What is Datamining

Data Mining" refers to the extraction of useful information from a bulk of data or data warehouses.

Data mining is the process of discovering patterns, correlations, and anomalies within large datasets to predict outcomes. Using a variety of techniques from statistics, machine learning, and database systems, data mining transforms raw data into useful information. Here are some key components and steps involved in data mining:

1. **Data Cleaning**: Removing noise and inconsistent data.
2. **Data Integration**: Combining data from multiple sources.
3. **Data Selection**: Retrieving relevant data for the analysis.
4. **Data Transformation**: Converting data into an appropriate format for mining.
5. **Data Mining**: Applying algorithms to extract patterns.
6. **Pattern Evaluation**: Identifying truly interesting patterns representing knowledge.
7. **Knowledge Representation**: Presenting the mined knowledge in a comprehensible way.

Techniques Used in Data Mining

- **Classification**: Assigning items to predefined categories or classes.
- **Regression**: Predicting a numeric value based on input data.
- **Clustering**: Grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- **Association Rule Learning**: Discovering interesting relations between variables in large databases.
- **Anomaly Detection**: Identifying unusual data points.
- **Sequential Pattern Mining**: Identifying regular sequences in data.

Applications of Data Mining

- **Market Analysis and Management**: Understanding customer behavior, product trends, and market basket analysis.
- **Risk Management**: Fraud detection, credit scoring, and risk assessment.
- **Healthcare**: Predicting disease outbreaks, patient diagnosis, and treatment effectiveness.
- **Finance**: Stock market analysis, financial forecasting, and investment strategies.
- **Manufacturing**: Predictive maintenance and quality control

Data mining (knowledge discovery from data)

  - **Extraction of** interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or **knowledge** from huge amount of data

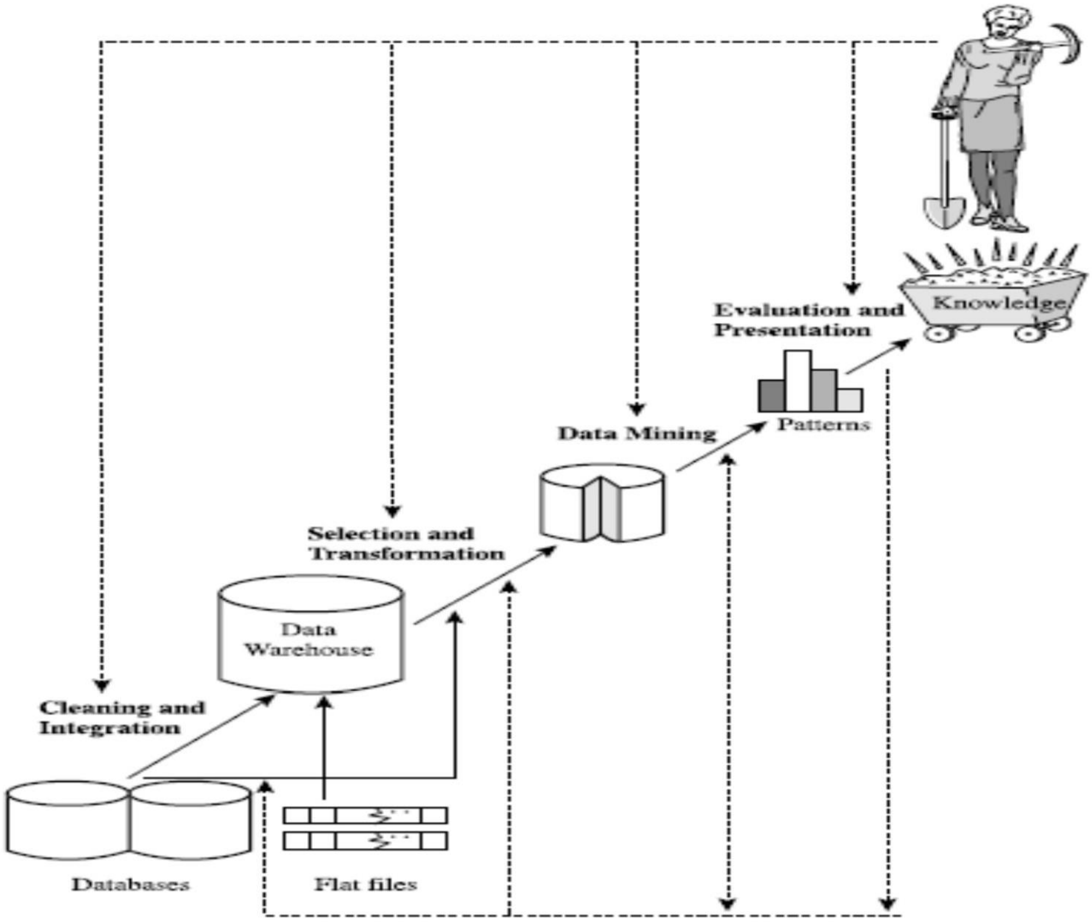
- Alternative names

  - Knowledge discovery (mining) in databases **(KDD),** knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Data mining is an essential step in KDD proces

What Is KDD?

Simply stated, data mining refers to *extracting or "mining" knowledge from large amounts of data*



KDD PROCESS STEPS:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Architecture of Data Mining

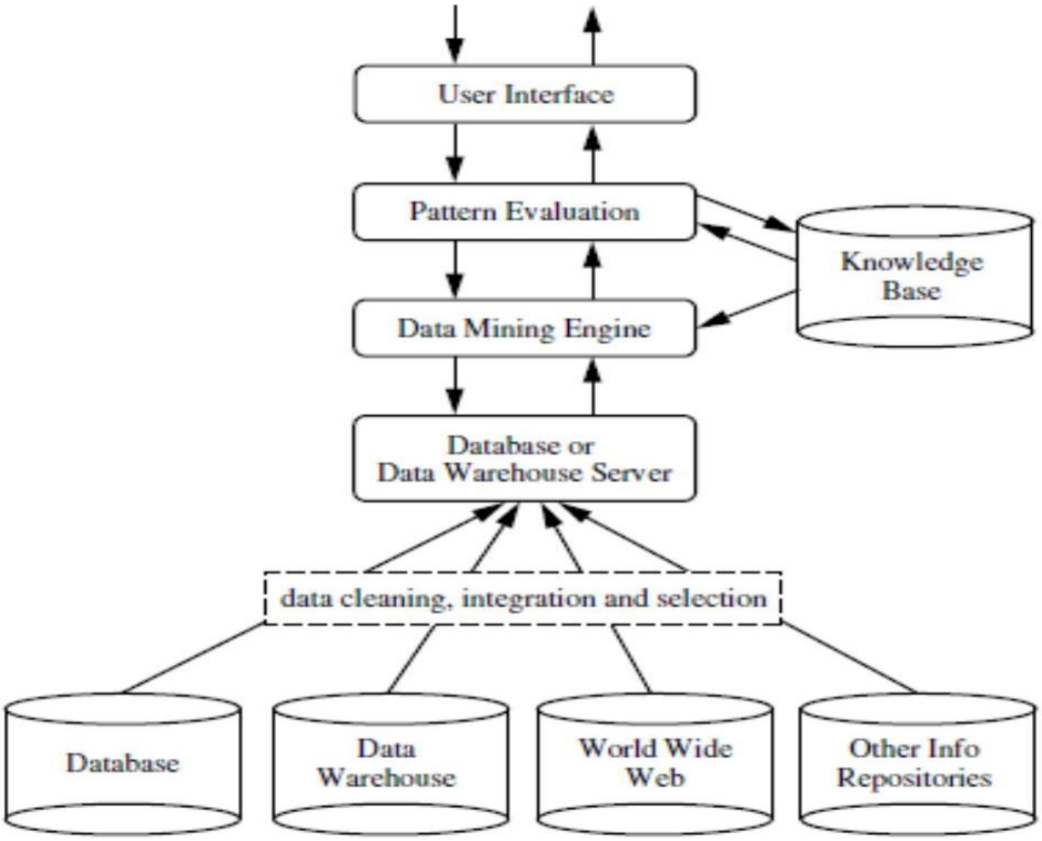


Figure: Architecture of a typical data mining system.

