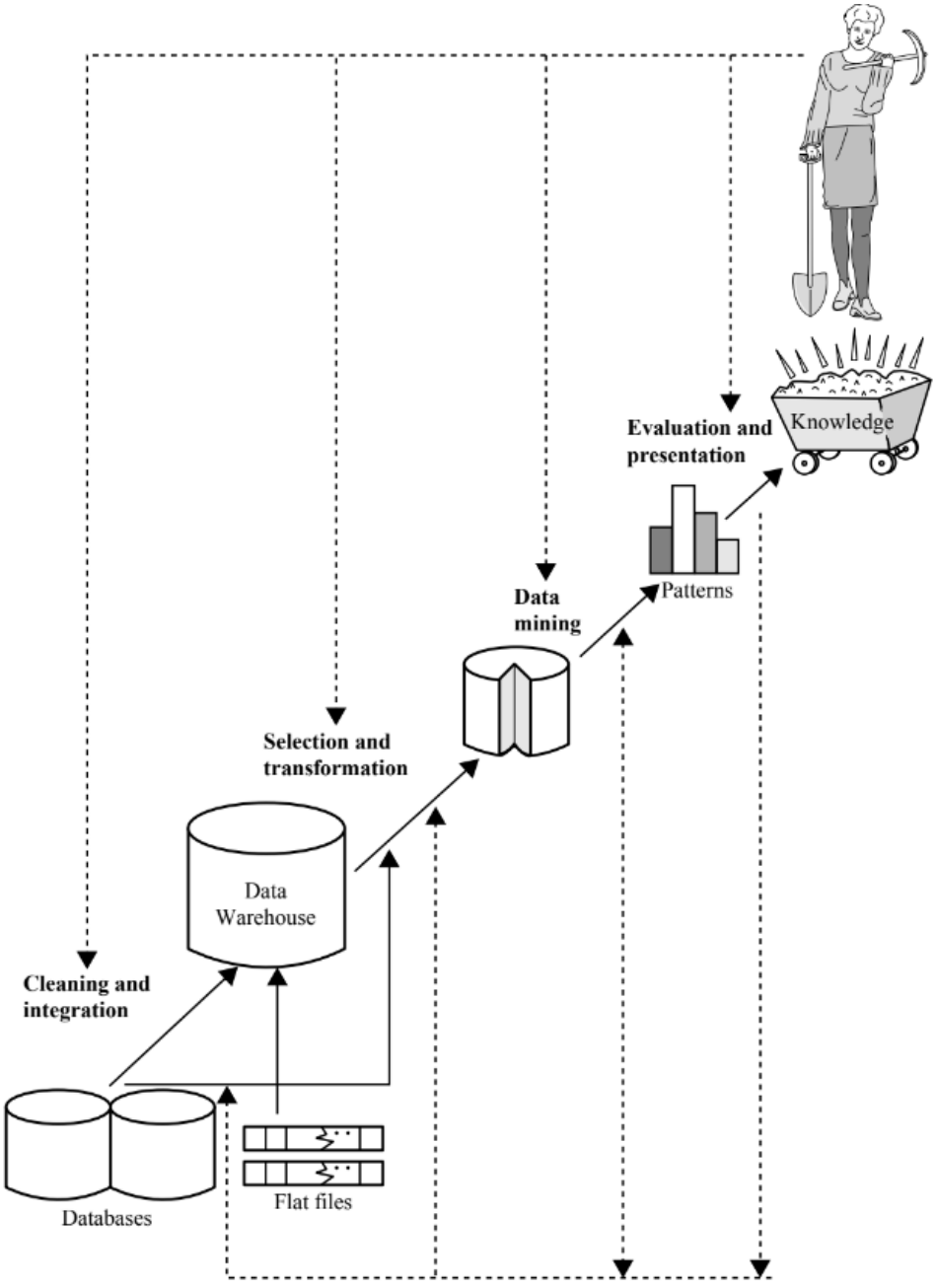


Figure 1.4 Data mining as a step in the process of knowledge discovery.

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)[3]

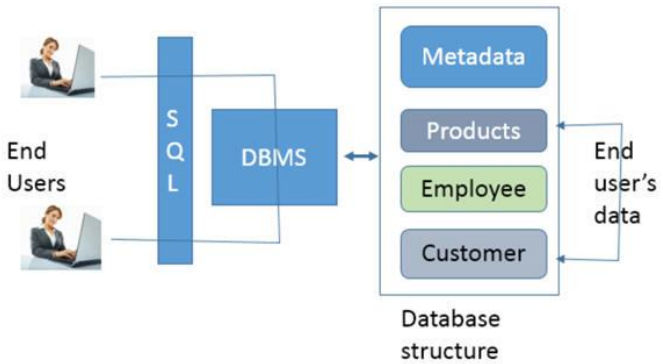
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)[4]
5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on *interestingness measures*—see Section 1.4.6)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

What Kinds of Data Can Be Mined?

