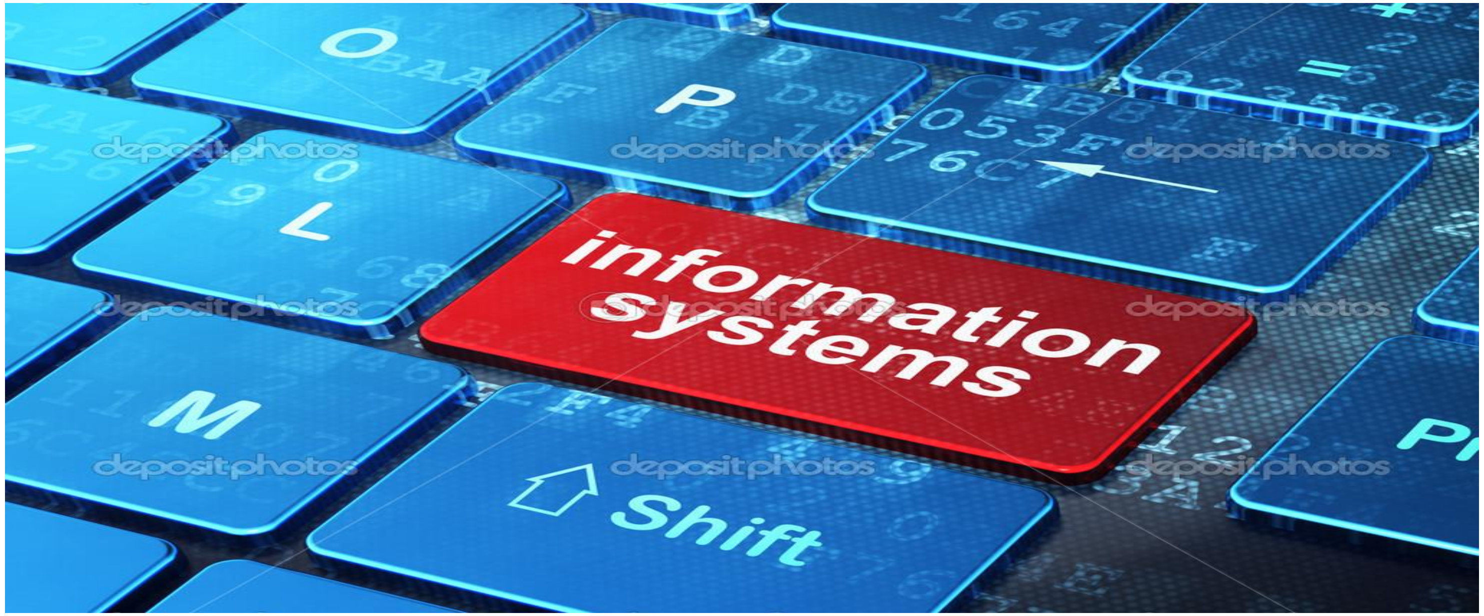
Advanced Topics in Information systems



SE204

Dr. Nelly Amer

Grades

- 60% on :

Attendance

Assignments

Quizzes

Midterm Exam

- Final term Exam: 40 %

Syllabus

- **Descriptive data analysis and visualization.**
- **Frequent pattern mining.**
- **Cluster and dimensionality reduction.**
- **Machine learning 1: Introduction.**
- **Machine learning 2: Classification.**
- **Machine learning 3: ROC analysis and regression.**
- **Machine learning 4: Deep learning.**

Descriptive data analysis and visualization.

- **Information system definition and elements.**
- **Data, information, and knowledge.**
- **Types of data.**
- **Data analysis.**
- **Data analysis and data analytics.**
- **Data analysis techniques.**
- **Descriptive data analysis methods.**
- **Descriptive data analysis phases.**
- **Data visualization.**
- **Descriptive data analysis and visualization tools.**

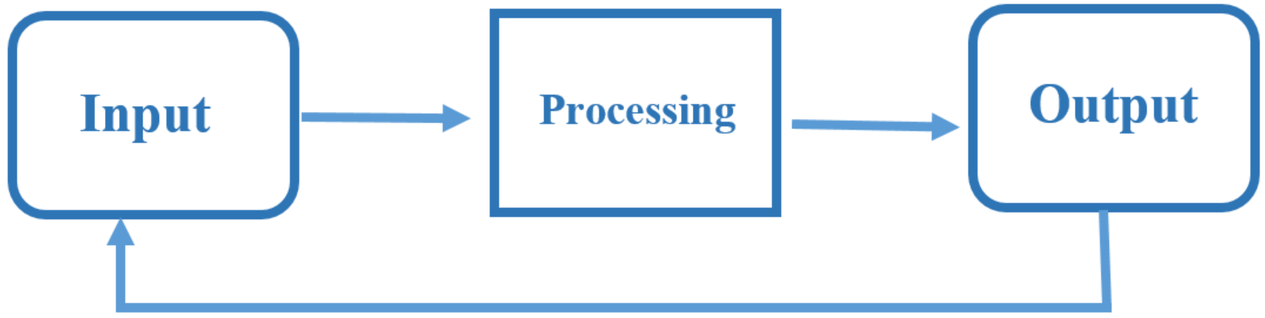
Information system definition and elements

Information system definition.

An information system is a set of interrelated components that collect, manipulate, store and disseminate information and provide a feedback mechanism to achieve a goal.

The feedback mechanism helps organizations achieve their goals by increasing profits, revenues, improving customer service, reduce costs, and supporting decision-making and control in organizations [1].

Information Systems



Input

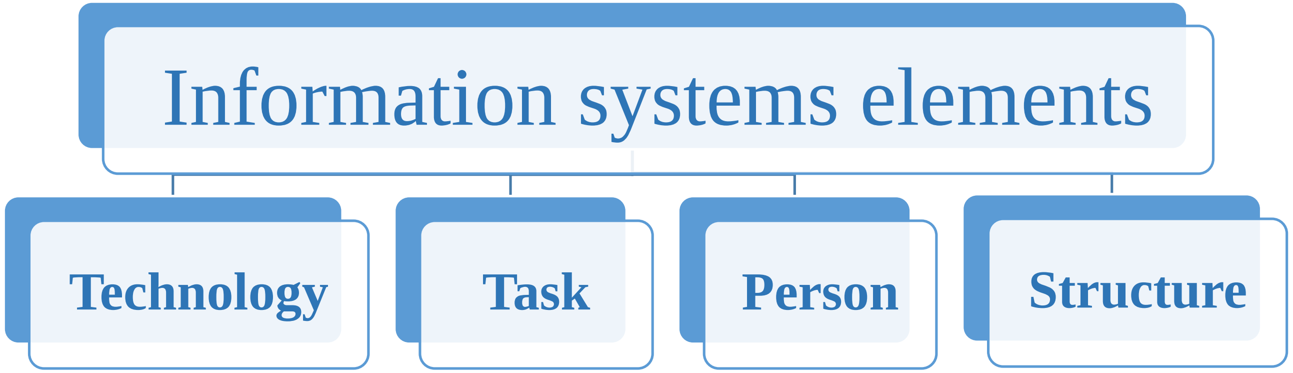
Processing

Output

Feedback

Information system definition and elements cont.