- **Database, data warehouse, World Wide Web, or other information repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

- **Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

- **Data mining engine:** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis

- **Pattern evaluation module:** This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns

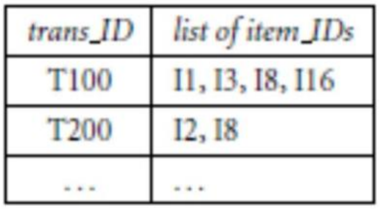
User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results

Data Mining—On What Kind of Data?

Types of data to be mined:

- Flat Files.
- Relational Databases.
- Data Warehouse.
- Transactional Databases.
- Multimedia Databases.
- Spatial Databases. (Maps)
- Time Series Databases. (Temporal data)
- World Wide Web (WWW)

- **Transactional Databases:** In general, a **transactional database** consists of a file where each record represents a transaction



trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

- **Advanced Data and Information Systems and Advanced Applications**

Relational database systems have been widely used in business applications. With the progress of database technology, various kinds of advanced data and information systems have emerged and are undergoing development to address the requirements of new applications.

- **Temporal Databases, Sequence Databases, and Time-Series Databases**

✓ A temporal database typically stores relational data that include time-related attributes. These attributes may involve several timestamps, each having different semantics.

✓ A sequence database stores sequences of ordered events, with or without a concrete notion of time. Examples include customer shopping sequences

✓ A time-series database stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly). Examples include data collected from the stock exchange

- **Spatial Databases and Spatiotemporal Databases**

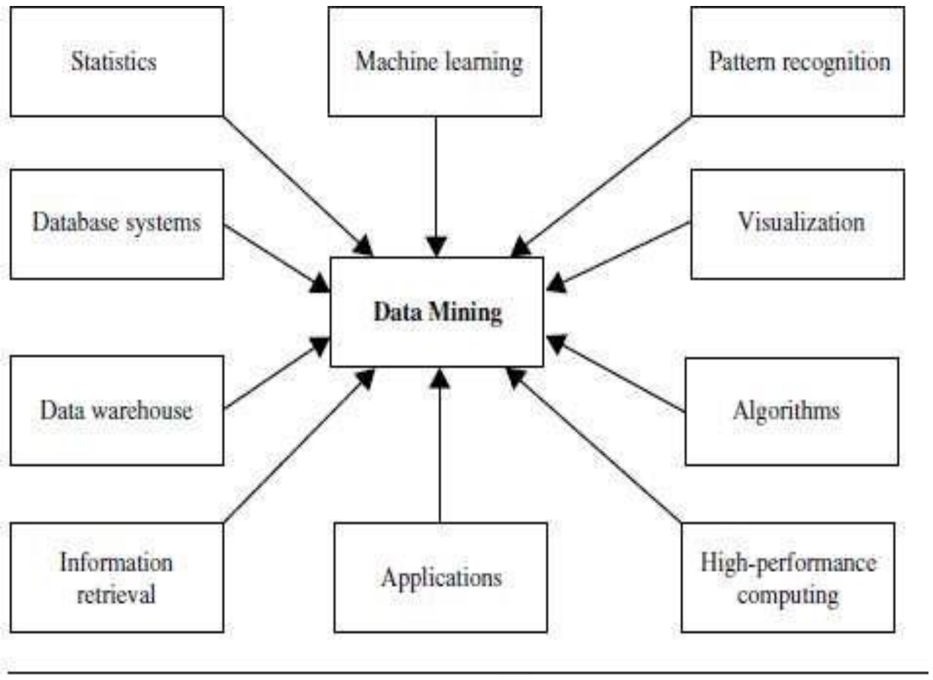
✓ Spatial databases contain spatial-related information. Examples include geographic (map) databases, very large-scale integration (VLSI) or computed-aided design databases, and medical and satellite image databases

✓ A spatial database that stores spatial objects That change with time is called spatio temporal database

- **Object-Relational Databases**

Conceptually, the object-relational data model inherits the essential concepts of object-oriented databases

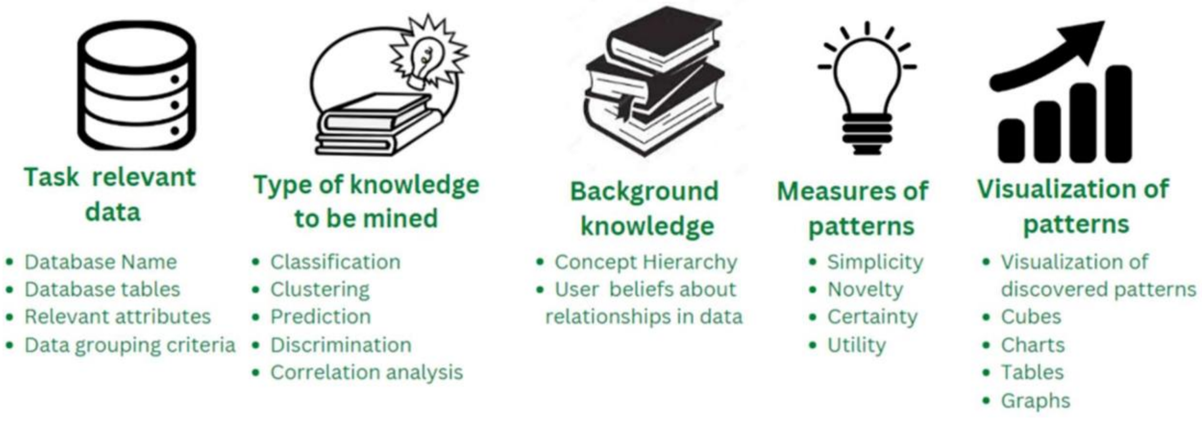
- ✓ Data mining technologies



Data mining adopts techniques from many domains.

Data mining Tasks

Data mining task primitives refer to the basic building blocks or components that are used to construct a data mining process. These primitives are used to represent the most common and fundamental tasks that are performed during the data mining process.



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

The Data Mining Task Primitives are as follows:

1. The 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 data may include specific attributes, variables, or characteristics that are relevant to the task at hand, such as customer demographics, sales data, or website usage statistics.

For **example**: Extracting the database name, database tables, and relevant required attributes from the dataset from the provided input database.

