DATA MINING

CHEKURI SRI SUMANTH,

ASST. PROFESSOR, CSE

INTRODUCTION TO DATA MINING

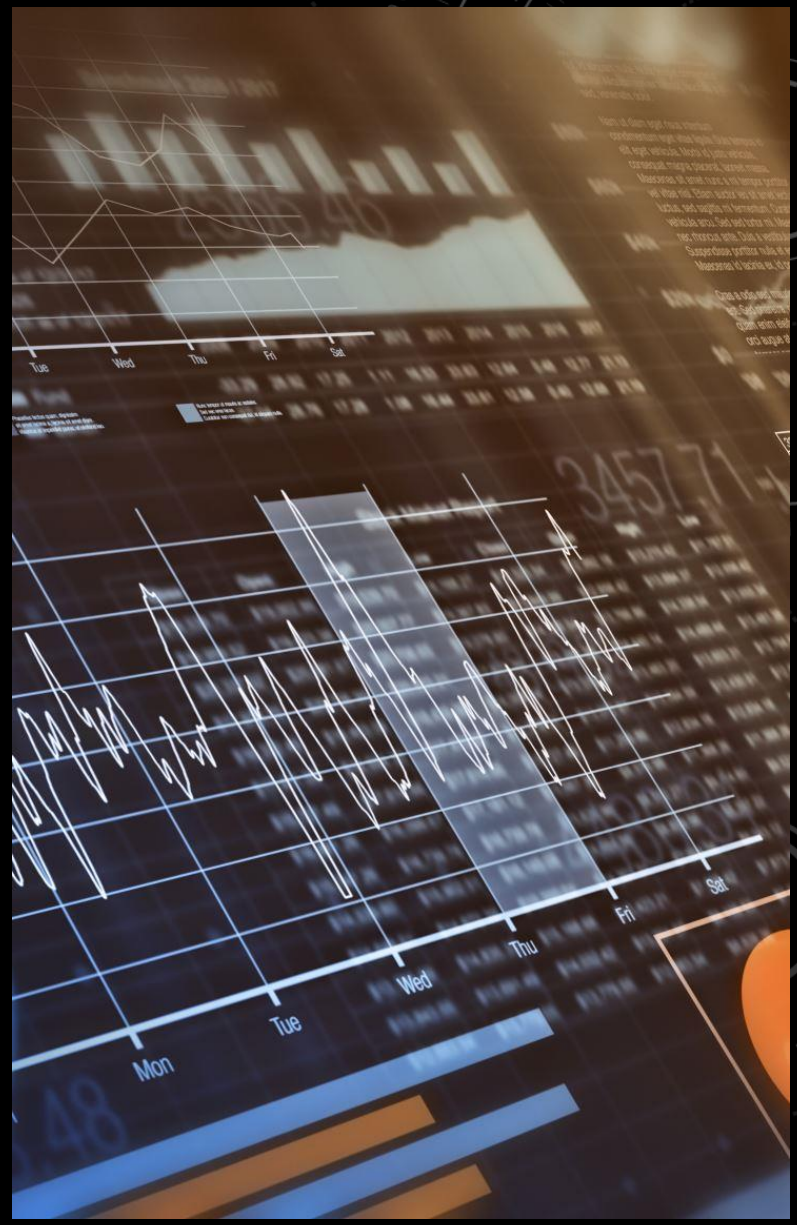
- **Definition**: Data mining is the process of discovering meaningful patterns, correlations, and insights from large volumes of data. It involves using various techniques and algorithms to extract valuable knowledge, hidden within structured or unstructured data.
- Key Points:
  - Data mining involves uncovering hidden information.
  - It uses statistical, mathematical, and machine learning techniques.
  - The goal is to turn raw data into actionable insights.