Information systems elements.



Technology: includes the hardware, software, and telecommunications equipment used to capture, process, store and disseminate information.

Task: activities necessary for the production of a good or service. These activities are supported by the flow of material, information, and knowledge between the different participants.

Person: The people component of an information system encompasses all the people directly involved in the system. These people include the **managers** who define the goals of the system, the **users**, the **data analysts**, and the **developers**.

Structure: the relationship between individuals people components. Thus, it encompasses hierarchical structures, relationships, and systems for evaluating people[1].

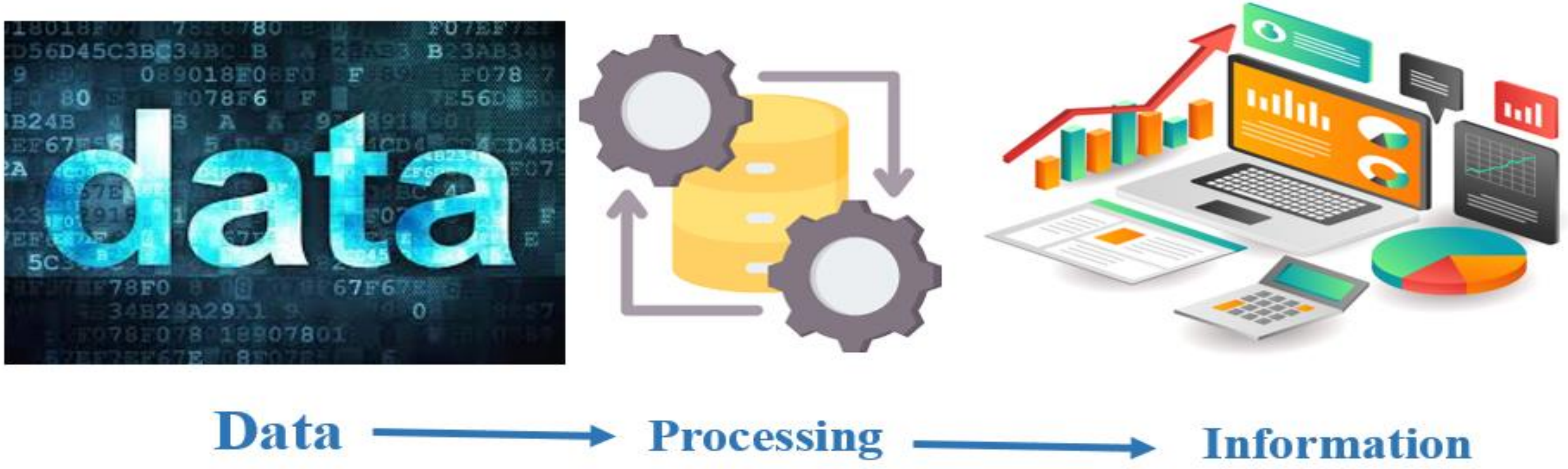
Data, information, and knowledge.

Data

Data is a collection of raw, unorganized facts and details like text, observations, figures, symbols and descriptions of things measured in terms of bits and bytes. Data does not carry any specific purpose or significance on its own.

Information

Data become information when analyzed, organized or classified, and possibly combined with other data in order to extract meaning, and to provide context, it provides answers to the questions: "who," "what," "where," and "when." it must be timely, accurate, and complete[2].



Data, information, and knowledge cont.

Knowledge.

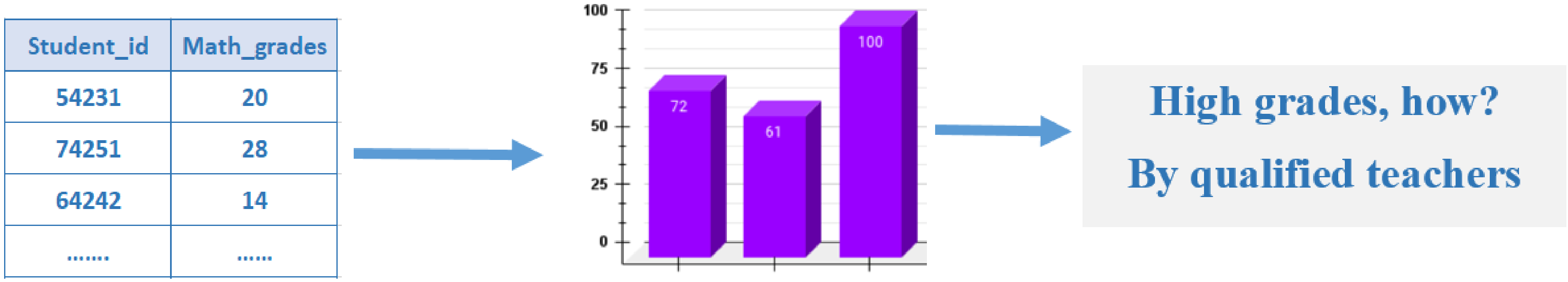
Knowledge can be defined as information combined with experience, context, and interpretation. Knowledge constitutes an additional semantic level derived from information, it provides answers to the question "how".

Data, information, and Knowledge.

Data: degrees of students in schools in final math exam in 6th grade, primary.

Information: percentage of success in math exam for school.

Knowledge: Schools with high success rates have qualified teachers in math, and vice versa.