2. **Kind of knowledge to be mined:** It refers to the type of information or insights that are being sought through the use of data mining techniques. This describes the data mining tasks that must be carried out. It includes various tasks such as classification, clustering, discrimination, characterization, association, and evolution analysis.

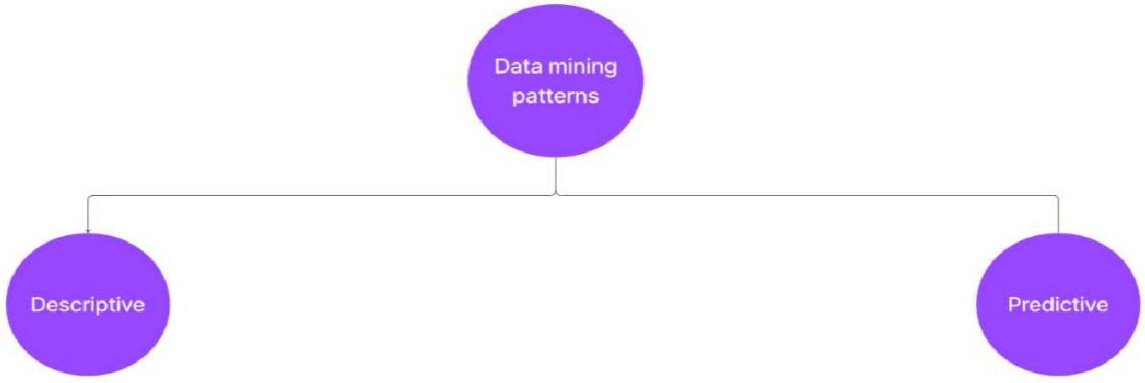
3. **Background knowledge to be used in the discovery process:** It refers to any prior information or understanding that is used to guide the data mining process. This can include domain-specific knowledge, such as industry-specific terminology, trends, or best practices, as well as knowledge about the data itself.

4. 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 frequency, confidence, or lift.

5. **Representation for visualizing the discovered pattern:** It refers to the methods used to represent the patterns or insights 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.

Data Mining Patterns

Based on the data functionalities, patterns can be further classified into two categories.

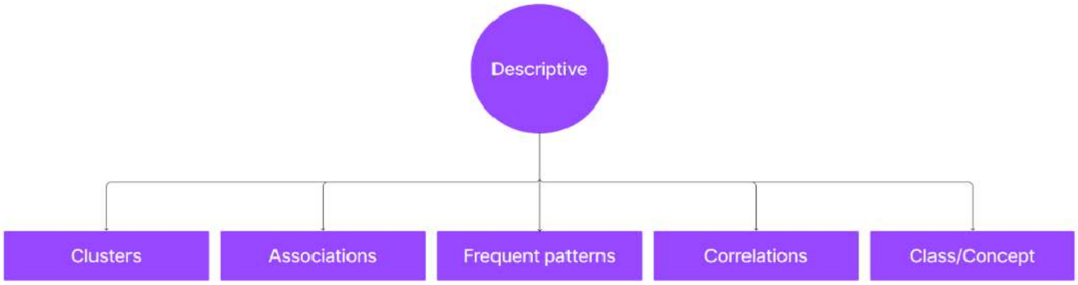


Descriptive data mining is often used to summarize or explore the data, and it can be used to answer questions such as: What are the most common patterns or relationships in the data? Are there any clusters or groups of data points that share common characteristics? What are the outliers in the data, and what do they represent?

Predictive Data Mining: This category of data mining is concerned with developing models that can predict future behavior or outcomes based on historical data.

Descriptive Tasks

Descriptive tasks aim to summarize and understand the characteristics of the data.



1. **Association Rule Learning**
   - **Objective**: Discover interesting relationships between variables in large datasets.
   - **Example**: Market basket analysis to find product purchase patterns.
2. **Clustering**
   - **Objective**: Group similar items or data points into clusters.
   - **Example**: Customer segmentation to identify distinct groups based on purchasing behavior.
3. **Summarization**
   - **Objective**: Provide a compact representation of the dataset.
   - **Example**: Generating summary statistics like mean, median, and standard deviation.
4. **Sequence Discovery**
   - **Objective**: Find sequential patterns in data.
   - **Example**: Analyzing web server logs to understand user navigation patterns on a website.

Predictive Tasks

Predictive tasks aim to predict unknown or future values of other variables of interest.

1. **Classification**
   - **Objective**: Assign items to predefined categories or classes.
   - **Example**: Classifying emails as spam or not spam.
2. **Regression**
   - **Objective**: Predict a continuous numeric value.
   - **Example**: Predicting housing prices based on features like location and size.
3. **Anomaly Detection**
   - **Objective**: Identify outliers or rare events in the data.
   - **Example**: Detecting fraudulent transactions in credit card data.
4. **Time Series Forecasting**
   - **Objective**: Predict future values based on previously observed values.
   - **Example**: Forecasting stock prices or sales over time.
5. **Recommendation Systems**
   - **Objective**: Predict preferences and recommend items to users.
   - **Example**: Suggesting movies or products to users based on their past behavior and preferences.

Hybrid Tasks

Some tasks can be seen as a combination of descriptive and predictive aspects.

1. **Dimensionality Reduction**
   - **Objective**: Reduce the number of variables under consideration while maintaining the dataset's integrity.
   - **Example**: Principal Component Analysis (PCA) to simplify data visualization and reduce computational complexity.
2. **Feature Selection**

- **Objective**: Identify the most relevant features for building predictive models.
- **Example**: Selecting key factors that influence customer churn in a subscription service.

Class/concept description: Data entries are associated with labels or classes. For instance, in a library, the classes of items for borrowed items include books and research journals, and customers' concepts include registered members and not registered members. These types of descriptions are class or concept descriptions. These class or concept definitions are referred to as class/concept descriptions.

Data Characterization: This refers to the summary of general characteristics or features of the class that is under the study. The output of the data characterization can be presented in various forms include pie charts, bar charts, curves, multidimensional data cubes.

Example: To study the characteristics of software products with sales increased by 10% in the previous years. To summarize the characteristics of the customer who spend more than$5000 a year at AllElectronics, the result is general profile of those customers such as that they are 40-50 years old, employee and have excellent credit rating.

Data Discrimination: It compares common features of class which is under study. It is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

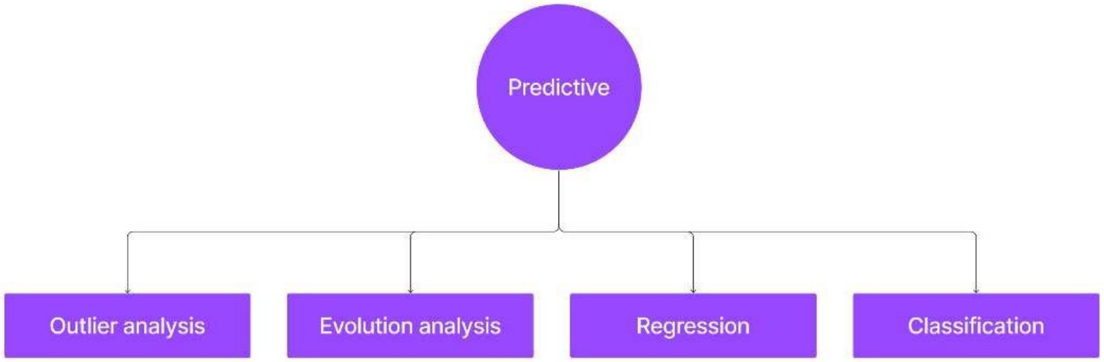
Example: we may want to compare two groups of customers those who shop for computer products regular and those who rarely shop for such products(less than 3 times a year), the resulting description provides a general comparative profile of those customers, such as 80% of the customers who frequently purchased computer products are between 20 and 40 years old and have a university degree, and 60% of the customers who infrequently buys such products are either seniors or youth, and have no university degree.

