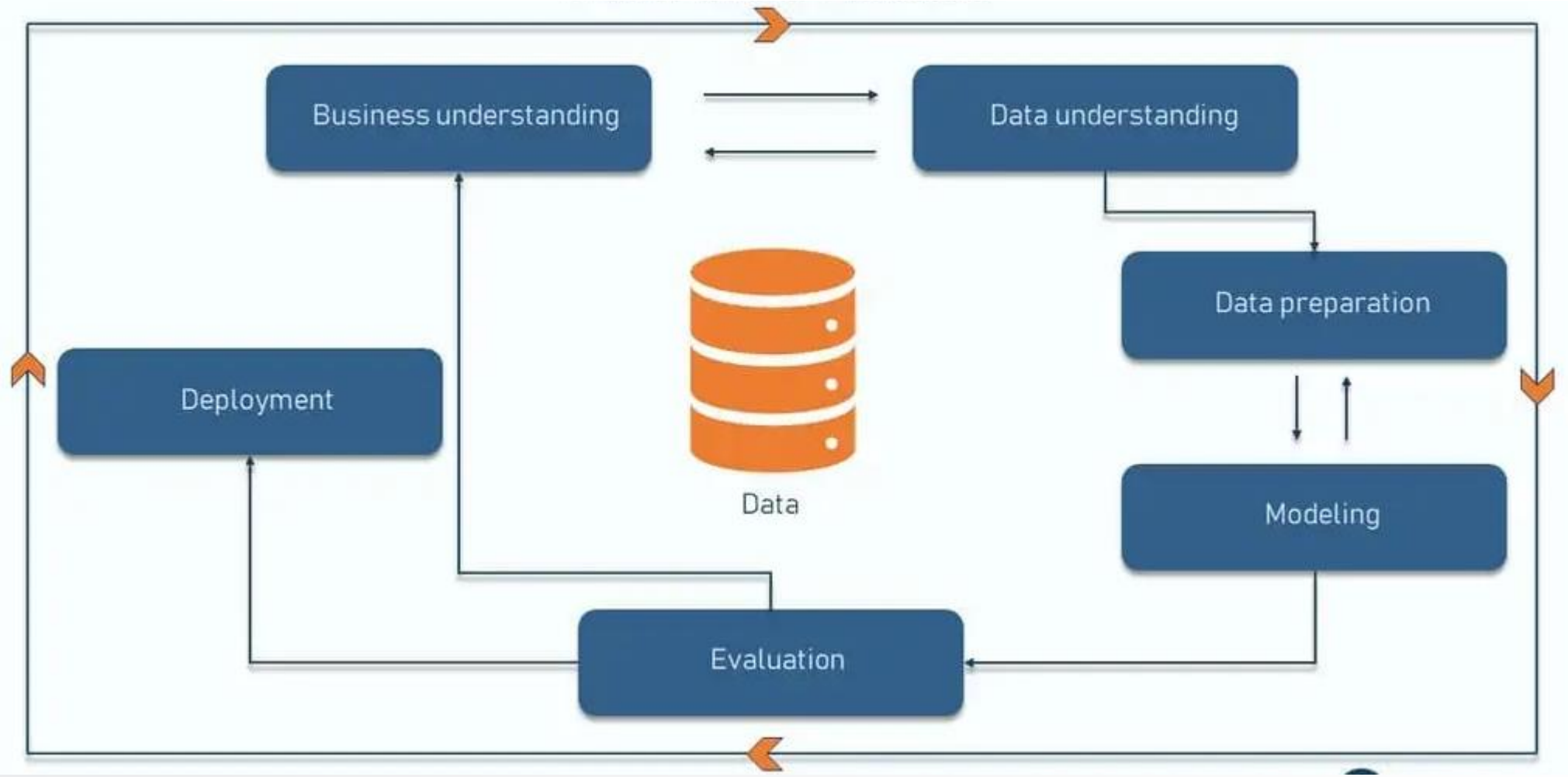
- Identifying rare or unusual data points (e.g., fraud detection)

Data Mining Tasks



Data Mining Process

CRoss Industry Standard Process for **Data Mining** (**CRISP-DM**)



Challenges

Data Quality: Incomplete, noisy, or biased data can impact results.

Scalability: Large datasets can be challenging to process efficiently.

Privacy and Security: Ensuring sensitive data is protected.

Interpretability: Making the results understandable to non-technical stakeholders.

Overfitting: Models that are too complex may not generalize well

References

1. Kothari, C.R. &Garg, G. (2019). *Research Methodology: Methods and Techniques*. New Age International Publishers, New Delhi

2. Goode, W.J. &Hatt, P.K. (2022). Methods in Social Research. McGraw Hill, London

3. Bhandarkar, P.L. & Wilkinson, T.S. (2016). Methodology and Techniques of Social Research. Himalaya Publishing House, Mumbai.

Thank You