Student_id	Math_grades
54231	20
74251	28
64242	14
…….	……

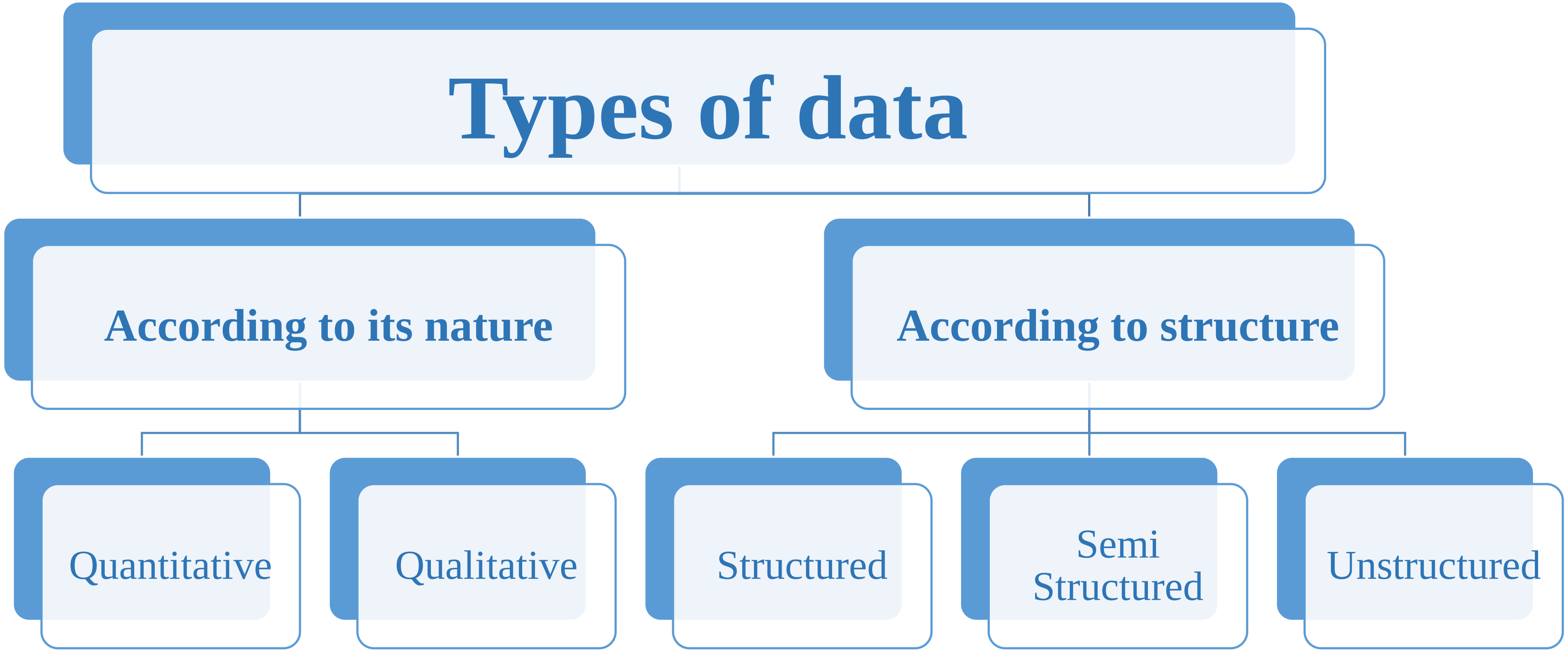
High grades, how?

By qualified teachers

Data **Information** **Knowledge**

Types of data.

Data types can be classified according to its nature and according to structure.



Types of data According to its nature.

Data can be classified according to its nature into : quantitative data, and qualitative data [4].

Quantitative (numeric)

Data that can be quantified and measured. This kind of data explains a trend or the results of research through numeric values.

This category of data can be subdivided into:

- **Discrete**: Data that consists of whole numbers (0, 1, 2, 3...). For example, the number of children in a family.
- **Continuous**: Data that can take any value within an interval. For example, people’s height (between 60 - 70 inches) or weight (between 90 and 110 pounds).

Types of data According to its nature cont.

Qualitative (categorical)

This kind of data is divided into categories based on non-numeric characteristics. It may or may not have a logical order, and it measures qualities and generates categorical answers.

It can be subdivided into :

- **Ordinal**: Meaning it follows an order or sequence. That might be the alphabet or the months of the year.
- **Categorical**: Meaning it follows no fixed order. For example, varieties of products sold.

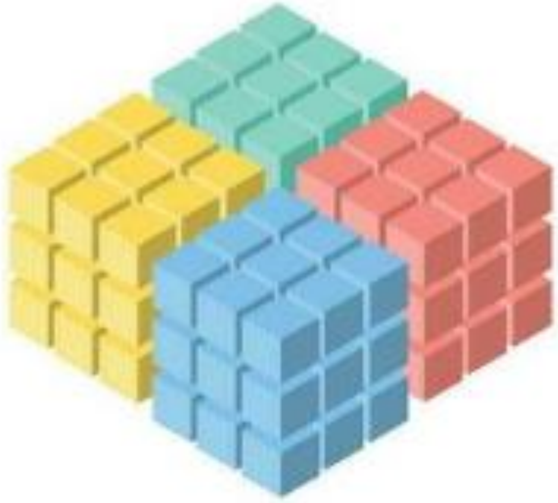
Types of data according to structure.

According to structure, data can be classified into : Structured data, semi structured data, and unstructured data.

Types of data cont.

Structured data.

- Structured data is data that can be stored in a table (rows and columns), and every instance in the table has the same structure (i.e., set of attributes.)
- Structured data is typically stored in a relational database (RDBMS) such as SQL Server, Oracle, and MySQL.
- Structured data can be managed using SQL (Structured Query Language).
- Structured data is often quantitative data, i.e., it usually consists of numbers or things that can be counted.
- Examples of structured data: consider the demographic data for a population, where each row in the table describes one person and consists of the same set of demographic attributes (name, age, date of birth, address, gender, education level, job status, credit card numbers, etc.) [3].



Types of data cont.

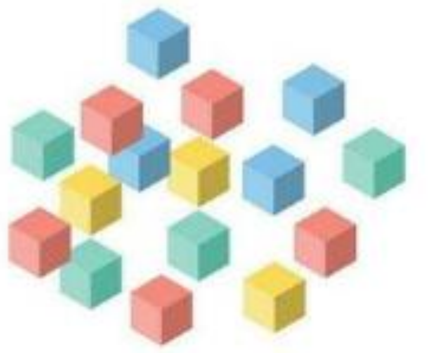
Semi structured data.

- This is the data which has some structure but there is no a data model, For example, a data set of webpages, with each webpage having a structure but this structure differing from one webpage to another.
- Semi structured data can be stored in NoSQL Databases; for non relational databases, such as MongoDB, Cassandra, and Couchbase.
- Semi structured data can be qualitative and quantitative data.
- Semi structured data examples: collections of human text (emails, tweets, text messages, posts, novels, etc.), XML(Extensible Markup Language) files, and JSON (**J**ava**S**cript **O**bject **N**otation) files[5].

Types of data cont.

Unstructured data.

- Unstructured data are data where each instance in the data set may have its own internal structure, and this structure is not necessarily the same in every instance, there is no data model, the data is stored in its native format.
- The unstructured data can be stored in NoSQL Databases; for non relational databases, such as MongoDB, Cassandra, and Couchbase.
- Structured data can be extracted from unstructured data using techniques from artificial intelligence (such as natural language processing and ML), digital signal processing, and computer vision.
- Unstructured data is often qualitative data, it can not be processed using conventional tools and methods.
- Unstructured data examples: sound, image, music, video, and multimedia files, audio files, and various other formats[5].



Data analysis

Data analysis definition.

data analysis is the process of identifying, cleaning, transforming, and modeling data to discover meaningful and useful information. The data is then crafted into a story via reports for analysis to support the critical decision-making process [6].