Frequent patterns: These are data points that occur more often in the dataset. There are many kinds of recurring patterns, such as frequent items, frequent subsequence, and frequent sub-structure.

Associations: It shows the relationships between data and pre-defined association rules. For instance, a shopkeeper makes an association rule that 70% of the time, when a football is sold, a kit is bought alongside. These two items can be combined together to make an association.

Correlations: This is performed to find the statistical correlations between two data points to find if they have positive, negative, or no effect.

Clusters: This is the formation of a group of similar data points. Each point in the collection is somewhat similar but very different from other members of different groups.



Classification: It helps predict the label of unknown data points with the help of known data points**.** For instance, if we have a dataset of X-rays of cancer patients, then the possible labels would be **cancer patient** and **not cancer patient**. These classes can be obtained by data characterizations or by data discrimination.

Regression: Unlike classification, regression is used to find the missing numeric values from the dataset. It is also used to predict future numeric values as well. For instance, we can find the behavior of the next year's sales based on the past twenty years' sales by finding the relation between the data.

Outlier analysis: Not all data points in the dataset need to follow the same behavior. Data points that don't follow the usual behavior are called outliers. Analysis of these outliers is called outlier analysis. These outliers are not considered while working on the data.

Evolution analysis: As the name suggests, those data points change their behavior and trends with time.

Data Mining Functionalities

There are a number of ***data mining functionalities***. These include characterization

- Data characterization and discrimination
- The mining of frequent patterns, associations
- correlations classification and regression
- clustering analysis
- outlier analysis

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: **Data characterization** is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

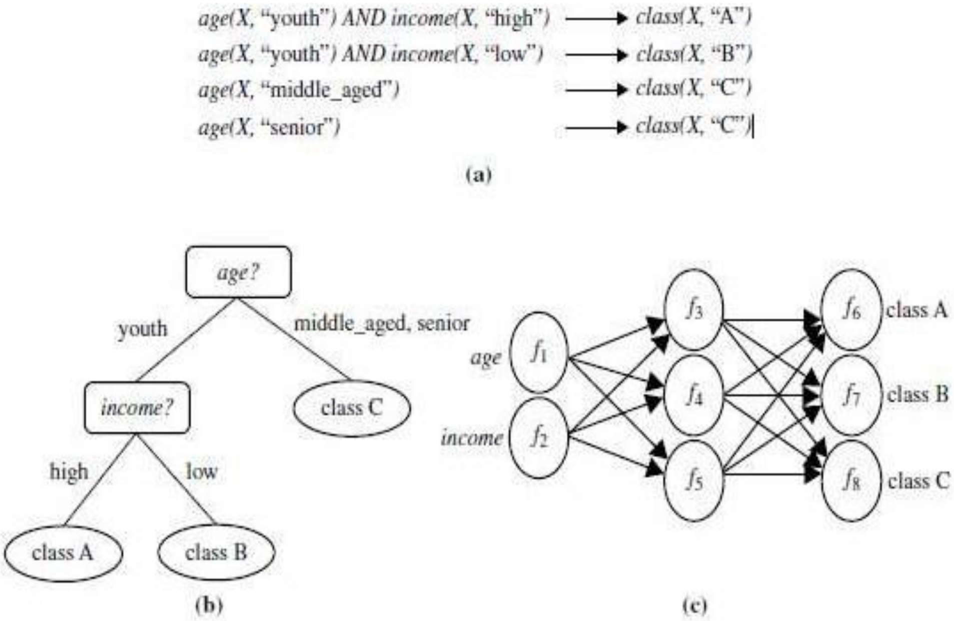
- **Data discrimination** is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.

- **Mining Frequent Patterns, Associations, and Correlations Frequent patterns**, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.

- **Association analysis.** Suppose that, as a marketing manager at *AllElectronics*, you want to know which items are frequently purchased together (i.e., within the same transaction).

- An example of such a rule, mined fromthe *AllElectronics* transactional database, is *buys*.*X*, “*computer*”/)*buys*.*X*, “*software*”/ [*support* D 1%, *confidence* D 50%], where *X* is a variable representing a customer. A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together.

- **Classification** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.



age(X, “youth”) AND income(X, “high”) → class(X, “A”)
age(X, “youth”) AND income(X, “low”) → class(X, “B”)
age(X, “middle_aged”) → class(X, “C”)
age(X, “senior”) → class(X, “C”)
(a)
age?
youth
middle_aged, senior
income?
class C
high
low
class A
class B
(b)
age
income
f_1
f_2
f_3
f_4
f_5
f_6
f_7
f_8
class A
class B
class C
(c)

Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Regression analysis

is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data.

Cluster Analysis

Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of ***maximizing the intraclass similarity and minimizing the interclass similarity.***

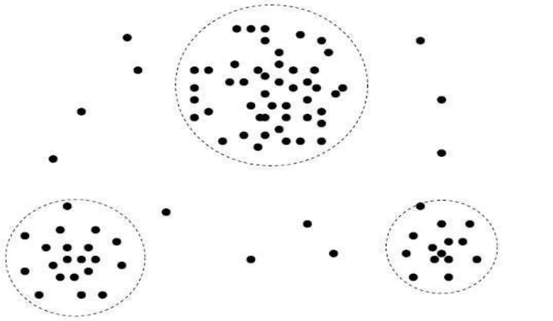


Figure 1.10 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

outlier analysis:

- Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

Tools and Technologies

- ETL Tools: Apache NiFi, Talend, Informatica
- Data Warehouses: Amazon Redshift, Google BigQuery, Microsoft Azure SQL Data Warehouse, Snowflake
- Data Mining Tools: RapidMiner, Weka, KNIME, SAS Enterprise Miner, IBM SPSS
- BI Tools: Tableau, Power BI, QlikView