4 Stages: Data Collection, Preparation, Exploration, Interpretation & analysis.

INTRODUCTION TO DATA MINING

Importance of Data Mining:

- Enhances Decision-Making: Data mining helps organizations make informed decisions based on data-driven insights.
- Identifies Trends: It identifies trends, patterns, and anomalies in data that may not be apparent through traditional analysis.
- Predictive Capabilities: Data mining enables predictive modeling, allowing organizations to anticipate future trends and behaviors.
- Customer Understanding: It helps in understanding customer preferences, behaviors, and needs.
- Competitive Advantage: Effective data mining can provide a competitive edge in various industries.



INTRODUCTION TO DATA MINING

Real-world Applications:

- Healthcare:
  - Patient diagnosis and treatment planning.
  - Identifying disease outbreaks and trends.
- Retail:
  - Customer segmentation and targeting.
  - Inventory optimization and demand forecasting.
- Finance:
  - Fraud detection and prevention.
  - Stock market analysis and risk assessment.

- **Marketing:**
  - Personalized marketing and recommendation systems.
  - Churn prediction and customer retention.
- **Manufacturing:**
  - Quality control and defect detection.
  - Supply chain optimization.
- **Social Media:**
  - Sentiment analysis and social network analysis.
  - Content recommendation.

DATA MINING AND KNOWLEDGE DISCOVERY (KDD)

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.