The two terms: data analysis and data analytics are often used interchangeably and could be confusing, and It's a common misconception that data analysis and data analytics are the same thing. **No they are not the same thing**, as we will show.

Data analysis and data analytics.

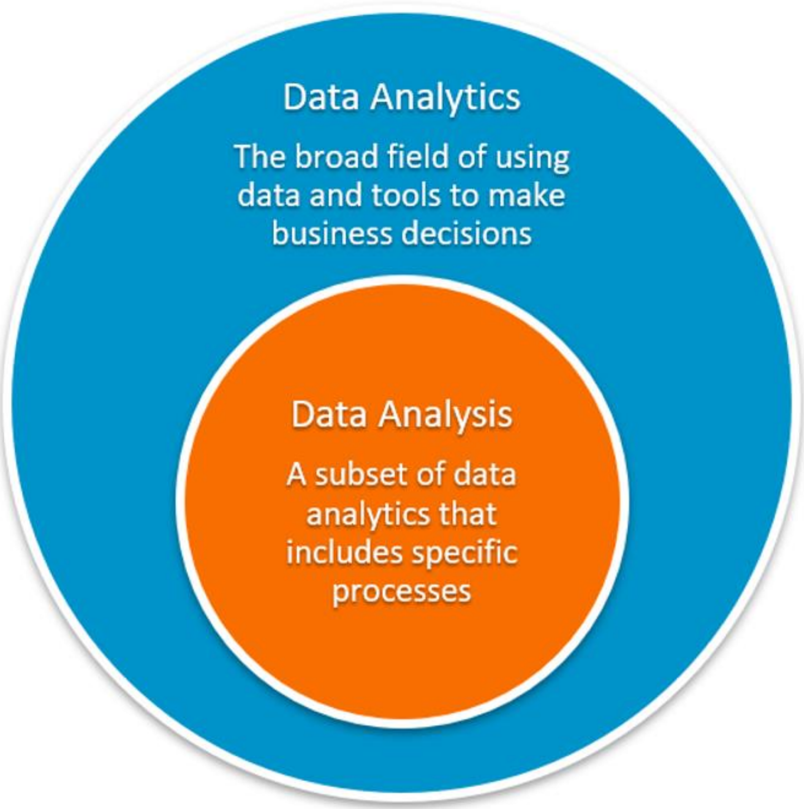
Data analytics and data analysis are closely related processes that involve extracting insights from data to make informed decisions, but there are some important differences between analysis and analytics [6].

Data analysis consists of cleaning, manipulating data, modeling, and questioning data to discover relevant information and gain **insights**, it is a vital part of data analytics.

Data analytics refers to a broad range of data-related activities and concepts. It is a process for translating basic facts and figures into specific **actions** by examining raw data assessments and perceptions in the context of organizational problem-solving and **decision-making.**

The purpose is to help businesses make better decisions and achieve greater success. Analytics uses data, machine learning, statistical analysis, and computer-based models to gain insight and make better decisions from collected data.

Analysis looks backwards, providing a historical view of what has happened. Analytics, on the other hand, models the future or predicts a result.



Data Analytics

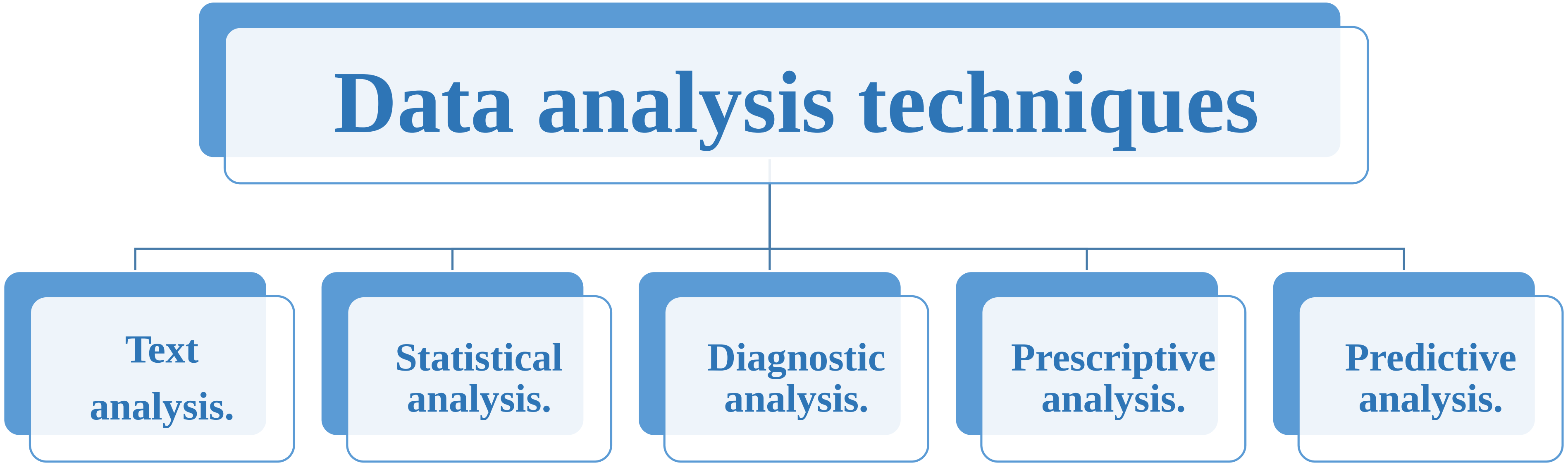
The broad field of using data and tools to make business decisions

Data Analysis

A subset of data analytics that includes specific processes

Data analysis techniques.

There are several data analysis techniques based on business and technology. The most common data analysis techniques are[7]:



Data analysis techniques

Text analysis.

Statistical analysis.

Diagnostic analysis.

Prescriptive analysis.

Predictive analysis.

Data analysis techniques cont.

Text analysis

Text analysis is the process of using computer systems to read and understand human-written text for business insights.

Text analysis **software** can classify, sort, and extract information from text to identify patterns, relationships, and other actionable knowledge.