Data Objects

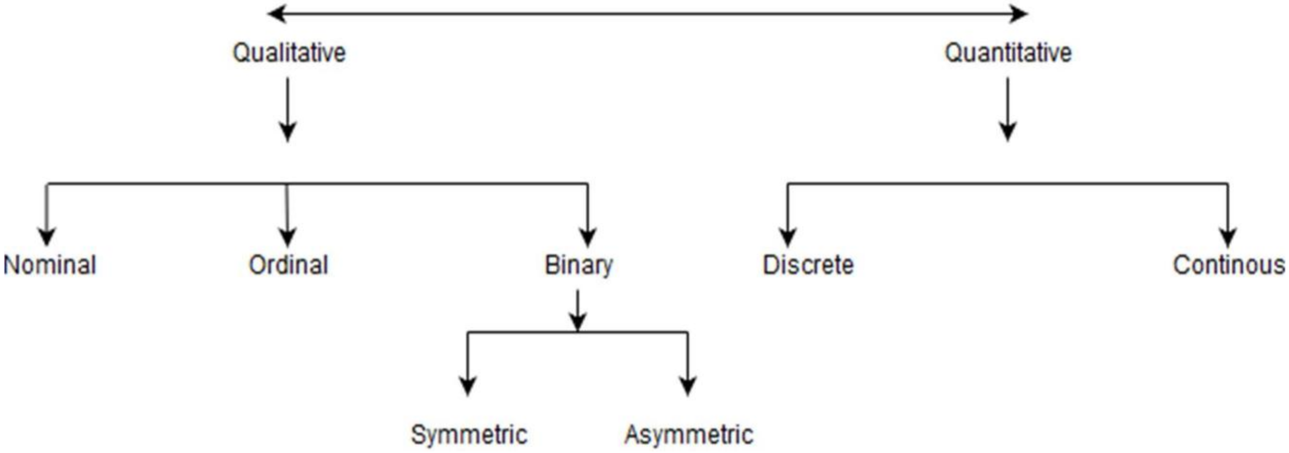
Data objects, often referred to as records, tuples, instances, or entities, are the primary units of information in a data warehouse or a data mining context. Each data object represents a real-world entity and is characterized by a set of attributes.

Attribute types

Attribute types and their characteristics is crucial for effective data preprocessing, analysis, and mining in data warehousing environments.

This is the First step of Data Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is description of attribute types.

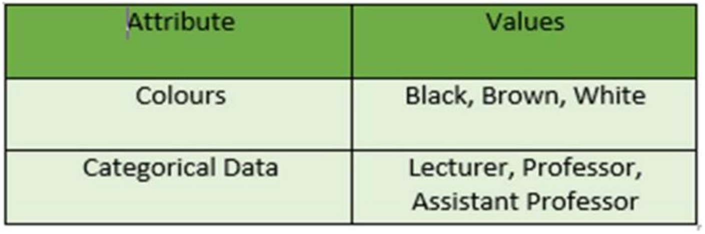
1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
2. Quantitative (Discrete, Continuous)



Qualitative Attributes:

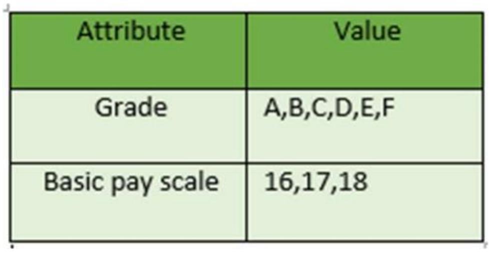
Nominal Attributes – related to names: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represent some category or state and that's why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values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

Example:



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

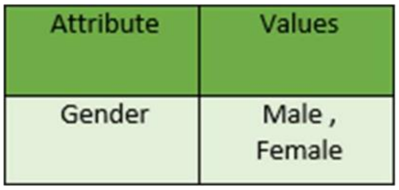
Ordinal Attributes: The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.



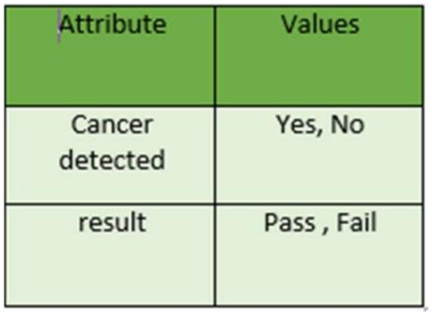
Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

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

i) **Symmetric:** Both values are equally important (Gender).
ii) **Asymmetric:** Both values are not equally important (Result).



Attribute	Values
Gender	Male , Female



Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Quantitative Attributes

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**

Interval Attributes

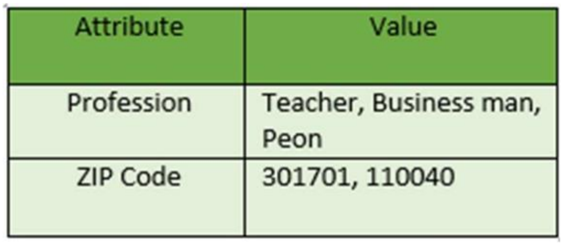
- **Definition**: Attributes with meaningful intervals between values but no true zero point.
- **Examples**: Calendar dates, temperature in Celsius or Fahrenheit.
- **Operations**: Mean, standard deviation, correlation.

Ratio Attributes

- **Definition**: Attributes with all the properties of interval attributes and a meaningful zero point.
- **Examples**: Age, salary, temperature in Kelvin, height, weight.
- **Operations**: All arithmetic operations, geometric mean, harmonic mean.

Discrete: Discrete data have finite values it can be numerical and can also be in categorical form. These attributes have finite or countably infinite set of values. Examples include the number of children, test scores, and product ratings.

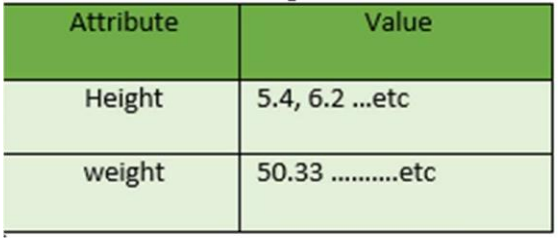
Example



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

Continuous: Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3. Examples include height, weight, and temperature.

Example :



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

The following properties (operations) of numbers are typically used to describe attributes.

1. **Distinctness** = and #
2. **Order<,** :S, >, and 2".
3. **Addition+** and -
4. **Multiplication*** and/

Different attribute types

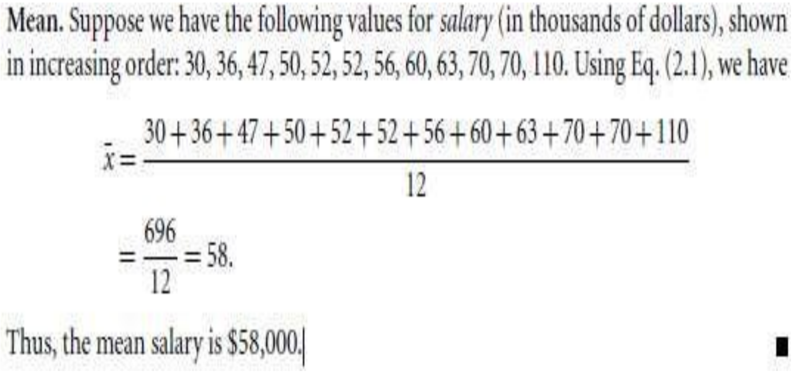
	Description	Examples	**Operations**
Nominal	The values of a nominal attribute is just different names; i.e., nominal values provide only enough information to distinguish one object from another. **(=≠)**	zip codes, employee ID numbers, eye color, gender	mode, entropy, Contingency correlation, x^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerais, *{good, better, best},* grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature **in** Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, Percent Variation

Statistical Descriptions of Data

Measuring the Central Tendency: Mean, Median, and Mode

The **mean** of this set of values is

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



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: Let’s find the median of the data from Example The data are already sorted in increasing order. There is an even number of observations (i.e., 12);

therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list).