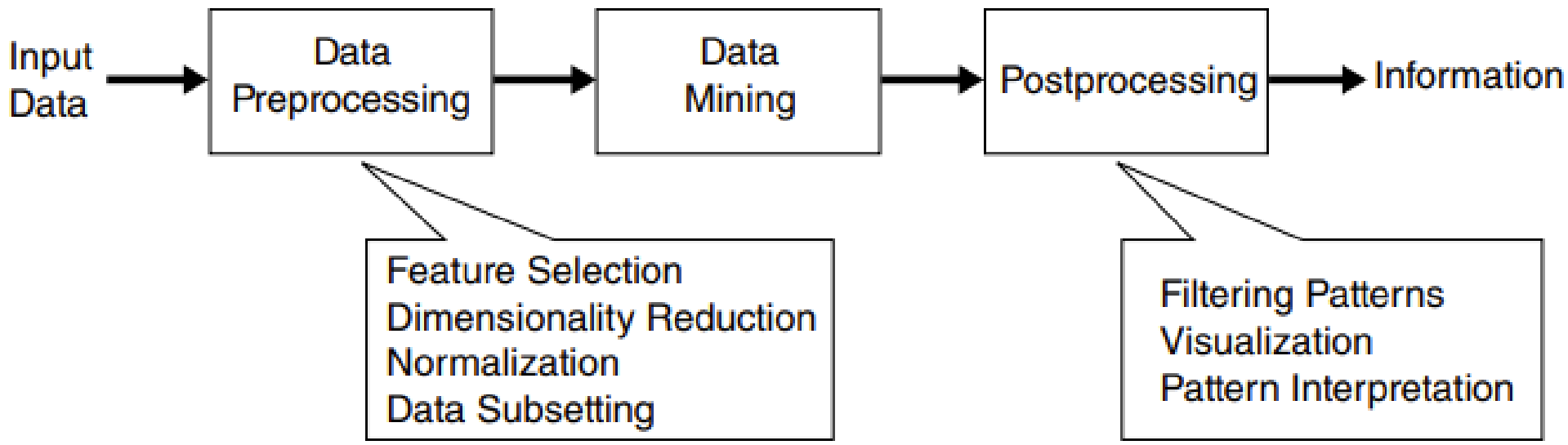
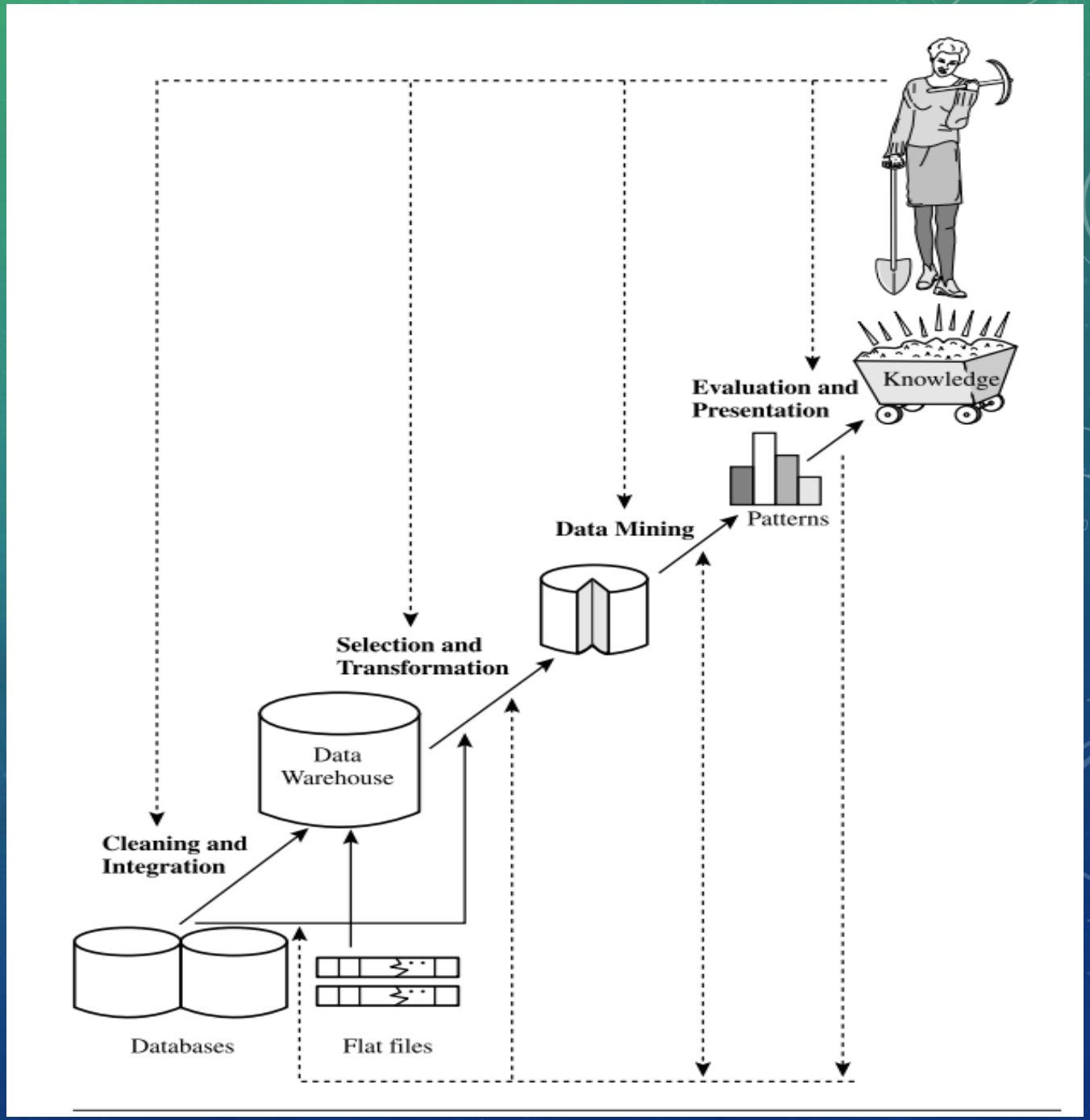


Figure 1.1. The process of knowledge discovery in databases (KDD).

KDD PROCESS


Evaluation and Presentation
Knowledge
Data Mining
Patterns
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

DATA MINING FUNCTIONALITIES

- Characterization and Discrimination
- Classification
- Prediction
- Association Analysis
- Cluster Analysis
- Outlier Analysis
- Evolution & Deviation Analysis