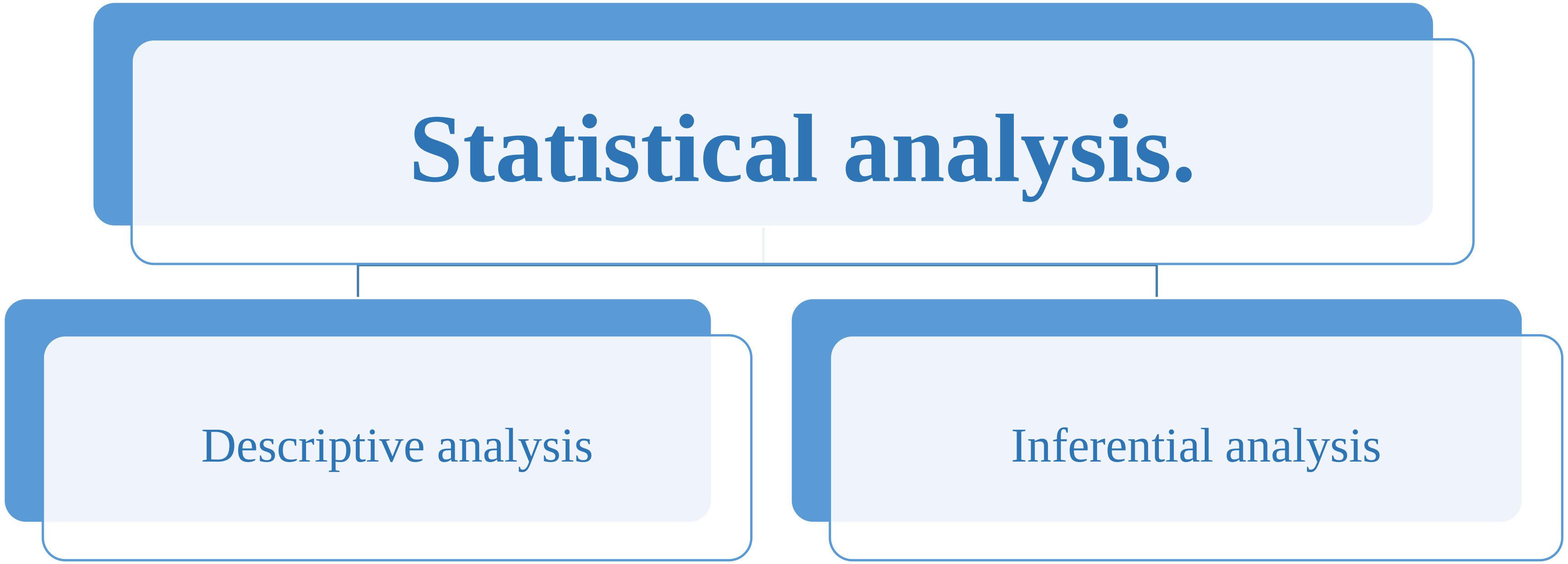
Text analysis is **important**, where Businesses use text analysis to extract actionable insights from various unstructured data sources as survey responses, emails, call center notes, product reviews, social media posts, product reviews, and any other feedback given in free text.

Data analysis techniques cont.

Statistical analysis

Statistical Analysis shows "What happened?" by using past data in the form of dashboards. Statistical analysis includes collection, analysis, interpretation, presentation, and modeling of data. **It analyses a set of data or a sample of data.**

There are two categories of Statistical analysis :



Statistical analysis.

Descriptive analysis

Inferential analysis

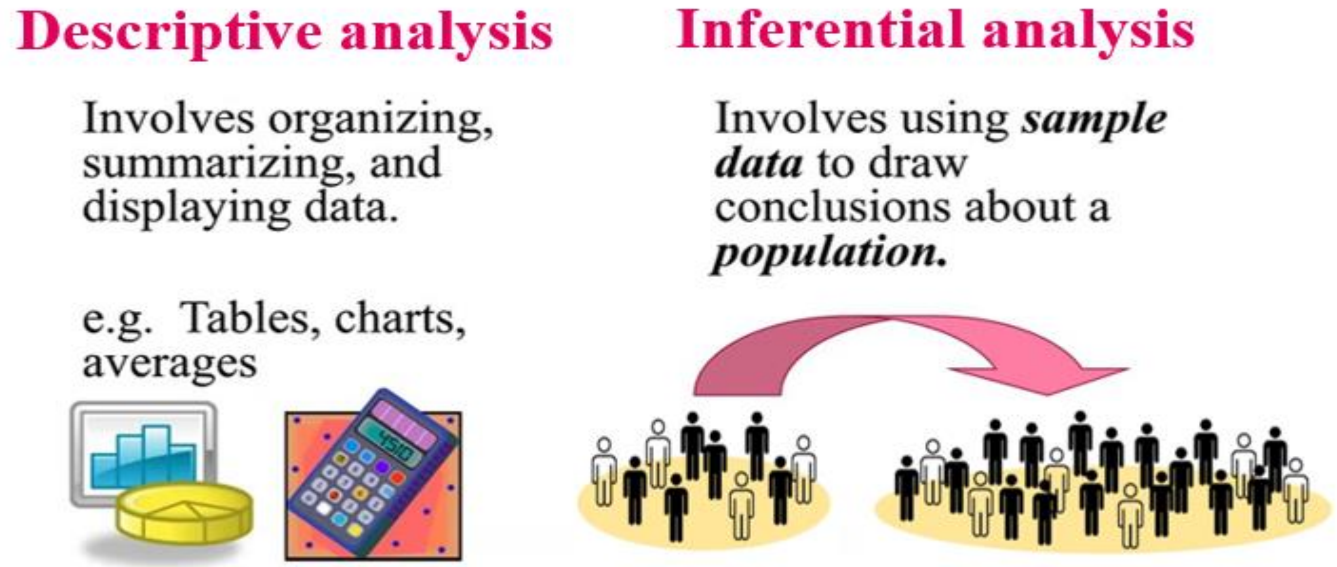
Statistical Analysis cont.

Descriptive analysis:

Descriptive analysis answers the question, “What happened?”, **it looks at past data to describe what has happened**. it analyses **complete data or a sample** of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data

Inferential analysis:

It analyses sample from complete data. It means; It uses a small sample to conclude a bigger population.



Descriptive analysis

Involves organizing, summarizing, and displaying data.

e.g. Tables, charts, averages

Inferential analysis

Involves using ***sample data*** to draw conclusions about a ***population.***

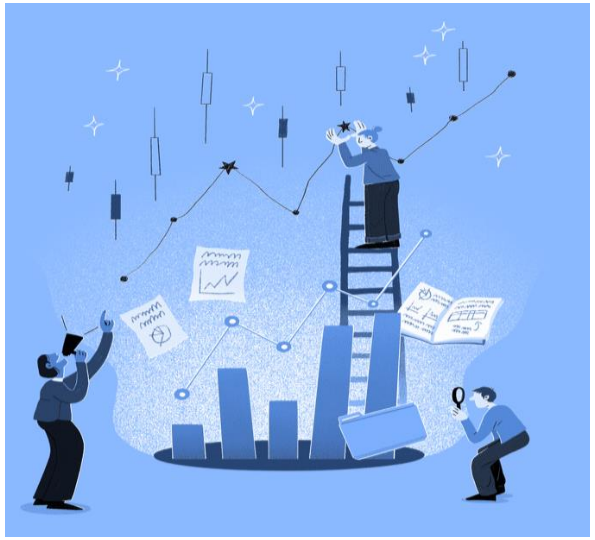
Data analysis techniques cont.