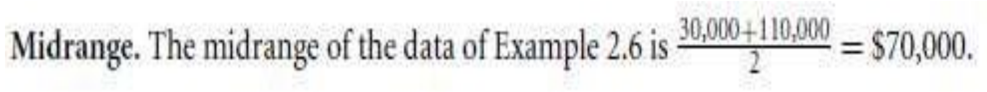
By convention, we assign the average of the two middlemost values as the median; that is,
52+56= 108=

2 2
54.

Thus, the median is $54,000.

Mode: identify which value in the data set occurs most often.

Range: which is the difference between the largest and smallest value in the data set, describes how well the central tendency represents the data



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

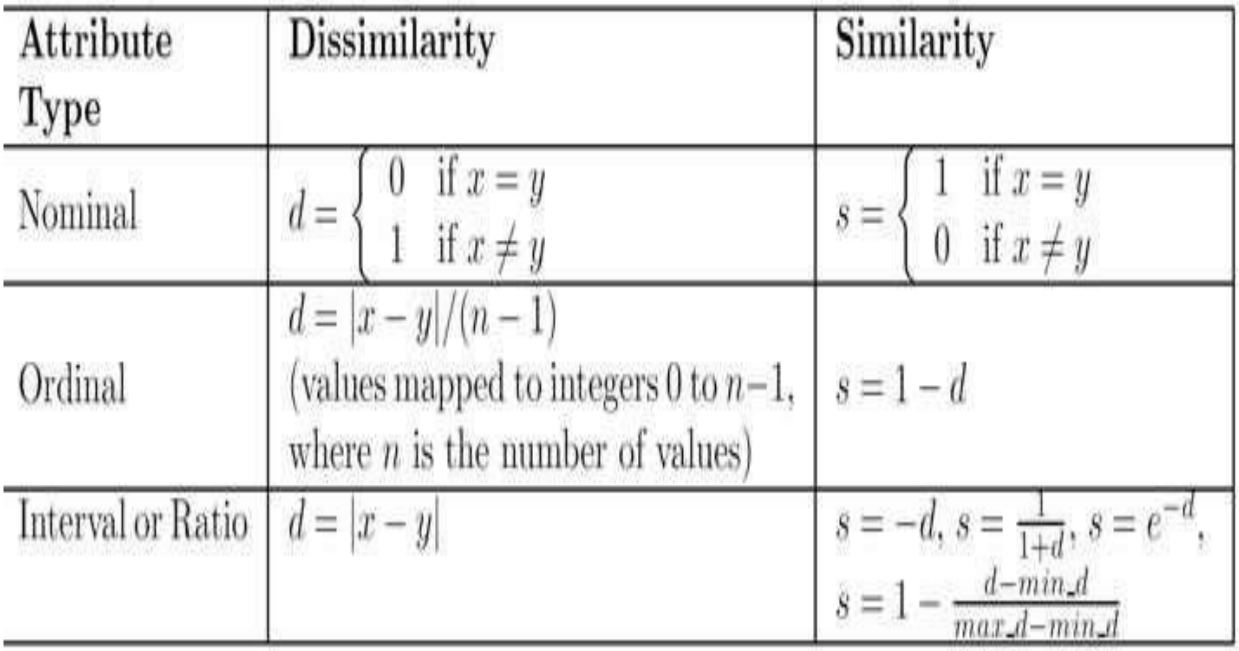
The **range** of the set is the difference between the largest (max()) and smallest (min()) values.

Measures of Similarity and Dissimilarity

- Similarity and dissimilarity are important because they are used by a number of data mining techniques, such *as* clustering, nearest neighbor classification, and anomaly detection
- Informally, the **similarity** between two objects is a numerical measure of the degree to which the two objects are alike. Consequently, similarities are *higher* for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

The **dissimilarity** between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities are *lower* for more similar

- ➢ Similarity and Dissimilarity between Simple Attributes



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

- Dissimilarities between Data Objects(multiple attributes)

Distances

- We first present some examples, and then offer a more formal description of distances in terms of the properties common to all distances. The Euclidean distance, d, between two points, x and y, in one-, two-, three-, or higher dimensional space, is given by the following familiar formula:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions and xk and yk are respectively, the kth attributes (components) of x and y

- The Euclidean distance measure given in above Equation is generalized by the **Minkowski** distance metric shown as

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r},$$

- where *r* is a parameter. The following are the three most common examples of Minkowski distances.

- ✓ *r* = 1. City block (Manhattan)distance. A common example is the Hamming **distance,** which is the number of bits that are different between two objects that have only binary attributes, i.e., between two binary vectors.(L1)
- ✓ *r* = 2. Euclidean distance(L2)
- ✓ *r* =$_\infty$Supremum distance. This is the maximum difference between any attribute of the objects. More formally, distance is defined by (L_∞)

$$d(\mathbf{x}, \mathbf{y}) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Distances, such as the Euclidean distance, have some well-known properties.

If d(x, y) is the distance between two points, x and y, then the following properties hold.

1. Positivity

(a) d(x, x) > =0 for all x and y

(a) d(x, Y) = 0 only if x = Y.

Ex: find dissimilarity among points p1(0,2) p2(2,0) p3(3,1) p4(5,1) Euclidean distance matri

	pl	p2	p3	p4
Pl	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

<u>Similarities between Data Objects</u>:

- If s(x, y) is the similarity between points x and y, then the typical properties of similarities are the following:

1. s(x,y) = 1 only if x =y (0 <= s <= 1)

2. s(x,y) = s(y, x) for all x and y. (Symmetry)

- similarity measure can easily be converted to a metric distance. The cosine and Jaccard similarity measures are two examples

- Examples of Proximity Measures

This section provides specific examples of some similarity and dissimilarity measures

<u>Similarity Measures for Binary Data</u>

Similarity measures between objects that contain only binary attributes are called similarity coefficients, and typically have values between 0 and 1. A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.

Let x and y be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binary vectors, Ieads to the following four quantities (frequencies:)

f00 : the number of attributes where x is 0 and y is 0

f01 : the number of attributes where x is 0 and y is 1

f10 : the number of attributes where x is 1 and y is 0

f11 : the number of attributes where x is 1 and y is 1

<u>Simple Matching Coefficient</u>

- One commonly used similarity coefficient is the simple matching coefficient (SMC), which is defined as

$$\text{SMC}= \frac{\text{number of matching attribute values}}{\text{number of attributes}}$$

$$=\frac{f11+f00}{f01+f10+f00+f11}$$

This measure counts both presences and absences equally.

X=1011010001

Y=1101000011

F00=3, f01=2, f10=2, f11=3   SMC=6/10=0.6

J=3/7=0.43

Jaccard coefficient

- The Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes. The Jaccard coefficient, which is often symbolized by J is given by the following equation:

  $$J = \frac{\text{no.of matching presences}}{\text{no of attributes not involved in 00 matches}}$$

  $$= \frac{f11}{f01+f10+f11}$$

Example of SMC and J

$\mathbf{x} = 1\;0\;0\;0\;0\;0\;0\;0\;0\;0$

$\mathbf{y} = 0\;0\;0\;0\;0\;0\;1\;0\;0\;1$

$f_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$f_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$f_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$f_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$\begin{aligned}\text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

$$\text{J} = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

- is one of the most common measure of document similarity. If x and y are two document vectors, then
- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\mathit{cos}(\, \mathbf{d}_1, \mathbf{d}_2\,) = \langle \mathbf{d}_1,\mathbf{d}_2 \rangle / \|\mathbf{d}_1\|\; \|\mathbf{d}_2\|\;,$$

where $\langle \mathbf{d}_1,\mathbf{d}_2 \rangle$ indicates inner product or vector dot product $\mathbf{d}_1'\mathbf{d}_2$ of vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, and || d || is the length of vector d.

