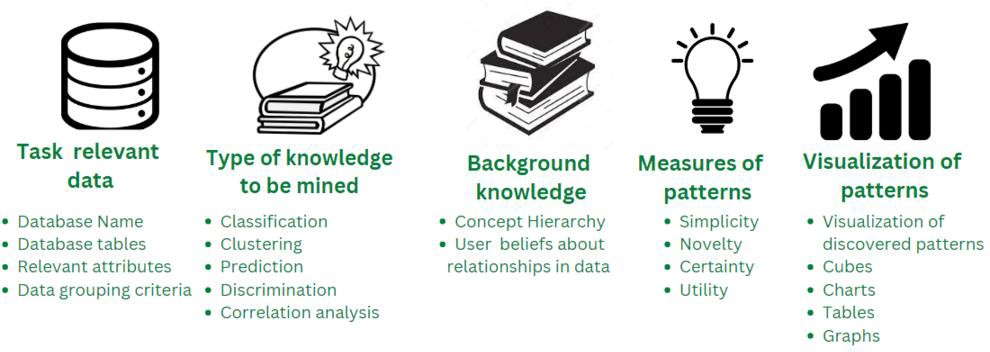
# UNIT-2 DATA MINING PRIMITIVES, LANGUAGE, ARCHITECTURE

**DATA MINING TASK PRIMITIVES:**

* **Data mining task primitives** are the basic building blocks that define a data mining task. They allow users to specify **what data to use**, **what kind of knowledge to discover**, and **how to evaluate the results**.
* A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives.
* A data mining query language can be designed to incorporate these primitives, allowing users to interact with data mining systems flexibly.

## The key primitives are:

* 1. Set of task-relevant data to be mined.
  2. Kind of knowledge to be mined.
  3. Background knowledge to be used in the discovery process.
  4. Interestingness measures and thresholds for pattern evaluation.
  5. Representation for visualizing the discovered patterns.



## Set of task-relevant data to be mined:

* + It refers to the specific data that is relevant and necessary for a particular task or analysis being conducted using data mining techniques. This specifies the data to be mined.
  + This data may include specific attributes, variables, or characteristics that are relevant to the task at hand, such as customer data, sales data, or website usage statistics.
  + The data selected for mining is typically a subset of the overall data available, as not all data may be necessary or relevant for the task.
  + **For example**: Extracting the database name, database tables, and relevant required attributes from the dataset from the provided input database.

## Kind of knowledge to be mined:

* + It refers to the type of information that are being required through the use of data mining techniques. This describes the data mining tasks or functions to be performed.
  + It includes various tasks such as classification, clustering, discrimination, characterization, association, and evolution analysis.
  + **For example**, it determines the task to be performed on the relevant data in order to mine useful information such as classification, clustering, prediction, discrimination, outlier detection, and correlation analysis.

## Background knowledge to be used in the discovery process:

* + It refers to any prior information or understanding that is used to guide the data mining process.
  + This can include domain-specific knowledge, such as industry-specific terminology, trends as well as knowledge about the data itself.
  + The use of background knowledge can help to improve the accuracy and relevance of the insights obtained from the data mining process.
  + **For example**, Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction. Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level. Such as:
    - **Rolling Up - Generalization of data:** Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.
    - **Drilling Down - Specialization of data:** Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

## Interestingness measures and thresholds for pattern evaluation:

* + It refers to the methods and criteria used to evaluate the quality and relevance of the patterns or insights discovered through data mining.
  + **Interestingness measures** are used to quantify the degree to which a pattern is considered to be interesting or relevant based on certain criteria, such as its support or confidence.
  + These measures are used to identify patterns that are meaningful or relevant to the task.
  + **Thresholds** for pattern evaluation are used to set a minimum level of interestingness that pattern must meet in order to be considered for further analysis or action.
  + These are used to evaluate the quality and relevance of the discovered patterns. They help to identify patterns that are truly meaningful and useful. Examples include:
    - **Support:** How often a pattern occurs in the data.
    - **Confidence:** How likely it is that a pattern is true.
  + **For example**: Evaluating the interestingness measures such as utility, certainty, and novelty for the data and setting an appropriate threshold value for the pattern evaluation.

## Representation for visualizing the discovered patterns:

* + It refers to the methods used to represent the patterns discovered through data mining in a way that is easy to understand and interpret.
  + Visualization techniques such as charts, graphs, and maps are commonly used to represent the data and can help to highlight important trends, patterns, or relationships within the data.
  + Visualizing the discovered pattern helps to make the patterns obtained from the data mining process more accessible and understandable to a wider audience, including non- technical stakeholders.
  + This specifies how the discovered patterns should be presented to the user. Common visualization techniques include:
    - **Rules:** Expressing patterns in the form of "if-then" statements.
    - **Tables:** Presenting patterns in a structured format.
    - **Charts and graphs:** Visualizing patterns using different types of diagrams.
    - **Decision trees:** Representing classification models in a tree-like structure.
  + **For example:** Presentation and visualization of discovered pattern of data using various visualization techniques such as bar graphs, charts, graphs, tables, etc.