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data
and transactional data.
Data mining can also be applied to other forms of data (e.g.,
data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).

Database Data

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

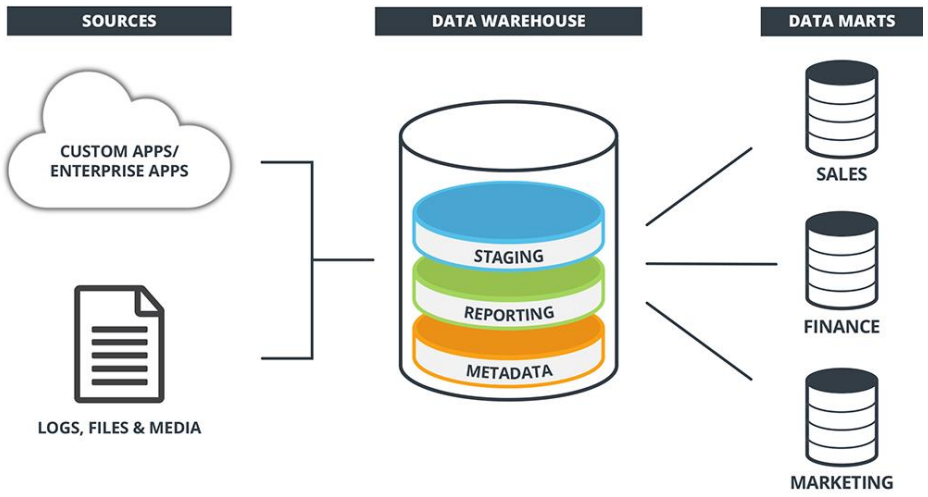


Data Warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). The data are stored to pro□vide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.

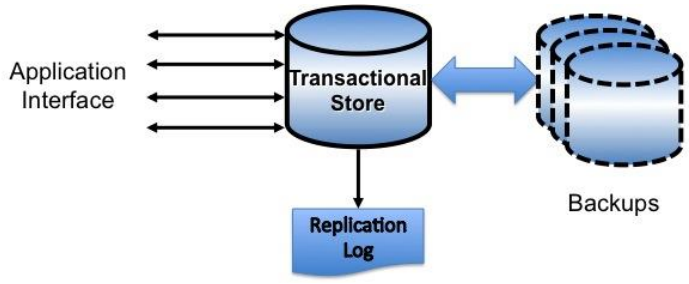
Involves ETL (Extract, Transform, Load) processes to gather, clean, and load data.
Often includes data from various operational databases to provide a centralized, consistent view.

A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count.



Transactional Data

In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page. A transaction typ ically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction. A trans actional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.



NEED OF DATA MINING

1. **Knowledge Discovery:**
    - **Objective:** Uncover hidden patterns and knowledge.
    - **Significance:** Reveals insights not apparent through traditional analysis.
2. **Business Intelligence:**
    - **Objective:** Extract actionable insights for decision-making.
    - **Significance:** Identifies market trends and improves strategic planning.
3. **Predictive Analytics:**
    - **Objective:** Forecast future trends based on historical data.
    - **Significance:** Enables proactive decision-making and resource optimization.
4. **Risk Management:**
    - **Objective:** Identify and mitigate potential risks.
    - **Significance:** Assesses and manages risks in various domains.
5. **Fraud Detection:**
    - **Objective:** Detect unusual patterns indicating fraud.
    - **Significance:** Essential for financial transactions and healthcare.
6. **Customer Relationship Management (CRM):**
    - **Objective:** Understand customer behavior and preferences.
    - **Significance:** Improves customer satisfaction and loyalty.
7. **Healthcare and Medical Research:**
    - **Objective:** Analyze patient data for insights.
    - **Significance:** Supports personalized medicine and disease diagnosis.
8. **Text and Sentiment Analysis:**
    - **Objective:** Analyze patterns in textual data.
    - **Significance:** Useful for social media monitoring and sentiment analysis.
9. **Supply Chain Optimization:**
    - **Objective:** Enhance efficiency and reduce costs.
    - **Significance:** Improves demand forecasting and logistics.
10. **Scientific Research:**
    - **Objective:** Analyze large datasets in scientific experiments.
    - **Significance:** Supports discoveries in fields like astronomy and genomics.