Example:

$\mathbf{d}_1$ = 3 2 0 5 0 0 0 2 0 0
$\mathbf{d}_2$ = 1 0 0 0 0 0 0 1 0 2

<**d**$_1$, **d**$_2$> = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

|| **d**$_1$ || = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{1/2}$ = (42) $^{1/2}$ = 6.481

|| **d**$_2$ || = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{1/2}$ = (6)$^{1/2}$ = 2.449

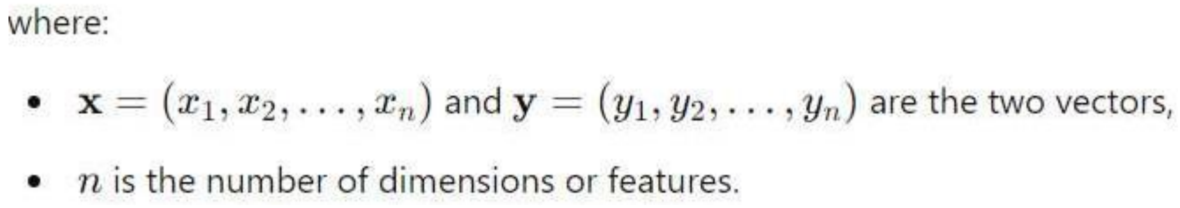
cos(**d**$_1$, **d**$_2$) = 0.3150.

Extended Jaccard Coefficient (Tanimoto Coefficient)

- The extended Jaccard coefficient can be used for document data and that reduces to the Jaccard coefficient in the case of binary attributes. The extended Jaccard coefficient is also known as the Tanimoto coefficient

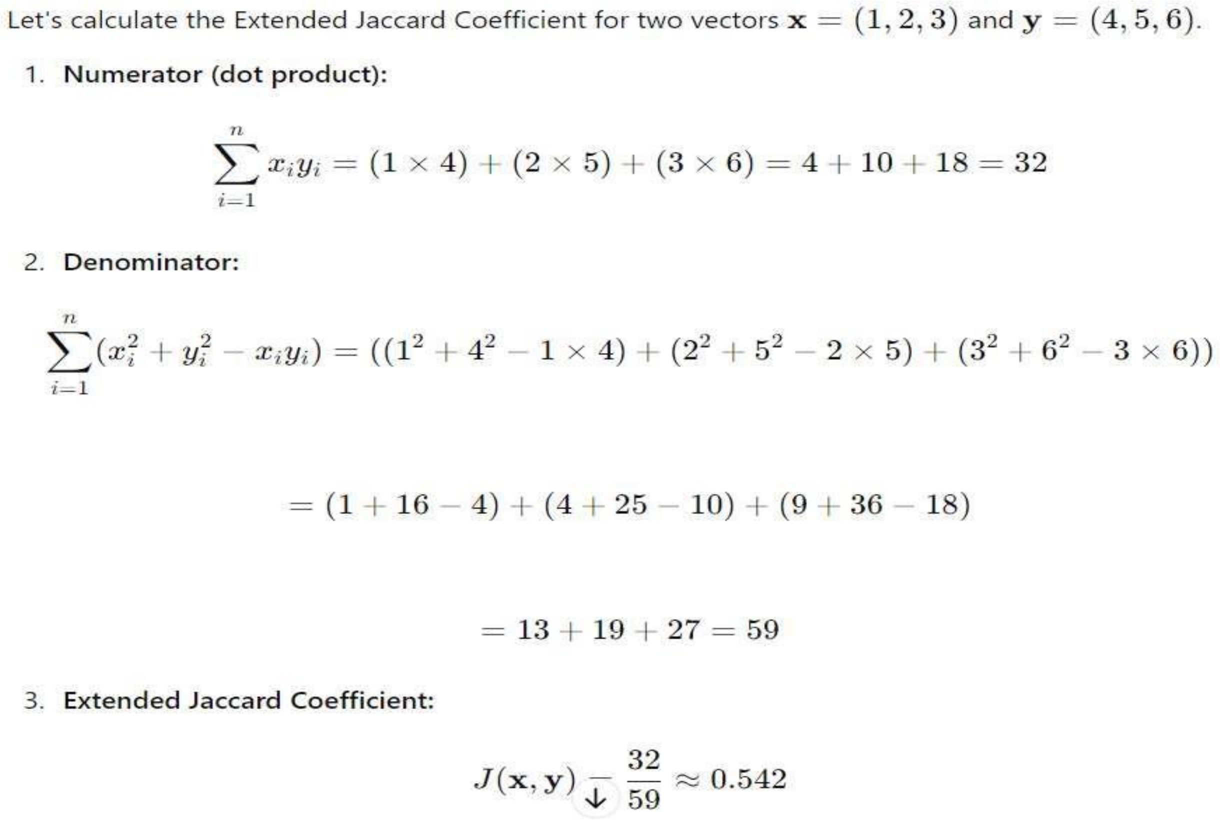
- The Extended Jaccard Coefficient, also known as the Tanimoto Coefficient, is a similarity measure that generalizes the Jaccard Index to real-valued vectors. It is commonly used in cases where data is represented by feature vectors, such as in information retrieval and machine learning.
- The formula for the Extended Jaccard Coefficient between two vectors x and y is:

$$J(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n}(x_i^2 + y_i^2 - x_i y_i)}$$



where:

- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are the two vectors,
- n is the number of dimensions or features.



Let's calculate the Extended Jaccard Coefficient for two vectors $\mathbf{x} = (1, 2, 3)$ and $\mathbf{y} = (4, 5, 6)$.

1. **Numerator (dot product):**

$$\sum_{i=1}^{n} x_i y_i = (1 \times 4) + (2 \times 5) + (3 \times 6) = 4 + 10 + 18 = 32$$