**DATA MINING TASKS with examples:**

|  |  |
| --- | --- |
| **Data Characterization:**  Summarizing general characteristics of data in a target class. | **Example:**  Finding the average salary and job roles of employees in an organization. |
| **Data Discrimination:**  Comparing two or more datasets to find differences. | **Example:**  Comparing the characteristics of high- performing and low-performing students. |
| **Association Rule Mining:**  Identifying relationships between different attributes in a dataset. | **Example:**  "Customers who buy bread often buy butter." |
| **Classification:**  Assigning data objects to predefined categories or classes. | **Example:**  Email classification into "Spam" and "Not Spam." |
| **Clustering:**  Grouping data into clusters based on similarities | **Example:**  Segmenting customers based on purchasing behaviour. |
| **Outlier Detection:**  Identifying data points that significantly differ from others. | **Example:**  Detecting fraudulent transactions in banking. |
| **Regression Analysis:**  Predicting continuous numerical values based on input data. | **Example:**  Predicting house prices based on features like location, size, and number of rooms |
| **Sequential Pattern Mining:**  Identifying patterns that occur in sequences over time. | **Example:**  Predicting stock price trends based on historical data. |
| **Deviation Analysis:**  Finding anomalies or changes in patterns over time. | **Example:**  Detecting unusual drops in sales for a product |

# DATA MINING QUERY LANGUAGE (DMQL):

* + Data Mining Query Language (DMQL) is a specialized query language used to define data mining tasks in databases. DMQL provides a structured way to extract patterns, similar to how SQL is used for querying databases.
  + It helps users specify:
    - **What data to mine**
    - **Which patterns to discover**
    - **How results should be presented**
  + Designing a comprehensive data mining language is challenging because data mining covers a wide spectrum of tasks, from data characterization to evolution analysis. Each task has different requirements.
  + The design of an effective data mining query language requires deep understanding of power, limitation, and fundamental mechanisms of various kinds of data mining tasks.
  + This facilitates a data mining system's communication with other information systems and integrates with the overall information processing environment.

**Basic Structure of DMQL:** A DMQL query consists of the following main components:

|  |  |  |
| --- | --- | --- |
| 1. | **Data to be Mined**: | Specifies the dataset. |
| 2. | **Type of Knowledge to be Discovered**: | Defines the type of pattern (e.g., association, classification, prediction, clustering) |
| 3. | **Background Knowledge:** | Includes additional rules, constraints, or hierarchies. |
| 4. | **Interestingness Measures:** | Defines criteria for meaningful patterns |
| 5. | **Presentation of Results:** | Specifies how results should be visualized |

**Basic syntax in DMQL:** DMQL acquires syntax like the relational query language, **SQL**. **Syntax -- To retrieve relevant dataset:**

1. use database (database\_name)
2. { use hierarchy (hierarchy\_name) for (attribute) }
3. (rule\_specified)
4. related to (attribute\_or\_aggreagate\_list)
5. from (relation(s)) [ where(condition) ]
6. [ order by(order\_list) ]
7. { with [ (type\_of) ] threshold = (threshold\_value) [ for(attribute(s)) ] }

## In the above syntax of Data Mining Query,

* + The **first line** retrieves the required database (database\_name).
  + The **second line** uses hierarchy, one has chosen(hierarchy\_name) with given attribute.
  + In **third line,** (**rule\_specified**) denotes the types of rules to be specified.
  + In **fourth line,** To find out the various specified rules, one must find the related set based on the attribute or aggregation which helps in generalization.
  + In **fifth line,** FROM and WHERE clauses makes sure of given condition being satisfied.
  + In **sixth line,** Then they are ordered using “order by” for a designated threshold value with respect to attributes.

**For the rule\_specified in DMQL,** The syntax is given below:

|  |  |
| --- | --- |
| **RULES** | **SYNTAX** |
| **Generalization:** | generalize data [into (relation\_name)] |
| **Association:** | find association rules [as (rule\_name)] |
| **Classification:** | find classification rules [as (rule\_name) ] according to [(attribute)] |
| **Characterization:** | find characteristic rules [as (rule\_name)] |

## Example-1 Clustering Analysis:

Clustering groups data into meaningful clusters based on similarities.

|  |  |
| --- | --- |
| **DMQL Query:**  mine clusters  from CustomerData using k-means  with k = 3; | **Explanation:**   * **mine clusters:** Specifies that clustering should be performed. * **from CustomerData:** Uses data from CustomerData table. * **using k-means:** Uses the k-means clustering algorithm. * **with k = 3:** Creates 3 clusters. |
| **Output:**  The system clusters customers into three groups:   1. Low-income customers 2. Middle-income customers 3. High-income customers | |

## Example-2: Association Rule Mining in DMQL

**Problem:** Suppose we have a retail dataset containing customer transactions, and we want to find frequent itemsets that customers often purchase together.

|  |  |
| --- | --- |
| **DMQL Query:**  use database RetailDB; mine association rules from TransactionData extracting frequent patterns  with support threshold = 30%  and confidence threshold = 70%; | **Output:**  The system will return frequent itemsets such as  {Milk, Bread} → {Butter}, Which means customers who buy Milk and Bread often buy Butter too. |