Technologies used for data mining

1. **Machine Learning Algorithms:**
   - **Purpose:** Enable computers to learn patterns from data and make predictions.
   - **Examples:** Decision trees, clustering algorithms, neural networks, and support vector machines.
2. **Data Warehousing:**
   - **Purpose:** Store and manage large volumes of structured data for analysis.
   - **Examples:** Data warehousing solutions like Amazon Redshift, Google BigQuery, and Microsoft Azure Synapse Analytics.
3. **Database Systems:**
   - **Purpose:** Efficiently store, retrieve, and manage structured data.
   - **Examples:** Relational Database Management Systems (RDBMS) like MySQL, PostgreSQL, and Oracle.
4. **Big Data Technologies:**
   - **Purpose:** Handle and process massive volumes of data.
   - **Examples:** Apache Hadoop, Apache Spark, and Apache Flink.
5. **Data Mining Tools:**
   - **Purpose:** Provide user-friendly interfaces for designing and executing data mining processes.
   - **Examples:** RapidMiner, KNIME, and Weka.
6. **Text Mining Tools:**
   - **Purpose:** Analyze and extract information from unstructured text data.
   - **Examples:** Natural Language Processing (NLP) libraries like NLTK (Natural Language Toolkit) and spaCy.
7. **Visualization Tools:**
   - **Purpose:** Present and interpret data patterns visually.
   - **Examples:** Tableau, Power BI, and D3.js.
8. **Cloud Computing Services:**
   - **Purpose:** Provide scalable and on-demand computing resources for data processing.
   - **Examples:** Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform.
9. **Data Preprocessing Tools:**
   - **Purpose:** Clean, transform, and prepare data for analysis.
   - **Examples:** Trifacta, DataWrangler, and OpenRefine.
10. **Statistical Analysis Tools:**
    - **Purpose:** Perform statistical analyses on data.
    - **Examples:** R, Python with libraries like NumPy and SciPy, and SAS.
11. **Data Integration Tools:**

- **Purpose:** Combine data from multiple sources for analysis.
- **Examples:** Informatica, Talend, and Microsoft SQL Server Integration Services (SSIS).

12. **Parallel Processing Frameworks:**
    - **Purpose:** Speed up data processing by performing tasks concurrently.
    - **Examples:** Apache Flink, Apache Storm, and Apache Beam.
13. **Data Mining APIs:**
    - **Purpose:** Provide programmatic access to data mining functionalities.
    - **Examples:** scikit-learn (for Python), Apache Mahout.
14. **Data Governance and Security Tools:**
    - **Purpose:** Ensure data quality, security, and compliance.
    - **Examples:** Collibra, IBM InfoSphere, and Varonis.

DATA MINING ISSUES

Mining Methodology:

1. **Diverse Knowledge Extraction:**
    - *Issue:* Data mining tasks vary widely.
    - *Challenge:* Ongoing emergence of new tasks requires continuous development of diverse data mining techniques.
2. **Multidimensional Space Exploration:**
    - *Issue:* Searching for patterns in multidimensional space.
    - *Challenge:* Enhancing data mining power through exploratory multidimensional data mining.
3. **Interdisciplinary Collaboration:**
    - *Issue:* Integrating methods from different disciplines.
    - *Challenge:* Incorporating knowledge from fields like information retrieval and natural language processing into data mining.
4. **Networked Environment Exploration:**
    - *Issue:* Objects in linked environments.
    - *Challenge:* Utilizing semantic links for boosting knowledge discovery in interconnected data objects.
5. **Handling Data Uncertainty:**
    - *Issue:* Presence of noise, errors, or uncertainty.
    - *Challenge:* Integrating techniques like data cleaning, preprocessing, and uncertainty reasoning with data mining.
6. **Pattern Evaluation and User Guidance:**
    - *Issue:* Not all patterns are interesting.

- *Challenge:* Developing techniques for assessing interestingness based on user-defined measures or constraints to guide the discovery process.

User Interaction:

1. **Interactive Mining:**
   - *Issue:* Need for a highly interactive process.
   - *Challenge:* Developing flexible user interfaces for dynamic interaction, exploration, and refinement of mining results.
2. **Incorporation of Background Knowledge:**
   - *Issue:* Utilizing domain-specific background knowledge.
   - *Challenge:* Incorporating knowledge, constraints, and rules to enhance pattern evaluation and guide the discovery process.
3. **Ad hoc Data Mining:**
   - *Issue:* Enabling users to pose ad hoc mining tasks.
   - *Challenge:* Developing high-level data mining query languages or user interfaces for flexible specification of mining tasks.
4. **Presentation and Visualization:**
   - *Issue:* Effective presentation of mining results.
   - *Challenge:* Adopting expressive knowledge representations, user-friendly interfaces, and visualization techniques.