2. **Denominator:**

$$\sum_{i=1}^{n}(x_i^2 + y_i^2 - x_i y_i) = ((1^2 + 4^2 - 1 \times 4) + (2^2 + 5^2 - 2 \times 5) + (3^2 + 6^2 - 3 \times 6))$$

$$= (1 + 16 - 4) + (4 + 25 - 10) + (9 + 36 - 18)$$

$$= 13 + 19 + 27 = 59$$

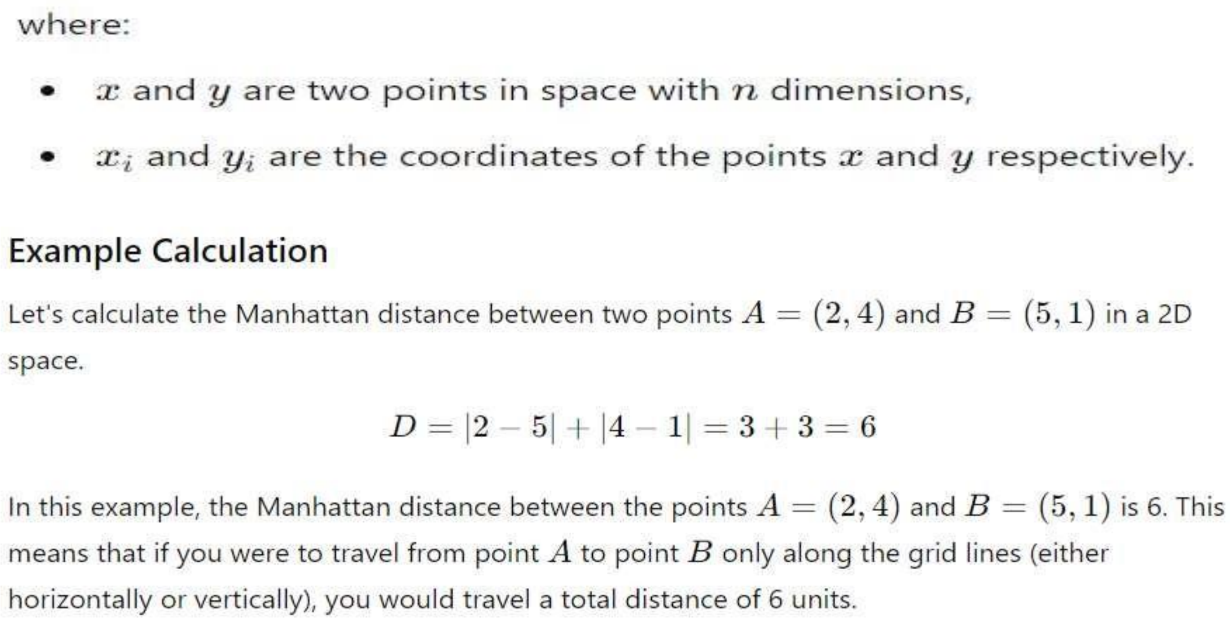
3. **Extended Jaccard Coefficient:**

$$J(\mathbf{x},\mathbf{y}) = \frac{32}{59} \approx 0.542$$

The Manhattan distance (L1 distance)

The Manhattan distance, also known as the L1 distance or taxicab distance, measures the distance between two points along the axes at right angles. It's defined by the formula:

$$D = \sum_{i=1}^{n} |x_i - y_i|$$



where:

- x and y are two points in space with n dimensions,
- x_i and y_i are the coordinates of the points x and y respectively.

Example Calculation

Let's calculate the Manhattan distance between two points $A = (2, 4)$ and $B = (5, 1)$ in a 2D space.

$$D = |2 - 5| + |4 - 1| = 3 + 3 = 6$$

In this example, the Manhattan distance between the points $A = (2, 4)$ and $B = (5, 1)$ is 6. This means that if you were to travel from point A to point B only along the grid lines (either horizontally or vertically), you would travel a total distance of 6 units.

The Minkowski distance

The Minkowski distance is a generalization of several different distance metrics in a normed vector space. It's defined by the formula:

$$D(p) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

where:

- x and y are two points in space with n dimensions,
- x_i and y_i are the coordinates of the points x and y respectively,
- p is a parameter that determines the type of distance metric.

The value of p determines the type of distance:

- $p = 1$: Manhattan distance (also known as L1 distance or taxicab distance)
- $p = 2$: Euclidean distance (also known as L2 distance)
- $p \to \infty$: Chebyshev distance (maximum distance)

Example Calculation

Let's calculate the Minkowski distance between two points $A = (2, 3, 5)$ and $B = (1, 1, 1)$ with different values of p.

1. **Manhattan Distance** ($p = 1$)**:**

$$D(1) = |2-1| + |3-1| + |5-1| = 1 + 2 + 4 = 7$$

2. **Euclidean Distance** ($p = 2$)**:**

$$D(2) = \sqrt{(2-1)^2 + (3-1)^2 + (5-1)^2} = \sqrt{1+4+16} = \sqrt{21} \approx 4.58$$

3. **Chebyshev Distance** ($p \to \infty$)**:**

$$D(\infty) = \max(|2-1|, |3-1|, |5-1|) = \max(1, 2, 4) = 4$$

These examples illustrate how the Minkowski distance can adapt to different metrics by changing the parameter p.

Data Preprocessing

Why Pre-process the Data -Data Cleaning-Data Integration-Data Reduction Data Transformation and Data Discretization

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising **data quality**, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

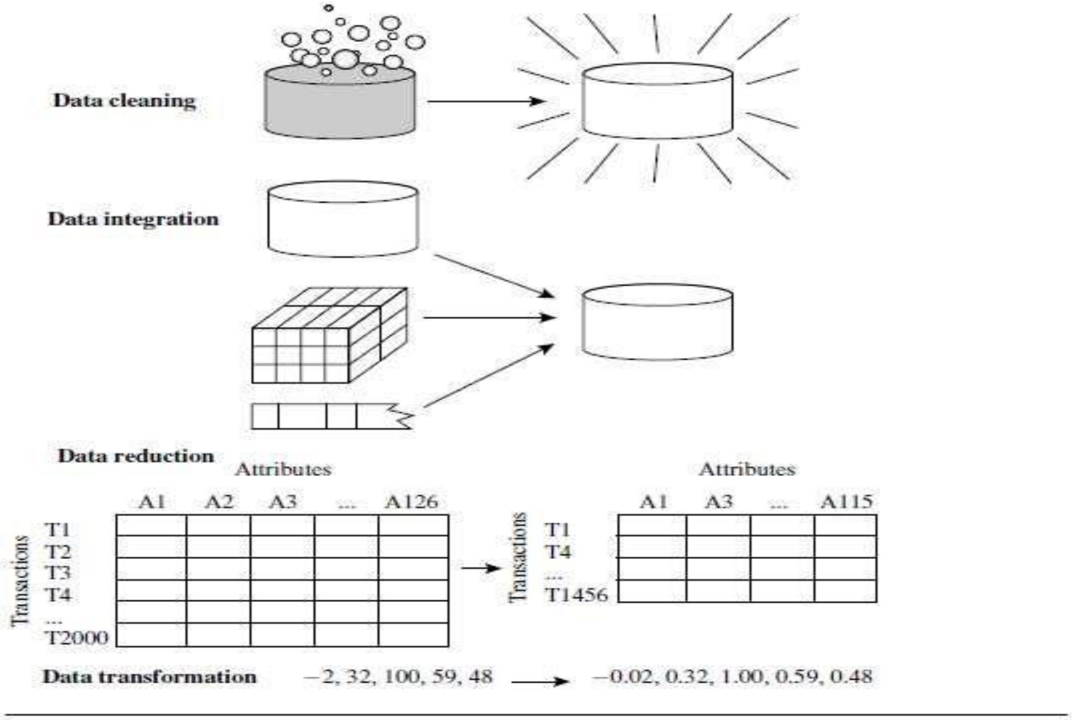


Figure 3.1 Forms of data preprocessing.

- Data Preprocessing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex ways.
    - Data Cleaning
    - Aggregation
    - Sampling
    - Dimensionality Reduction
    - Feature subset selection
    - Feature creation
    - Discretization and Binarization
    - variable Transformation.