## Explanation of the Query:

|  |  |
| --- | --- |
| **use database RetailDB;** | Specifies the database that contains the transaction data. |
| **mine association rules** | Instructs the system to perform association rule mining. |
| **from TransactionData** | Specifies table (or dataset) that contains transaction data. |
| **extracting frequent patterns** | Indicates that we want to find itemsets that appear frequently. |
| **with support threshold = 30%** | Means that an itemset must appear in at least 30% of the transactions to be considered frequent. |
| **and confidence threshold = 70%;** | Specifies that the confidence level for a rule must be at least 70% (e.g., "if a customer buys bread, there is a 70% chance they also buy butter"). |

**Example-3: Concept Hierarchy Definition**

A concept hierarchy allows data to be viewed at different levels of abstraction. For example, we can define a hierarchy for location data (Country → State → City).

|  |  |
| --- | --- |
| **DMQL Query:**  define hierarchy location\_hierarchy on Customer(Location)  as (Country, State, City); | **Explanation:**   * Defines a new hierarchy named location\_hierarchy. * Specifies that the hierarchy applies to the Location attribute in the Customer table. * Defines the levels of hierarchy from the highest (Country) to the lowest (City). |

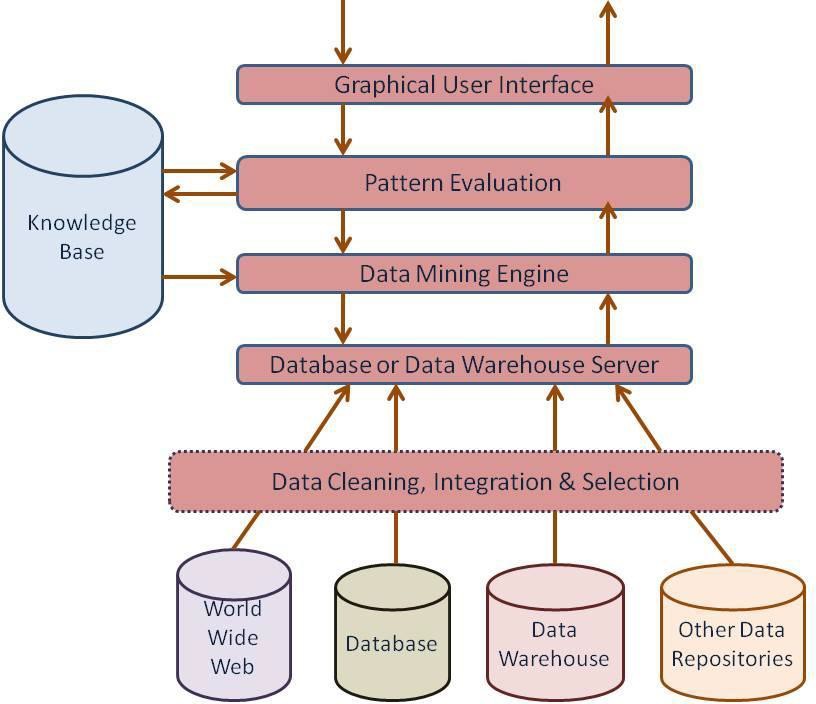
# ADVANTAGES OF DMQL:

1. **Standardized Query Language:** DMQL provides a structured and systematic approach to data mining, making it easier to define and execute mining tasks.
2. **Integration with Databases:** It can be integrated with traditional databases and data warehouses, allowing seamless data retrieval and mining.
3. **Supports Multiple Data Mining Functions:** DMQL supports classification, clustering, association rule mining, and other tasks.
4. **User-Friendly & Simplifies Complex Queries:** It provides easier way for users to specify complex mining operations compared to traditional programming or SQL-based queries.

# DISADVANTAGES OF DMQL:

1. **Complex Syntax for Beginners:** While it simplifies some tasks, DMQL can still be challenging for users without experience in query languages.
2. **Performance Issues with Large Datasets:** Executing DMQL queries on massive datasets can be slow if not optimized properly.
3. **Requires Strong Understanding of Data Mining Concepts:** Users must be familiar with data mining techniques to write effective DMQL queries.

# DATA MINING ARCHITECTURE:



## Components of Data Mining Architecture:

1. **Data Sources:**
   1. Database, World Wide Web(WWW), and data warehouse are parts of data sources.
   2. The data in these sources may be in the form of plain text, spreadsheets, or other forms of media like photos or videos.
   3. WWW is one of the biggest sources of data.

## Database Server:

* 1. The database server contains the actual data ready to be processed.
  2. It performs the task of handling data retrieval as per the request of the user.

## Data Mining Engine:

* 1. It is one of the core components of the data mining architecture that performs all kinds of data mining techniques like association, classification, clustering etc.

## Pattern Evaluation Modules:

* 1. They are responsible for finding interesting patterns in the data and sometimes they also interact with the database servers for producing the result of the user requests.

## Graphic User Interface:

* 1. Since the user cannot fully understand the complexity of the data mining process so graphical user interface helps the user to communicate effectively with the data mining system.

## Knowledge Base:

* 1. Knowledge Base is an important part of the data mining engine that is quite beneficial in guiding the search for the result patterns.
  2. Data mining engines may also sometimes get inputs from the knowledge base.
  3. This knowledge base may contain data from user experiences.
  4. The objective of the knowledge base is to make the result more accurate and reliable.