- **Characterization and Discrimination**:
  - **Class/Concept Description**: Characterization involves describing the general features or properties of a dataset, while discrimination focuses on finding features that distinguish one class or group from another within the dataset.
  - **Example**: In a healthcare dataset, characterization might involve calculating average patient ages and identifying common medical conditions. Discrimination, on the other hand, could be used to find differences in health indicators between patients with diabetes and those without diabetes.
- **Classification**:
  - **Class/Concept Description**: Classification is the process of assigning predefined categories or labels to data instances based on their features or attributes.
  - **Example**: Spam email classification involves categorizing incoming emails as either "spam" or "not spam" based on features like keywords, sender, and content.
- **Prediction**:
  - **Class/Concept Description**: Prediction involves forecasting or estimating future values or outcomes based on historical data and patterns.
  - **Example**: Stock price prediction uses historical stock price data, trading volume, and other financial indicators to predict the future price of a stock.
- **Association Analysis**:
  - **Class/Concept Description**: Association analysis aims to discover interesting relationships or associations between items or attributes in a dataset.
  - **Example**: Market basket analysis can identify associations between products that are frequently purchased together, such as finding that customers who buy bread are also likely to buy butter.

- **Cluster Analysis**:
  - **Class/Concept Description**: Cluster analysis groups data points that are similar to each other into clusters or segments, helping to uncover natural patterns or groupings in the data.
  - **Example**: Customer segmentation in marketing can involve clustering customers based on their purchase history and behavior to create targeted marketing strategies for different customer groups.
- **Outlier Analysis**:
  - **Class/Concept Description**: Outlier analysis identifies data points that deviate significantly from the norm or expected behavior in a dataset.
  - **Example**: Anomaly detection in network security can detect unusual network traffic patterns that may indicate a cyberattack.
- **Evolution & Deviation Analysis**:
  - **Class/Concept Description**: Evolution analysis examines how data changes over time, while deviation analysis focuses on identifying deviations or anomalies in time series data.
  - **Example**: Tracking the monthly sales performance of a product (evolution analysis) and identifying a sudden and unexplained drop in sales in a particular month (deviation analysis).

CLASSIFICATION OF DATA MINING SYSTEMS

There is a large variety of data mining systems available. Data mining systems may integrate techniques from the following –

- Spatial Data Analysis
- Information Retrieval
- Pattern Recognition
- Image Analysis
- Signal Processing
- Computer Graphics
- Web Technology
- Business
- Bioinformatics

CLASSIFICATION OF DATA MINING SYSTEMS

A data mining system can be classified according to the following criteria –

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines

Apart from these, a data mining system can also be classified based on the kind of (a) databases mined, (b) knowledge mined, (c) techniques utilized, and (d) applications adapted.

CLASSIFICATION BASED ON THE DATABASES MINED

This classification is based on different criteria:

- Data Models:
  - Relational mining systems
  - Transactional mining systems
  - Object-relational mining systems
  - Data Warehouse mining systems
- Type of Data:
  - Spatial Data
  - Time-Series Data
  - Text Data
  - Stream Data
- Type of Application:
  - Multimedia Data Mining Systems
  - World Wide Web mining Systems

CLASSIFICATION BASED ON THE TYPE OF KNOWLEDGE MINED

A data mining system categorized based on the kind of knowledge mind may have the following functionalities:

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

CLASSIFICATION BASED ON THE TECHNIQUES UTILIZED

Based on the degree of user interactions involved:

- Autonomous Systems
- Interactive Exploratory Systems
- Query Driven Systems

CLASSIFICATION ACCORDING TO THE APPLICATIONS ADAPTED:

- Finance
- Telecommunication
- DNA
- Stock Market
- E-mail
- Integrated Application specific systems

DATA MINING TASK PRIMITIVES

- **A data mining task can be specified in the form of a data mining query, which is input to the data mining system.**
- **These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process or examine the findings from different angles or depths**

Set of task-relevant data to be mined (the relevant attributes or dimensions)

Kind of knowledge to be mined (data mining functions to be performed)

Background knowledge to be used in the discovery process (knowledge base – concept hierarchy, user beliefs)

Interestingness measures and thresholds for pattern evaluation (Interestingness measures for association rules are ‘support’ and ‘confidence’)

Representation for visualizing the discovered patterns (Tables, charts, graphs, decision trees, cubes)