Diagnostic analysis

Diagnostic analysis answers the question, **“Why did this happen?”** by finding the cause from the insight found in statistical analysis.

This analysis is useful to identify behavior patterns of data. **If a new problem arrives in your business process**, then you can look into this analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

Predictive analysis



Predictive analysis answers the question, “**What might happen in the future?”**

It shows "what is likely to happen" by using previous data, it makes predictions about future outcomes based on current or past data.

Data analysis techniques cont.

Prescriptive analysis

Prescriptive analysis answers the question, **“What should we do next?”**, it recommends the best actions to take.

Prescriptive analysis **combines the insight from all previous analysis** to determine **which action to take in a current problem or decision**. Most data-driven companies are utilizing prescriptive analysis because predictive and descriptive analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

Data analysis techniques cont.

Descriptive analysis

Example:

- Looking at sales data to see how sales have changed over time.

Diagnostic analysis

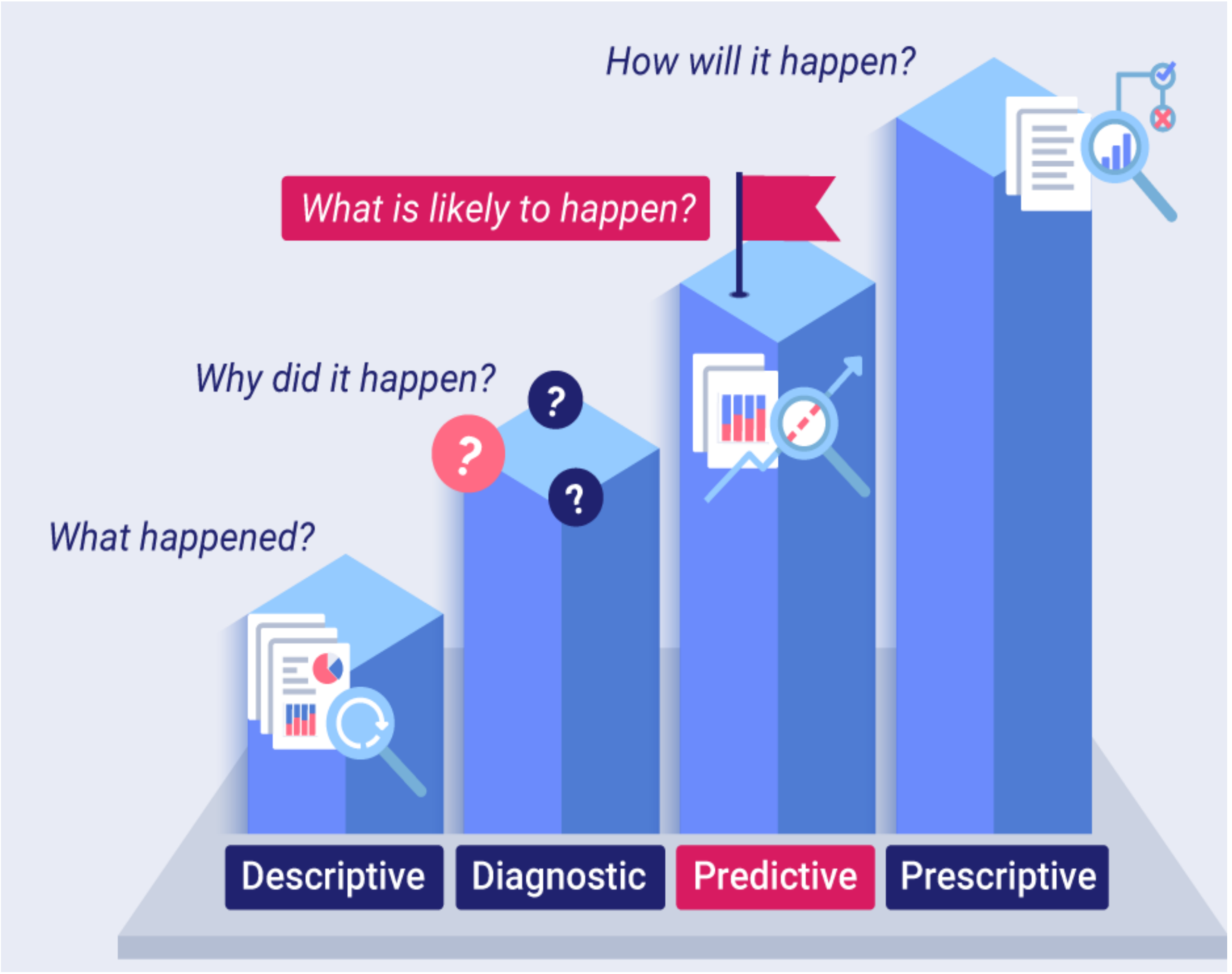
Example: Figuring out why sales dropped by looking at different factors.

Predictive analysis

Example: Predicting next month’s sales based on past trends.

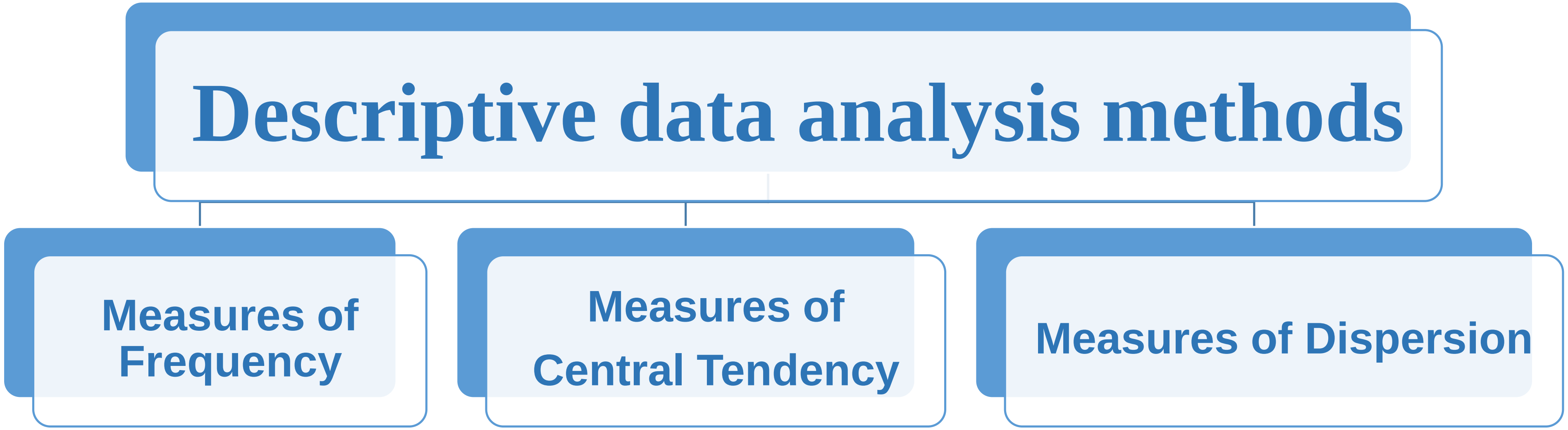
Prescriptive analysis

Example: Suggesting the best marketing strategy to use for the next campaign.