# VARIOUS ARCHITECTURES OF DATA MINING SYSTEMS:

* Data mining systems can be categorized into different architectures based on **how they interact with databases**, **how they process data**, and **their level of integration**.
* In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

## The list of Integration Schemes is as follows -

1. **No-Coupling Architecture:**
   * It is a **Standalone data mining system** that does not interact with a database or data warehouse. In this, Data is stored separately in local files. Data mining is performed independently without direct access to databases. Users must manually extract data before running mining algorithms.

|  |  |
| --- | --- |
| **Pros:**   * Simple to implement. * Useful for small datasets or experimental research | **Cons:**   * Inefficient for large datasets. * No real-time or dynamic querying support. |

## Loose Coupling Architecture:

* + In this, **Data mining system accesses databases indirectly.** In this, Data is extracted from databases or data warehouses. The mining system processes the extracted data separately. Results are stored back in the system or displayed to the user.

|  |  |
| --- | --- |
| **Pros:**   * More flexible than no-coupling architecture. * Can work with multiple data sources. | **Cons:**   * Not optimized for real-time mining. * Additional steps are required for data   extraction and storage. |

## Semi-Tight Coupling Architecture:

* + In this, **Partial integration of the data mining system with databases or data warehouses.** In this, Data mining components use database functionalities (like indexing, query optimization). Here, some pre-processing and mining tasks are performed within the database. The mining system still works as a separate module.

|  |  |
| --- | --- |
| **Pros:**   * Faster processing than loose coupling. * Takes advantage of database capabilities. | **Cons:**   * Still not fully integrated, leading to inefficiencies. |

## Tight Coupling Architecture:

* + In this, **Fully integrated with database systems or data warehouses.** In this, Data mining is performed directly inside the database system. It uses SQL-based queries or built-in mining functions. No need for separate data extraction.

|  |  |
| --- | --- |
| **Pros:**   * High efficiency and performance. * Supports real-time mining. * Better scalability for big data applications. | **Cons:**   * Complex implementation. * Requires modifications to existing   database systems. |

# CONCEPT DESCRIPTION: CHARACTERIZATION AND COMPARISIONS

* **Concept Description** refers to the process of summarizing and characterizing a set of data to extract meaningful patterns, trends, and relationships. It involves generating summarized information about a dataset.
* It a descriptive task that aims to provide summaries of data. It involves characterizing and comparing different concepts or classes within a dataset.

## Types of Concept Description:

1. **Characterization:**
   * It summarizes the general characteristics of a target class or concept. This involves identifying common features, patterns, and trends within the data. It uses statistical measures such as mean, median, mode, and standard deviation.
   * It provides a general summary of a specific set of data (e.g., customers who frequently purchase a product).
   * **For example:** "The majority of high-spending customers are aged 25-40 and prefer online shopping."

## Discrimination / Comparison:

* + It compares two or more classes/datasets to identify their differences and similarities. It helps in distinguishing one category from another based on key attributes.
  + **For example:** "Customers who purchase luxury items tend to have higher annual incomes compared to those who buy budget items."

## Techniques of Concept Description:

1. Attribute-Oriented Induction (AOI)
2. Data summarization/ Statistical Summarization
3. Rule-Based Description
4. Visualization

## Applications of Concept Description:

1. **Understanding customer behaviour:** It means customer segmentation in marketing.
2. **Analysing market trends:** Comparing product sales across different regions to identify trends and opportunities.
3. **Identifying risk factors:** Characterizing individuals with a high risk of developing a certain disease based on their medical history and lifestyle factors.

# CHARACTERIZATION:

* **Characterization** is a fundamental concept in data mining used to describe and summarize the general features of a specific dataset. It helps businesses, researchers, and decision- makers understand key properties of data and make data-driven decisions.
* By using techniques like **attribute-oriented induction (AOI)**, **statistical summarization**, **rule-based descriptions**, and **visualization** can identify trends, behaviours, and patterns efficiently.
* **For example**, in a retail business, characterization might help describe frequent customers based on their purchasing patterns and shopping behaviour.

## Steps in Characterization: (Data Characterization Process)

1. **Data Selection:** Choose relevant data for characterization.
2. **Data Pre-processing:** Clean, normalize, and prepare data for analysis.
3. **Attribute-Oriented Induction (AOI):** Generalize raw data into higher-level concepts.
4. **Data Summarization:** Find statistics like mean, median, and frequency.
5. **Pattern Extraction and Rule Generation:** Identify trends, associations, and classification rules.
6. **Visualization & Presentation:** Represent data using charts, graphs, and reports.

## Advantages of Characterization:

1. It helps in understanding key trends and patterns in the data.
2. Summarized data helps businesses and researchers to simplify decision making.
3. It helps in personalized marketing and targeted campaigns.

## Applications of Characterization:

1. **Retail and Marketing:** Identifying the characteristics of high-valued customers. **Example:** Finding most high-value customers are aged 25-40. They prefer online shopping over in-store purchases. They purchase electronics and fashion items frequently.
2. **Healthcare:** Characterize diabetic patients based on lifestyle and medical history. **Example:** Findings Most diabetic patients are over 40 years old. Many have a family history of diabetes. A majority have sitting lifestyles with high carbohydrate intake.
3. **Banking and Finance:** It characterizes individuals likely to default on loans. **Example:** Findings defaulters often have **low credit scores**. Many have **multiple high-interest loans**. A significant percentage have **unstable employment histories**.