DATA MINING TASK PRIMITIVES

Task relevant data

- Database Name
- Database tables
- Relevant attributes
- Data grouping criteria

Type of knowledge to be mined

- Classification
- Clustering
- Prediction
- Discrimination
- Correlation analysis

Background knowledge

- Concept Hierarchy
- User beliefs about relationships in data

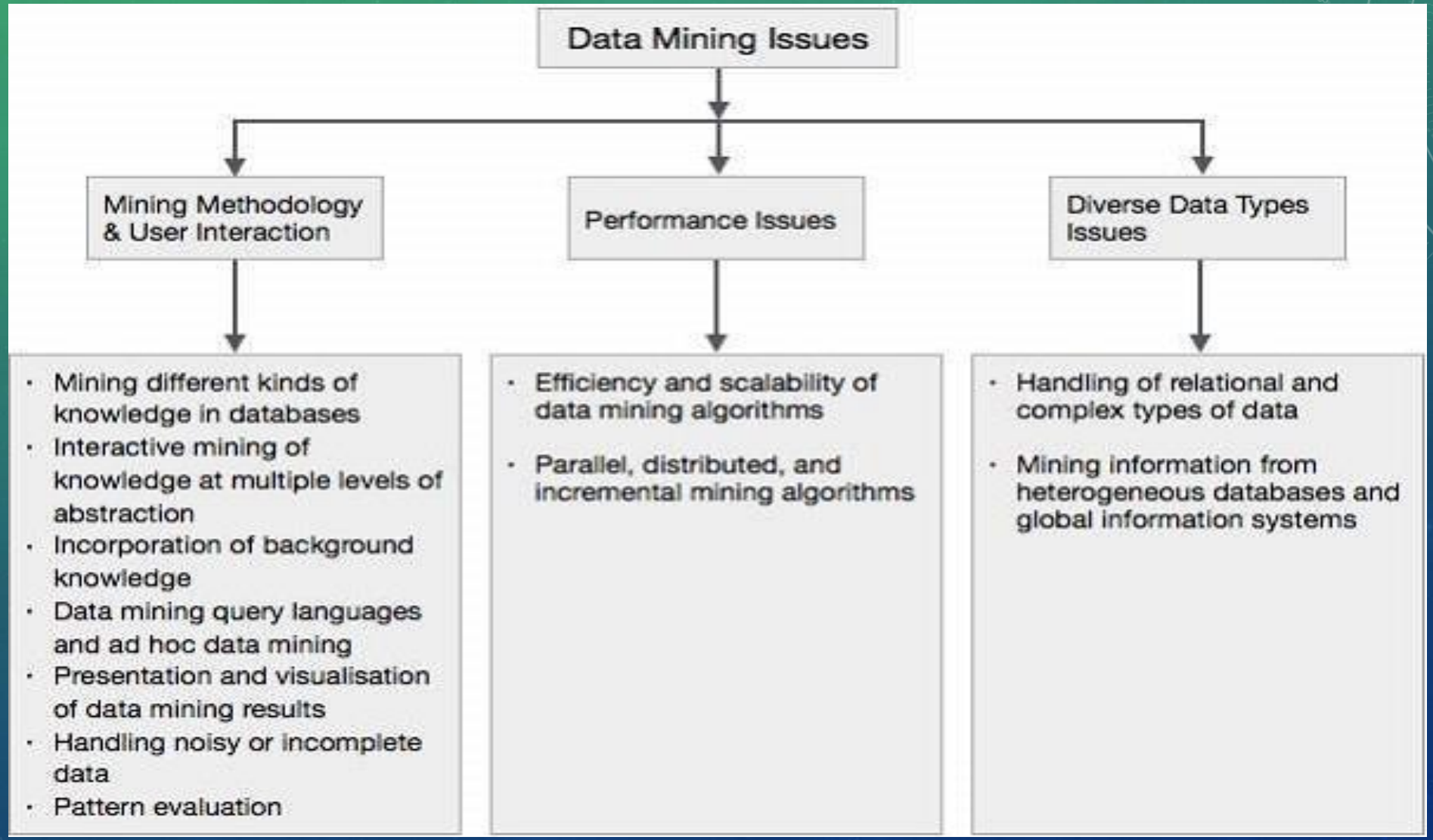
Measures of patterns

- Simplicity
- Novelty
- Certainty
- Utility

Visualization of patterns

- Visualization of discovered patterns
- Cubes
- Charts
- Tables
- Graphs

MAJOR ISSUES IN DATA MINING



DATA PREPROCESSING

Why Data Pre-processing?

- Data in the real world is dirty. That is it is incomplete or noisy or inconsistent.
- Incomplete: means lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

e.g., occupation=“ ”

- Noisy: means containing errors or outliers

e.g., Salary=“-10”

- Inconsistent: means containing discrepancies in codes or names

e.g., Age=“42” Birthday=“03/07/1997”

e.g., Was rating “1,2,3”, now rating “A, B, C”

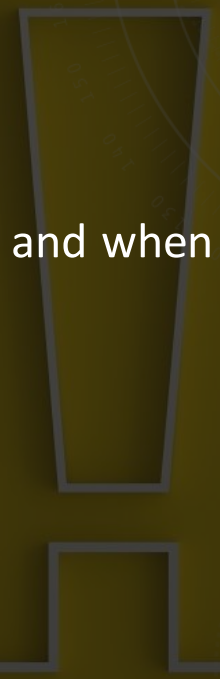
e.g., discrepancy between duplicate records

DATA PREPROCESSING

Why Is Data Dirty?

Data is dirty because of the below reasons.

1. Incomplete data may come from
   - "Not applicable" data value when collected
   - Different considerations between the time when the data was collected and when it is analyzed.
   - Human / hardware / software problems
2. Noisy data (incorrect values) may come from
   - Faulty data collection instruments
   - Human or computer error at data entry
   - Errors in data transmission
3. Inconsistent data may come from
   - Different data sources
   - Functional dependency violation (e.g., modify some linked data)
4. Duplicate records also need data cleaning



DATA PREPROCESSING

Why Is Data Pre-processing Important?

- If there is No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprise the majority of the work of building a data warehouse

Multi-Dimensional Measure of Data Quality

A well-accepted multidimensional view has the following properties:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Value added
- Interpretability
- Accessibility

MAJOR TASKS IN DATA PRE-PROCESSING

Data cleaning

- Data Cleaning includes, filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Data integration

- Data Integration includes integration of multiple databases, data cubes, or files.

Data transformation