Descriptive data analysis methods.

The main descriptive data analysis methods are: measures of frequency, measures of central tendency, and measures of dispersion [6].



Descriptive data analysis methods

Measures of Frequency

Measures of Central Tendency

Measures of Dispersion

Descriptive data analysis methods cont.

Measures of Frequency.

The **frequency** of a value is the number of times that value appears in a data set. Measurements of Frequency show how many times each score occurs. The **main goal of frequency** measurements is to provide something like a **count or a percentage.**

Frequency table is a tabulation of data values that displays the number of times each value or group of values occurs in the dataset.

Frequency distribution tables can be used for both **categorical and numeric variables**. **Continuous variables** should only be used with class intervals.

Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a **relative frequency distribution**.

Frequency distributions can be **represented** by a histogram, or a pie chart.

Measurements of Frequency cont.

Example 1:

Here are the temperatures at midday for 7 days (in °C)

23, 24, 24, 23, 24, 25, 21

Represent the above data by the frequency table.

Temperature	Frequency
21	1
23	2
24	3
25	1
total	7

Measurements of Frequency cont.

Example 2:

The grades obtained in an English test by a class of 15 students are given below.

A, C, A, B, B, D, F, D, A, D, F, B, C, D, C

Represent the above-given grades in a **relative frequency table**.
Solution:

Grades	Frequency	Relative frequency
A	3	3/15*100=20%
B	3	3/15*100=20%
C	3	3/15*100=20%
D	4	4/15*100=26.67%
F	2	2/15*100=13.33%
	15	100%

Measurements of Frequency cont.

Example 3:

The following data represents the age of **25** employees .

36, 32, 48, 41, 38, 28, 37, 30, 58, 44, 34, 38, 43, 50, 40, 45, 55, 59, 29, 43, 32, 39, 48, 57, 46

Make a frequency distribution table.

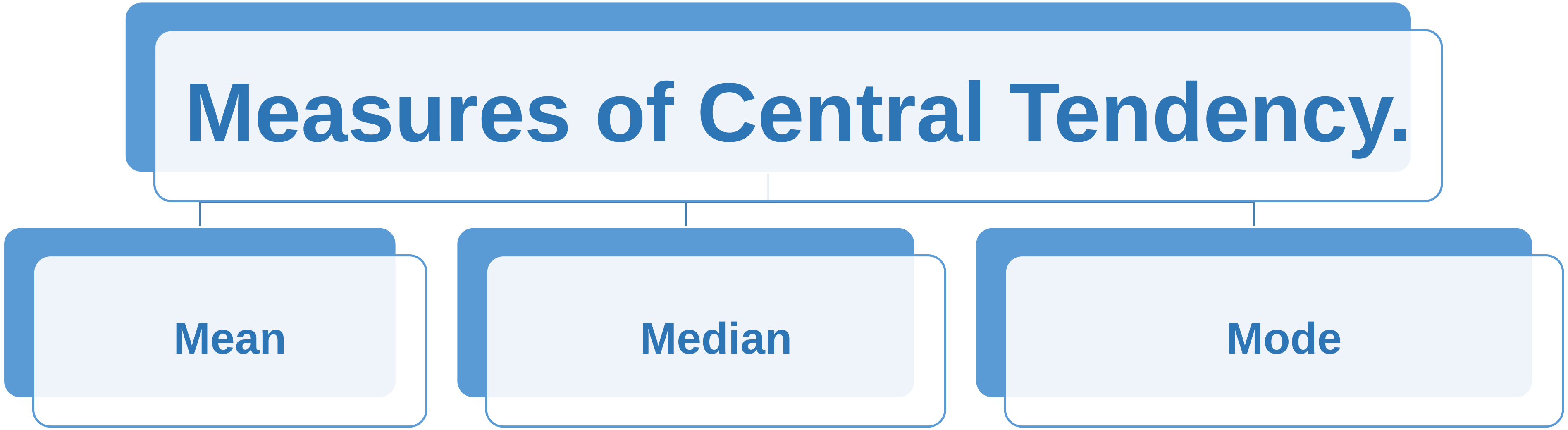
Solution:

Age interval	Frequency
20-30	2
30-40	9
40-50	9
50-60	5
	25

Measures of Central Tendency.

Finding the central tendency is crucial in descriptive analysis.

Three standards: mean (average), median, and mode are used to calculate central tendency.



Measures of Central Tendency.

Mean

Median

Mode

Measures of Central Tendency cont.

Mean.

The mean is commonly known as average, if there are n number of values in a dataset and the values are x_1, x_2, ..., x_n , then the mean is calculated as:

$$\boldsymbol{\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots + x_n}{n}}$$

Example 1:

Compute the mean for the following data:

5, 6, 2, 4, 7, 8, 3, 5, 6, 6

Solution:

Mean$= \frac{\mathbf{5+6+2+4+7+8+3+5+6+6}}{\mathbf{10}} = \frac{\mathbf{52}}{\mathbf{10}}$=5.2

Mean drawback.

Mean is susceptible to the influence of outliers. Also, mean is only meaningful if the data is normally distributed, or at least close to looking like a normal distribution.

Measures of Central Tendency cont.

Mean cont.

Mean drawback.

Mean is susceptible to the influence of outliers, outliers have a significant impact on the mean, as they can skew it towards their extreme values.

Example 2

Here the data for salaries of **10** individuals in month, Compute the mean :

8.000, 8.700, 10.500, 8.000, 9.000, 10.000, 8.500, 10.200, 7.900, **150.000**

Solution:

$$\textbf{Mean}=\frac{\mathbf{8000+8700+10500+8000+9000+10000+8500+10200+7900+150{,}000}}{\mathbf{10}}=\frac{\mathbf{230{,}800}}{\mathbf{10}}=\mathbf{23.800}$$

Measures of Central Tendency cont.

Median.

The median is the middle score for a dataset that has been sorted according to the values of the data, it is not affected by the outliers.

Odd Number of Observations

If the total number of observations given is odd, then the formula to calculate the median is:

Median =$\left(\frac{n+1}{2}\right)^{th}$ $term$, n is , where n is the number of observations.

Even Number of Observations

If the total number of observation is even, then the median formula is:

Median =$\frac{\left(\frac{n}{2}\right)^{th} term+\left(\frac{n}{2}+1\right)^{th} term}{2}$, where n is the number of observations.

Median cont.