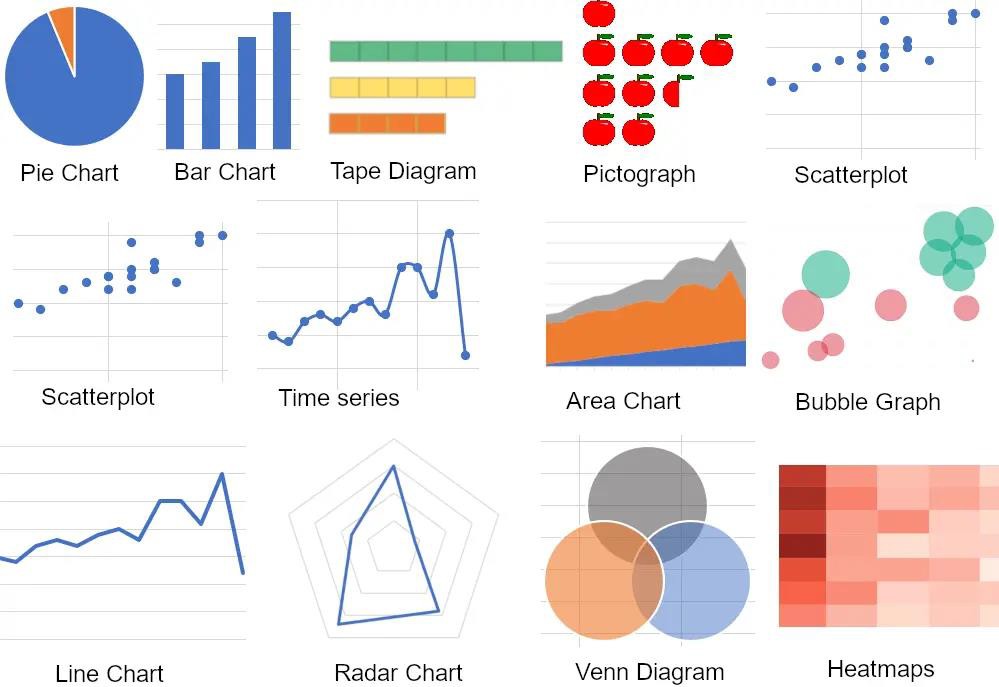
## Techniques of Characterization:

1. **Attribute-Oriented Induction (AOI):**
   * A technique that uses **data generalization** to extract characteristic rules and summarize data. It generalizes data by replacing specific values with higher-level concepts. **Example:** If dataset contains "Age = 21, 22, 23," it can be generalized to "young adult."
   * Data Generalization involves summarizing data to a higher level of abstraction. **For example**, instead of describing individual customers, we might describe customer segments based on its purchasing behaviour.

## Statistical Summarization:

* + It provides an overview of data using statistical measures, such as mean, median, and mode. Techniques like calculating summary statistics, creating histograms, and generating box plots provides a visual and quantitative overview of the data's distribution and key features. **Example:** "The average salary of employees in the IT department is $80,000, with a standard deviation of $10,000."
    - **Mean** (Average): The central value of a numerical attribute.
    - **Median**: The middle value when data is sorted.
    - **Mode**: The most frequently occurring value.
    - **Standard Deviation**: Measures data spread or variability.

1. **Rule-Based Description:** It uses association rules or classification rules. **For example: If** Age > 30 **and** Income > $50,000 **then** Likely to Buy a Luxury Car. This rule helps businesses target high-income individuals for luxury car promotions.
2. **Visualization Techniques:** Charts, graphs, and other visual representations can effectively communicate the characteristics of a class, making it easier to understand the dataset. They are used to display summarized data. **Example:** A bar chart showing the percentage of customers from different age groups in a store.



# COMPARISON / DISCRIMINATION:

* + Comparison is also called **Discrimination**. It is the process of analysing and contrasting two or more groups of data to identify **differences, distinguishing features, and patterns**.
  + It helps businesses, researchers, and analysts to find key differences in customer behaviour, product performance, or market trends. It supports decision-making by analysing different data segments.
  + Using statistical techniques, rule-based methods, and machine learning models, organizations can make data-driven decisions, optimize marketing strategies, detect fraud, and improve healthcare outcomes.
  + **For example**: In marketing, comparison can help differentiate **high-spending** and **low- spending** customers based on their purchasing behaviour.

## Steps in the Comparison Process:

1. **Select Data Groups:** Define the two or more datasets to compare.
2. **Pre-process Data:** Clean, filter, and transform data to ensure accuracy.
3. **Choose Comparison Method:** Statistical, machine learning, rule-based, or visualization techniques.
4. **Analyse Differences:** Identify key distinguishing features and trends.
5. **Present Findings:** Use tables, charts, graphs, or rules to explain the differences.

## Applications of Comparison:

1. **Customer Segmentation in Marketing:** Compare high-value and low-value customers. **Findings:** High-value customers shop frequently and prefer premium brands. Low-value customers shop occasionally and buy discount products.
2. **Fraud Detection in Banking:** Compare fraud and valid transactions.