Efficiency and Scalability:

1. **Efficient Mining Algorithms:**
   - *Issue:* Algorithms must be efficient and scalable.
   - *Challenge:* Developing algorithms with predictable running times, short processing durations, and real-time execution capabilities.
2. **Parallel and Distributed Mining:**
   - *Issue:* Huge datasets and distributed data.
   - *Challenge:* Developing parallel and distributed mining algorithms for efficient processing, utilizing techniques like cloud and cluster computing.
3. **Incremental Mining:**
   - *Issue:* Mining updates without starting from scratch.
   - *Challenge:* Incorporating incremental data mining methods that modify knowledge based on new data updates.

OTHER BASIC ISSUES

1. **Data Quality:**
   - Issue: Incomplete, noisy, or inconsistent data.
2. **Data Privacy and Security:**
   - Issue: Concerns about unauthorized access and misuse of sensitive information.
3. **Ethical Concerns:**
   - Issue: Potential misuse of data for unethical purposes.
4. **Lack of Domain Knowledge:**
   - Issue: Data miners may lack domain-specific expertise.
5. **Scalability:**
   - Issue: Handling large volumes of data efficiently.
6. **Complexity of Algorithms:**
   - Issue: Some algorithms are complex and hard to understand.
7. **Data Diversity:**
   - Issue: Diverse data sources and formats.
8. **Interpretability:**
   - Issue: Black-box nature of some advanced models.
9. **Bias and Fairness:**
   - Issue: Biased results due to imbalanced datasets.
10. **Dynamic Nature of Data:**
    - Issue: Data changes over time, affecting model relevance.
11. **Data Ownership and Access:**
    - Issue: Challenges in accessing and sharing proprietary data.
12. **Integration with Business Processes:**
    - Issue: Difficulty in integrating mining results into existing processes.
13. **Cost and Resource Constraints:**
    - Issue: Data mining can be resource-intensive.
14. **Overfitting and Model Generalization:**
    - Issue: Models may perform poorly on new data.

APPLICATIONS OF DATA MINING

1. **Business and Marketing:**
    - **Objective:** Understand customer behavior, preferences, and market trends.
    - **Applications:** Customer segmentation, targeted marketing, demand forecasting, and churn prediction.
2. **Healthcare:**
    - **Objective:** Enhance patient care, disease diagnosis, and medical research.
    - **Applications:** Predictive modeling for disease outbreaks, personalized medicine, fraud detection in healthcare claims.
3. **Finance and Banking:**
    - **Objective:** Improve risk management, fraud detection, and customer satisfaction.
    - **Applications:** Credit scoring, fraud detection, anomaly detection in transactions, customer relationship management.
4. **Retail and E-Commerce:**
    - **Objective:** Optimize inventory, pricing, and customer experience.
    - **Applications:** Recommender systems, market basket analysis, dynamic pricing, inventory management.
5. **Telecommunications:**
    - **Objective:** Enhance network performance, reduce churn, and improve customer satisfaction.
    - **Applications:** Network optimization, customer churn prediction, fraud detection.
6. **Manufacturing and Supply Chain:**
    - **Objective:** Optimize production processes, reduce costs, and improve efficiency.
    - **Applications:** Predictive maintenance, quality control, supply chain optimization, demand forecasting.
7. **Education:**
    - **Objective:** Improve student performance, educational planning, and resource allocation.
    - **Applications:** Learning analytics, student retention prediction, personalized learning.
8. **Fraud Detection and Cybersecurity:**
    - **Objective:** Identify and prevent fraudulent activities, enhance cybersecurity.
    - **Applications:** Credit card fraud detection, network intrusion detection, anomaly detection.