- Data Transformation includes normalization and aggregation.

Data reduction

- Data reduction is achieved by obtaining reduced representation of data in volume but produces the same or similar analytical results.

Data Discretization

- Data Discretization is part of data reduction but with particular importance, especially for numerical data.

MAJOR TASKS IN DATA PRE-PROCESSING

Data cleaning

- Data Cleaning includes, filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Data integration

- Data Integration includes integration of multiple databases, data cubes, or files.

Data transformation

- Data Transformation includes normalization and aggregation.

Data reduction

- Data reduction is achieved by obtaining reduced representation of data in volume but produces the same or similar analytical results.

Data Discretization

- Data Discretization is part of data reduction but with particular importance, especially for numerical data.

1. DATA CLEANING:

- The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  - **Ignore the tuples:**
    This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  - **Fill the Missing values:**
    There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- **(b). Noisy Data:**
  Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  - **Binning Method:**
    This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  - **Regression:**
    Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  - **Clustering:**
    This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. DATA TRANSFORMATION:

- This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:
- **Normalization:**
  It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

- **Attribute Selection:**
  In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

- **Discretization:**
  This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

- **Concept Hierarchy Generation:**
  Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. DATA REDUCTION:

- Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:
- **Feature Selection:** This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).
- **Feature Extraction:** This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).
- **Sampling:** This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.
- **Clustering:** This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.
- **Compression:** This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

DATA INTEGRATION