**Findings:** Fraud transactions are often international, high-value, and made at unusual times. Valid transactions match customer spending history and location.

1. **Medical Diagnosis:** Compare patients with and without a disease.

**Findings:** Diabetic patients have higher glucose levels and sitting lifestyles. Non-diabetic patients have normal glucose levels and exercise regularly.

1. **Stock Market Analysis:** Compare performing and non-performing stocks.

**Findings:** Performing stocks have high trading volume and positive earnings reports. Non- performing stocks show low demand and declining revenue.

## Techniques for Comparison:

1. **Attribute-Oriented Induction (AOI):** Similar to AOI used in characterization, but instead of summarizing, it finds key differences. Attributes are generalized and compared between two datasets. **Example:**

|  |  |
| --- | --- |
| **Group-1:** High-income customers | **Group-2:** Low-income customers |
| **Finding:** High-income customers frequently buy luxury items, whereas low-income customers prefer budget-friendly products. | |

1. **Statistical Comparison:** Uses statistical measures to compare datasets. Such as, mean, median, mode to find the central tendency of different groups. **Example:**

|  |  |
| --- | --- |
| **Group-1 (Online Shoppers):** Average purchase = $150 | **Group-2 (In-store Shoppers):** Average purchase = $100 |
| **Finding:** Online shoppers spend more per transaction than in-store shoppers. | |

1. **Discriminant Analysis:** A classification technique that finds attributes that best differentiate between two groups. **Example:**

|  |  |
| --- | --- |
| **Group-1:** Loan Defaulters  Defaulters have low credit scores, high and unstable employment | **Group-2:** Non-Defaulters  Non-defaulters have high credit scores and stable jobs. |

1. **Rule-Based Comparison:** Uses **association rules** and **decision trees** to extract rules differentiating datasets. **Example:**

|  |  |
| --- | --- |
| **Rule 1:**  If the customer is under 30 **and** has an active social media presence, they are likely to prefer digital banking. | **Rule 2:**  If the customer is over 50, they are more likely to visit physical bank branches. |

1. **Machine Learning-Based Comparison:** Uses models like Decision Trees, Random Forest, and Neural Networks to find key distinguishing factors. **Example:**

|  |  |
| --- | --- |
| **Group-1:** Spam Emails  Spam emails often contain words like "FREE," "WINNER," and "CASH PRIZE." | **Group-2:** Non-Spam Emails  Non-spam emails have more personalized greetings and business language |

## Advantages of Comparison:

* 1. It helps in business strategy and decision-making.
  2. It distinguishes different consumer groups for personalized marketing.
  3. It finds anomalies that indicate fraud activity.

**Difference between Data Characterization and Data Comparison:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Characterization** | **Data Comparison** |
| **Definition** | Summarizes and describes the general characteristics of a dataset. | Compares two or more datasets to identify differences. |
| **Purpose** | Provides an overview of the key features of a dataset. | Highlights differences between groups or datasets. |
| **Techniques Used** | * Attribute-Oriented Induction (AOI) * Statistical Summarization * Visualization | * Statistical Analysis * Discriminant Analysis * Rule-Based Comparison |
| **Example** | "The average income of customers in this dataset is $70,000, and most purchases come from people aged 25- 40." | "Customers with an income above  $70,000 buy luxury items, while those with lower income prefer budget-friendly products." |
| **Output Type** | Generalized patterns, summaries, and trends. | Differences, distinguishing features, and contrasting patterns. |
| **Use Cases** | * Understanding customer behaviour * Summarizing medical records * Identifying key attributes in a dataset | * Customer segmentation (e.g., high vs. low spenders) * Fraud detection (e.g., fraud vs. valid transactions) * Disease classification (e.g., diabetic vs. non-diabetic patients) |

# DATA GENERALIZATION:

* **Data Generalization** is the process of summarizing data by replacing relatively low level values with higher level concepts. It is a form of descriptive data mining.
* It simplifies large datasets by summarizing data and identifies trends that may not be visible in raw data. It reduces data storage needs by eliminating unnecessary details. It enhances privacy by masking specific details in sensitive data.

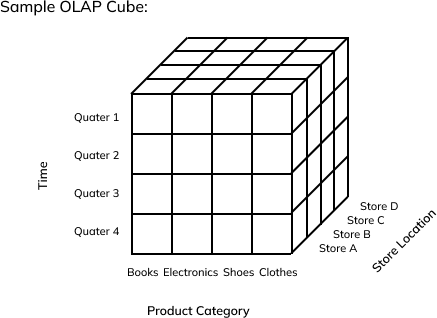
**For example:** Instead of listing every single value, data is grouped into **ranges or categories**.

|  |  |
| --- | --- |
| **Original Data** | **Generalized Data** |
| Age: 22 | Age: 20-30 |
| Age: 27 | Age: 20-30 |
| City: New York | City: USA |
| Salary: $55,000 | Salary: $50K - $60K |

**There are two basic approaches of data generalization:** OLAP approach & AOI approach

## Data cube approach (OLAP - Online Analytical Processing approach):

* + It is an efficient approach as it is helpful to make the past selling graph.
  + In this approach, computation and results are stored in the Data cube.
  + It uses Roll-up and Drill-down operations on a data cube.
  + These operations typically involve aggregate functions, such as count(), sum(), average(), and max().
  + These materialized views can then be used for decision support, knowledge discovery, and many other applications.