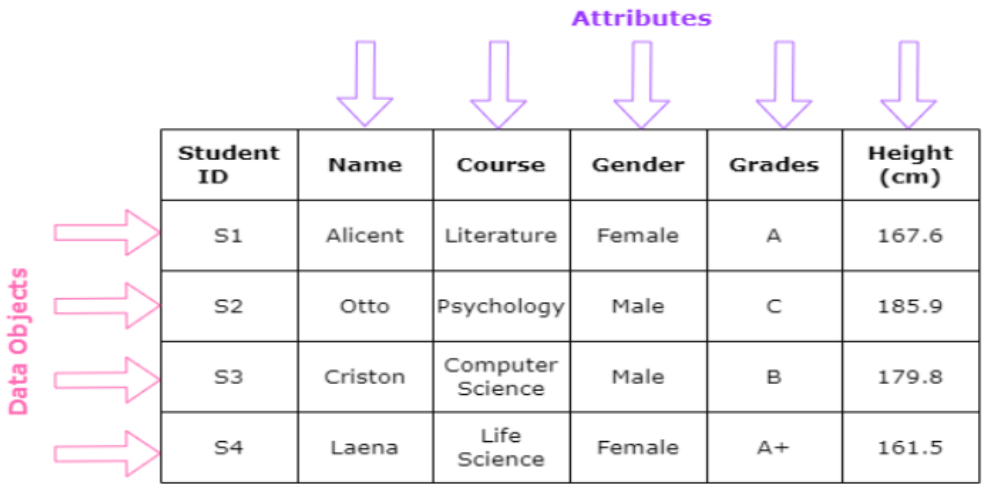
9. **Government and Public Sector:**
    - **Objective:** Enhance decision-making, public safety, and resource allocation.
    - **Applications:** Crime pattern analysis, traffic management, public health monitoring.
10. **Human Resources:**
    - **Objective:** Optimize recruitment, employee retention, and workforce planning.
    - **Applications:** Employee performance analysis, talent acquisition, attrition prediction.
11. **Energy and Utilities:**
    - **Objective:** Improve energy efficiency, predictive maintenance, and resource optimization.
    - **Applications:** Predictive maintenance of equipment, energy consumption forecasting.
12. **Environmental Science:**
    - **Objective:** Analyze environmental data for research and conservation efforts.
    - **Applications:** Climate modeling, biodiversity monitoring, pollution control.
13. **Social Media and Web Analytics:**
    - **Objective:** Understand user behavior, sentiment analysis, and content recommendation.
    - **Applications:** Social network analysis, sentiment analysis, personalized content recommendations.
14. **Sports Analytics:**
    - **Objective:** Enhance team performance, player analysis, and fan engagement.
    - **Applications:** Player performance analysis, injury prediction, fan engagement strategies.

Data Objects and Attribute Types

Data sets are made up of data objects. A data object represents an entity—in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a university database, the objects may be students, professors, and courses. Data objects are typically described by attributes. Data objects can also be referred to as samples, examples, instances, data points, or objects. If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

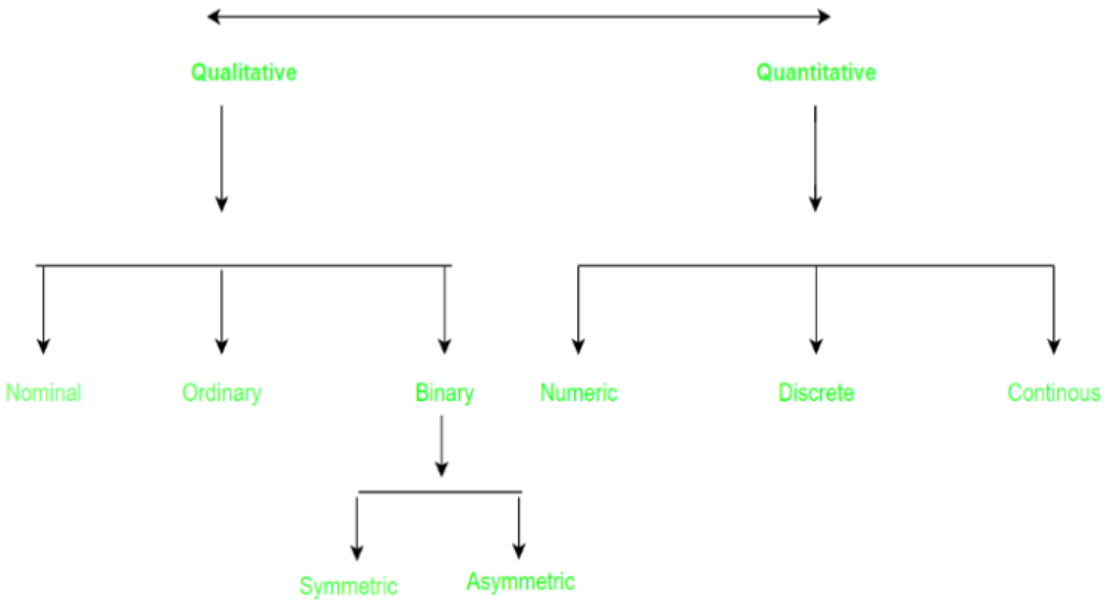
What Is an Attribute?

An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature. The term dimension is commonly used in data warehousing. Machine learning literature tends to use the term feature, while statisticians prefer the term variable. Data mining and database professionals commonly use the term attribute, and we do here as well. Attributes describing a customer object can include, for example, customer ID, name, and address. Observed values for a given attribute are known as observations. A set of attributes used to describe a given object is called an attribute vector (or feature vec□tor). The distribution of data involving one attribute (or variable) is called univariate. A bivariate distribution involves two attributes, and so on.



Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

TYPES OF ATTRIBUTES



Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values of the nominal attribute.

Example :

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

2. Binary Attributes: Binary data has only 2 values/states. For Example, yes or no, affected or unaffected, true or false.

- **Symmetric:** Both values are equally important (Gender).