## Attribute Oriented Induction (AOI):

* + It is an online data analysis, query oriented and generalization based approach.
  + In this approach, we perform generalization on basis of different values of each attributes within the relevant data set. After that same tuple are merged and their respective counts are accumulated in order to perform aggregation.

## Attribute oriented induction approach uses two method :

* + 1. Attribute removal.
    2. Attribute generalization.

## Attribute Removal:

* + **Attribute removal** is a technique used in data generalization where specific attributes in a dataset are eliminated to simplify the data. This helps in reducing complexity, improving privacy, and highlighting only the most relevant features.
  + It involves removing less important or sensitive attributes from a dataset to make analysis more efficient.

|  |  |
| --- | --- |
| **Why Use Attribute Removal?**   * Few attributes make data easier to analyze. * Sensitive data is removed, protecting user identities. * Less data to process means faster computations and improves performance. | **When to Use Attribute Removal?**   * **When certain attributes are irrelevant** (**e.g.,** customer names in sales trend analysis). * **When privacy concerns exist** (**e.g.,** removing personal information like A/c no.). * **When data redundancy is high** (**e.g.,** removing one column if another provides similar information). |

**Example:** In Customer Data, removing name and Phone number. We have removed because:

* + **Privacy** → Phone numbers & names are sensitive.
  + **Relevance** → Name doesn’t affect salary analysis.
  + **Efficiency** → Fewer columns mean faster processing.

**Before Attribute removal:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Age** | **City** | **Salary** | **Phone Number** |
| John | 28 | New York | $50K | 123-456-7890 |
| Sarah | 35 | Chicago | $65K | 987-654-3210 |

**After Attribute Removal (Removing Name & Phone Number):**

|  |  |  |
| --- | --- | --- |
| **Age** | **City** | **Salary** |
| 28 | New York | $50K |
| 35 | Chicago | $65K |

1. **Attribute Generalization:**
   * **Attribute Generalization** is a technique in data generalization where specific attribute values are replaced with higher-level, more abstract concepts using concept hierarchies. This helps in summarizing data and identifying patterns efficiently.
   * It involves replacing detailed values with generalized categories based on pre-defined hierarchies. It simplifies large datasets by reducing complexity. It uses less space by summarizing attributes.

**Example:** Instead of dealing with exact ages, we group them into broader categories.

|  |  |  |
| --- | --- | --- |
| **Original Age** | **Generalized Age (Level 1)** | **Generalized Age (Level 2)** |
| 23 | 20-30 | Young |
| 37 | 30-40 | Middle-aged |
| 65 | 60-70 | Senior |

## Techniques of Attribute Generalization:

1. **Concept Hierarchy-Based Generalization:** Data is grouped into hierarchical levels for abstraction. The higher levels provide more generalization.

## Example: Location Generalization

|  |  |  |
| --- | --- | --- |
| **City** | **Generalized Level 1** | **Generalized Level 2** |
| New York | New York (State) | USA |
| Los Angeles | California | USA |
| Paris | Île-de-France | France |

1. **Range-Based Generalization:** Numerical data is grouped into ranges instead of individual values. This is useful when analysing salary distributions rather than focusing on individual salaries.

**Example: Income Generalization**

|  |  |
| --- | --- |
| **Exact Salary** | **Generalized Salary** |
| $52,000 | $50K - $60K |
| $98,000 | $90K - $100K |

# DATA SUMMARIZATION:

* **Data summarization** is a key technique in data mining that helps in extracting useful information from large datasets by providing a compact representation of the data.
* It enables to understand large datasets quickly, reduces storage requirements, and helps in decision-making. It involves computing statistical measures, aggregating data, and creating visual or textual summaries that highlight key patterns, trends, and distributions.

## Techniques of Data Summarization:

1. **Statistical Measures:** Summarization begins with computing basic statistical measures:
   * **Mean (Average):** The central value of a dataset.
   * **Median:** The middle value when data is sorted.
   * **Mode:** The most frequently occurring value.
   * **Variance & Standard Deviation:** Measures of data dispersion.
2. **Aggregation:** Combining multiple data points into a single value to provide a higher-level summary. For example:
   * Summing up monthly sales data to get yearly totals.
   * Averaging customer ratings for a product.

## Clustering and Segmentation:

* + **Clustering:** Grouping similar data points together to form patterns.
  + **Segmentation:** Dividing data into meaningful groups based on predefined attributes.

## Visualization Techniques:

* + **Histograms & Bar Charts:** Show frequency distributions.
  + **Pie Charts:** Represent proportions of categories.
  + **Box Plots:** Indicate data spread and outliers.
  + **Heat-maps:** Display correlations between variables.

## Data Cube (OLAP Summarization):

* + Online Analytical Processing (OLAP) cubes store pre-aggregated data for fast retrieval.
  + Used for multidimensional data analysis (e.g., sales by region, time, and product).

## Applications of Data Summarization:

1. **Business:** Understanding customer behaviour, sales trends, and market performance.
2. **Healthcare:** Summarizing patient records for better diagnosis.
3. **Finance:** Summarizing stock market data for investment decisions.
4. **Social Media:** Aggregating trends from large-scale text data.