Attribute	Values
Gender	Male , Female

- **Asymmetric:** Both values are not equally important (Result)

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

3. Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.

- An **interval-scaled** attribute has values, whose differences are interpretable. Data can be added and subtracted at an interval scale but cannot be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day, we cannot say that one day is twice as hot as another day.
- A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

2. **Discrete :** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.

-

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

- **Example :**

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

	Discrete Nominal	Discrete Nominal	Binary Discrete Nominal	Ordinal Discrete Nominal	Continuous Ratio-scaled
Student ID	**Name**	**Course**	**Gender**	**Grades**	**Height (cm)**
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Basic Statistical Descriptions of Data

For data preprocessing to be successful, it is essential to have an overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers. This section discusses three areas of basic statistical descriptions. We start with mea□sures of **central tendency** which measure the location of the middle or center of a data distribution. Intuitively speaking, given an attribute, where do most of its values fall? In particular, we discuss the **mean, median, mode, and midrange**. In addition to assessing the central tendency of our data set, we also would like to have an idea of the **dispersion of the data**. That is, how are the data spread out? The most common data dispersion measures are the **range, quartiles, and interquartile range; the five-number summary and boxplots; and the variance and standard deviation of the data** Finally, we can use many **graphic displays of basic statistical descriptions to visually inspect our data** . Basic Statistical Descriptions of Data packages **include bar charts, pie charts, and line graphs**. Other popular displays of data summaries and distributions include quantile plots, quantile–quantile plots, histograms, and scatter plots.

The most common and effective numeric measure of the "center" of a set of data is the *(arithmetic) mean*. Let $x_1, x_2, \ldots, x_N$ be a set of N values or *observations*, such as for some numeric attribute X, like *salary*. The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{2.1}$$

This corresponds to the built-in aggregate function, *average* (**avg()** in SQL), provided in relational database systems.

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$
$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width, \tag{2.3}$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are

lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of the median interval.

The *mode* is another measure of central tendency. The **mode** for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**. At the other extreme, if each data value occurs only once, then there is no mode.

Mode. The data from Example 2.6 are bimodal. The two modes are \$52,000 and \$70,000. ■

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation:

$$mean - mode \approx 3 \times (mean - median). \tag{2.4}$$

This implies that the mode for unimodal frequency curves that are moderately skewed can easily be approximated if the mean and median values are known.

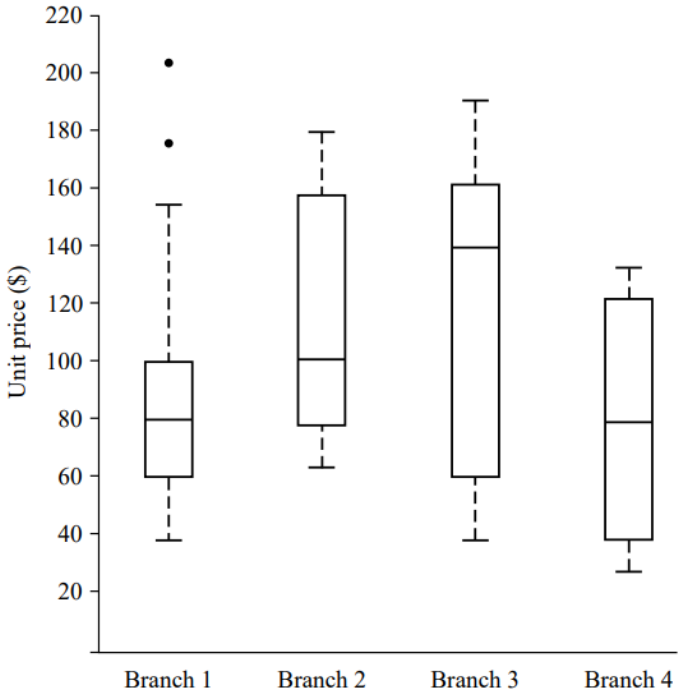
The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set. This measure is easy to compute using the SQL aggregate functions, max() and min().

Midrange. The midrange of the data of Example 2.6 is $\frac{30,000+110,000}{2} = \$70,000$. ■

1. **Range:**
   - **Definition:** The difference between the maximum and minimum values in a dataset.
   - **Calculation:** Range = Maximum Value - Minimum Value.
   - **Significance:** Provides a simple measure of the overall spread of data but is sensitive to extreme values.
2. **Quartiles:**
   - **Definition:** Divides a dataset into four equal parts, with each part representing 25% of the data.
   - **Calculation:** Q1 (First Quartile) - 25th percentile, Q2 (Second Quartile or Median) - 50th percentile, Q3 (Third Quartile) - 75th percentile.
   - **Significance:** Helps identify the central tendency and spread of the middle 50% of the data.
3. **Interquartile Range (IQR):**
   - **Definition:** The range between the first and third quartiles (Q3 - Q1).
   - **Calculation:** IQR = Q3 - Q1.
   - **Significance:** A robust measure of spread that is less affected by extreme values compared to the range.
4. **Five-Number Summary:**
   - **Definition:** Consists of the minimum value, Q1, Q2 (median), Q3, and the maximum value.

- **Significance:** Provides a concise summary of the distribution, offering insights into central tendency and spread.

5. **Boxplots (Box-and-Whisker Plots):**
   - **Definition:** Graphical representation of the five-number summary, displaying the distribution's central tendency and spread.
   - **Components:** Box represents the interquartile range (IQR), and whiskers extend to the minimum and maximum values.
   - **Significance:** Visually depicts the spread, skewness, and potential outliers in the data.



2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

Variance and Standard Deviation

Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2, \tag{2.6}$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

Variance and standard deviation. In Example 2.6, we found $\bar{x} = \$58{,}000$ using Eq. (2.1) for the mean. To determine the variance and standard deviation of the data from that example, we set $N = 12$ and use Eq. (2.6) to obtain

$$\begin{aligned}\sigma^2 &= \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2 \\ &\approx 379.17 \\ \sigma &\approx \sqrt{379.17} \approx 19.47.\end{aligned}$$

■

Data mining functions:

1. **Generalization:**
   - **Definition:** The process of summarizing and condensing large volumes of data into more compact, understandable forms.
   - **Purpose:** Helps in identifying trends, patterns, and relationships in the data, making it more manageable and interpretable.
2. **Association and Correlation Analysis:**
   - **Definition:** Identifying relationships, associations, or connections among different variables or items in a dataset.

- **Purpose:** Reveals patterns such as items frequently bought together in retail transactions, aiding in decision-making and business strategy.

3. **Classification:**
   - **Definition:** Assigning predefined categories or labels to new, unseen data based on patterns learned from a labeled dataset.
   - **Purpose:** Used for tasks like spam filtering, image recognition, and credit scoring by predicting the class of new instances.
4. **Cluster Analysis:**
   - **Definition:** Grouping similar data points or objects into clusters or segments based on certain similarity criteria.
   - **Purpose:** Helps identify natural groupings within data, facilitating insights into underlying structures or patterns.
5. **Outlier Analysis:**
   - **Definition:** Identifying data points that deviate significantly from the overall pattern or distribution in a dataset.
   - **Purpose:** Useful for detecting anomalies, errors, or unexpected patterns, aiding in quality control and anomaly detection.

Measuring Data Similarity and Dissimilarity

In data mining applications, such as clustering, outlier analysis, and nearest-neighbor classification, we need ways to assess how alike or unalike objects are in comparison to one another. For example, a store may want to search for clusters of customer objects, resulting in groups of customers with similar characteristics (e.g., similar income, area of residence, and age). Such information can then be used for marketing. A cluster is a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters. Outlier analysis also employs clustering-based techniques to identify potential outliers as objects that are highly dis□similar to others. Knowledge of object similarities can also be used in nearest-neighbor classification schemes where a given object (e.g., a patient) is assigned a class label (relating to, say, a diagnosis) based on its similarity toward other objects in the model.

similarity and dissimilarity measures, which are referred to as **measures of proximity**. Similarity and dissimilarity are related. A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unalike. The higher the similarity value, the greater the similarity between objects. (Typically, a value of 1 indicates complete similarity, that is, the objects are identical.) A dissimilarity measure works the opposite way. It returns a value of 0 if the objects are the same (and therefore, far from being dissimilar). The higher the dissimilarity value, the more dissimilar the two objects are.

Data matrix (or *object-by-attribute structure*): This structure stores the n data objects in the form of a relational table, or n-by-p matrix (n objects $\times p$ attributes):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}. \tag{2.8}$$

- **Dissimilarity matrix** (or *object-by-object structure*): This structure stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}, \tag{2.9}$$

Similarity/Dissimilarity for Single Attributes

p and *q* are the attribute values for two data objects

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

Proximity calculation for Nominal attributes:

- For example, binary attribute, Gender = {Male, female} where Male is equivalent to binary 1 and female is equivalent to binary 0.
- Similarity value is 1 if the two objects contains the same attribute value, while similarity value is 0 implies objects are not at all similar.