# MINING CLASS COMPARISIONS:

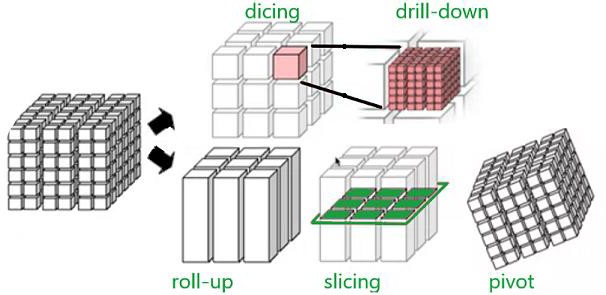
* Mining class comparisons is a technique used to analyse and compare different classes or groups of data based on their attributes. The goal is to identify discriminative features that distinguish one class from another.
* Class discrimination or class comparison mines descriptions that distinguish a target class from its contrasting classes. The target and contrasting classes must be comparable because they share similar dimensions and attributes.
* **For example,** the three classes - person, address, and item are not comparable. But the sales in the last three years are comparable classes. Also, we can compare computer science candidates versus physics candidates.

**The general procedure for class comparison is as follows:**

1. **Data Collection:** The set of relevant data in the database and data warehouse is collected by query Processing and partitioned into a target class and one or a set of contrasting classes.
2. **Dimension relevance analysis:** If there are many dimensions and analytical comparisons are desired, then dimension relevance analysis should be performed on these classes and only highly relevant dimensions are included in the further analysis.
3. **Synchronous Generalization:** The process of generalization is performed upon the target class to the level controlled by the user, which results in a prime target class relation. The concepts in the contrasting class are generalized to the same level as those in the prime target class relation, forming the prime contrasting class relation or cuboid.
4. **Drilling Down, Rolling Up and other OLAP adjustments:** Synchronous and asynchronous drill-down, roll-up & other OLAP operations such as, slicing, dicing and pivoting can be performed on target and contrasting classes based on user instructions.
5. **Presentation of the derived comparison:** The resulting class comparison description can be visualized in the form of tables, charts, and rules. This presentation usually includes a "contrasting" measure (such as count %) that reflects the comparison between the target and contrasting classes.

The OLAP operations are:

1. **Roll-up**: Aggregates data to a higher level of hierarchy.
2. **Drill-down**: Breaks down data to a more detailed level.
3. **Slice**: Filters data along a single dimension.
4. **Dice**: Filters data based on multiple dimensions.
5. **Pivot (Rotate)**: Re-orients the data view for better analysis.



## Example of Mining Class Comparison:

Suppose we would like to compare the general properties between the old customer and new customer of Royal Electronics which deals in computer’s electronic products, given the attributes name, gender, product, birthplace, birthdate, residence and phone.

**Target class: old customer**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Gender** | **Product** | **Birthplace** | **Birthdate** | **Residence** | **Phone** |
| Rakesh | M | Printer | Jodhpur | 14/07/1993 | 3/A C.H.B. | 9925852890 |
| Sumit | M | Scanner | Jaipur | 22/06/2002 | A.G. colony | 9875894102 |
| Minal | F | Keyboard | Jodhpur | 11/08/1992 | Bank colony | 9928509928 |
| …… | …… | …… | …… | …… | …… | …… |

**Contrasting class: new customer**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Gender** | **Product** | **Birthplace** | **Birthdate** | **Residence** | **Phone** |
| Harish | M | Mouse | Jodhpur | 14/07/1993 | 11/B-road | 9941852791 |
| Priya | F | Monitor | Jaipur | 22/06/2002 | Jain colony | 9460768333 |
| …… | …… | …… | …… | …… | …… | …… |

**Initial working relations: the target class vs. the contrasting class.**

1. **Data Collection:** In this, we select two set of task relevant data. One for the initial target class working relation and the other for the initial contrasting class working relation.
2. **Dimension relevance analysis:** Now, this analysis is performed on the two classes of data. After this analysis, irrelevant or weakly relevant dimensions such as name, gender, product, residence and phone are removed from the resulting class. Only highly relevant attributes are included in the subsequent analysis.
3. **Synchronous Generalization:** Here, the generalization is performed on the target class to the levels controlled by user, forming the prime target class relation. The contrasting class is also generalized to the same levels as those in the prime target class, forming the prime contrasting class relation.

**Prime generalized relation for the**

**target class:** old customer

**Prime generalized relation for the**

**contrasting class:** new customer

|  |  |
| --- | --- |
| **Birthplace** | **Age-Range** |
| Jodhpur | 25-30 |
| Jaipur | 20-25 |
| Others | Over 30 |

|  |  |
| --- | --- |
| **Birthplace** | **Age-Range** |
| Jodhpur | 18-25 |
| Jodhpur | 18-25 |
| Others | Over 30 |

1. **Drilling Down, Rolling Up and other OLAP adjustments:** The OLAP operations are performed on the target and contrasting class, based on the user’s instruction to adjust the level of abstraction.
2. **Presentation of the derived comparison:** Finally, the resulting class comparison is presented in form of tables, graphs or rules.