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

- Here, Similarity value let it be denoted by p, among different objects are as follows.

$$p(Ram, sita) = 0$$
$$p(Ram, Laxman) = 1$$

Note : In this case, if q denotes the dissimilarity between two objects i *and* j with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

BINARY

Object x	Object y: 1	Object y: 0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

Similarity Measure with Symmetric Binary

- To measure the similarity between two objects defined by symmetric binary attributes using a measure called symmetric binary coefficient and denoted as $\mathcal{S}$ and defined below

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

The dissimilarity measure, likewise can be denoted as $\mathcal{D}$ and defined as

$$\mathcal{D} = \frac{Number\ of\ mismatched\ attribute\ values}{Total\ number\ of\ attributes}$$

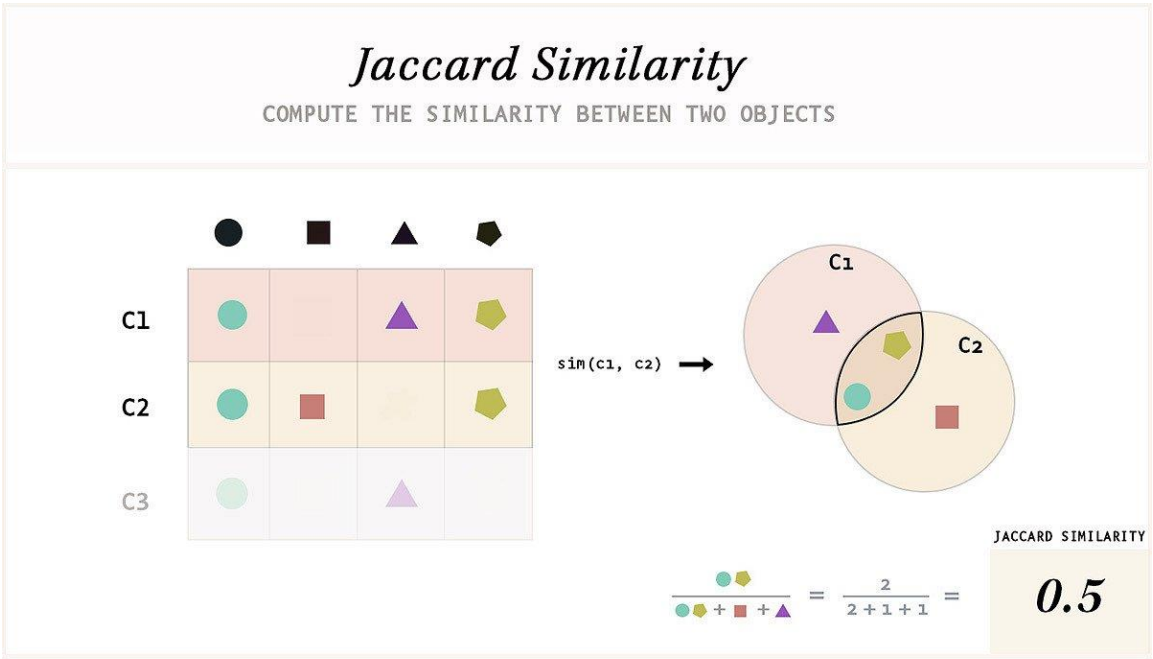
or

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

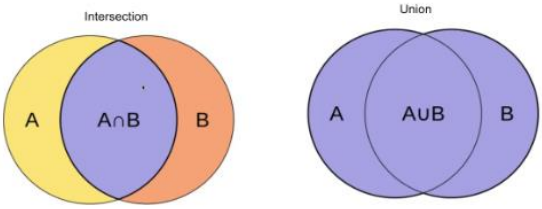
Note that, $\mathcal{D} = 1 - \mathcal{S}$

Proximity Measure with Asymmetric Binary

Example 2: Jaccard Coefficient



Jaccard coefficient



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

	User1	User2	User3
Books	B1	B1	A1
	B2	B3	A2
	B3	A1	A3
	B4	A4	B4

Number of books read by both User1 and User2 = 2

Number of books read by User1 or User2 = 6

J(User1, User2) = 2/6 = 33.33%

Similarly,

J(User2,User3) = 1/7= 14.28%

J(User1,User3) = 1/7= 14.28%

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

Proximity Measure with Categorical Attribute

- Binary attribute is a special kind of nominal attribute where the attribute has values with two states only.
- On the other hand, categorical attribute is another kind of nominal attribute where it has values with three or more states (e.g. color = {Red, Green, Blue}).
- If $s(x, y)$ denotes the similarity between two objects x an , then

$$s(x, y) = \frac{Number\ of\ matches}{Total\ number\ of\ attributes}$$

and the dissimilarity $d(x, y)$ is

$$d(x, y) = \frac{Number\ of\ mismatches}{Total\ number\ of\ attributes}$$

- If m = number of matches and a = number of categorical attributes with which objects are defined as

$$s(x, y) = \frac{m}{a} \quad \text{and} \quad d(x, y) = \frac{a-m}{a}$$