Example 1:

Find the median of the following:

4, 17, 77, 25, 22, 23, 92, 82, 40, 24, 14, 12, 67, 23, 29

Solution:

- Order the numbers ascending, or descending, we have:

4, 12, 14, 17, 22, 23, 23, **24**, 25, 29, 40, 67, 77, 82, 92,

- There are 15 numbers. Our middle is the term no (15+1)/2 , i. e, the eighth term, the median value of this set of numbers is 24.

Example 2:

Find the median of the following:

3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29

Solution

- Order the numbers ascending, or descending, we have:

3, 5, 7, 12, 13, 14, **21, 23**, 23, 23, 23, 29, 40, 56

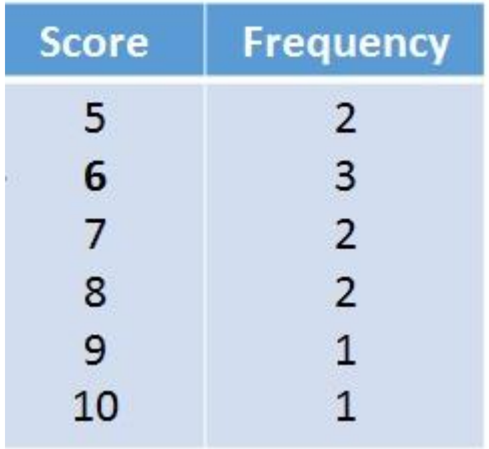
- There are 14 numbers, our middle is the average of the two terms no (14)/2=7, (14)/2+1=8 , i. e, the median value of this set of numbers is (21+23) /2=22.

Measures of Central Tendency cont.

Mode.

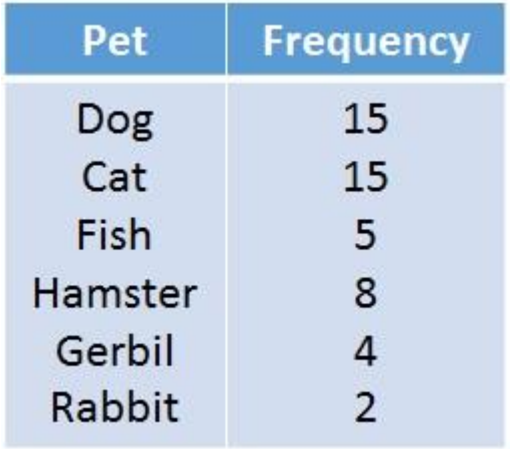
The mode is the most frequently occurring value in a dataset. On a histogram representation, the highest bar denotes the mode of the data. For example the mode for the data 3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48 is 16

.

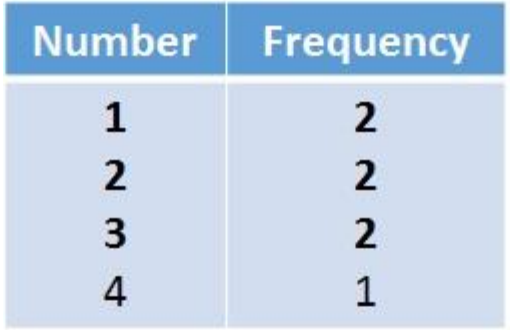


Score	Frequency
5	2
6	3
7	2
8	2
9	1
10	1

Uni mode

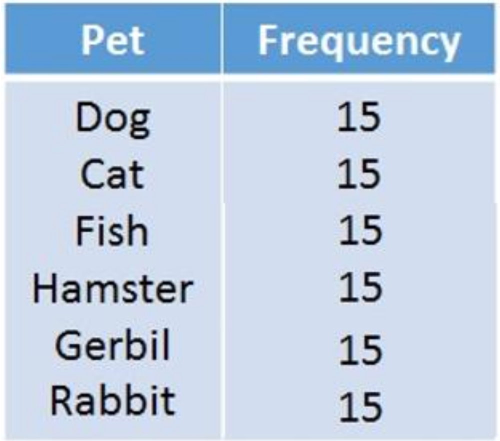


Pet	Frequency
Dog	15
Cat	15
Fish	5
Hamster	8
Gerbil	4
Rabbit	2



Number	Frequency
1	**2**
2	**2**
3	**2**
4	1

Multi mode



Pet	Frequency
Dog	15
Cat	15
Fish	15
Hamster	15
Gerbil	15
Rabbit	15

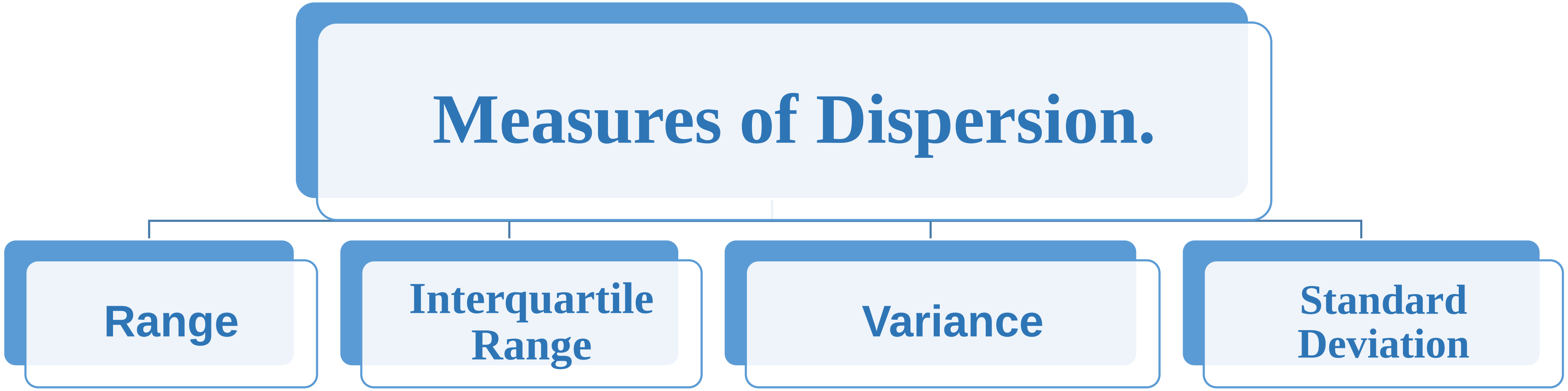
No mode

Descriptive data analysis methods cont.

Measures of Dispersion.

Looking at a central point (mean, median, or mode) may not help in understanding the actual shape of a distribution. Therefore, we often look at the spread, or the dispersion, of a distribution.

The most common measures of dispersion is:



Measures of Dispersion cont.

Range

The range is the difference between largest observation and smallest observation

Range = largest observation- smallest observation

The main **disadvantage** of the range is that it is affected by **extremes or outliers**, resulting in an inaccurate image of the most likely range.

For example for the data 3, 6, 9,16, 27, 27, 37, 48

The range is 48-3=45

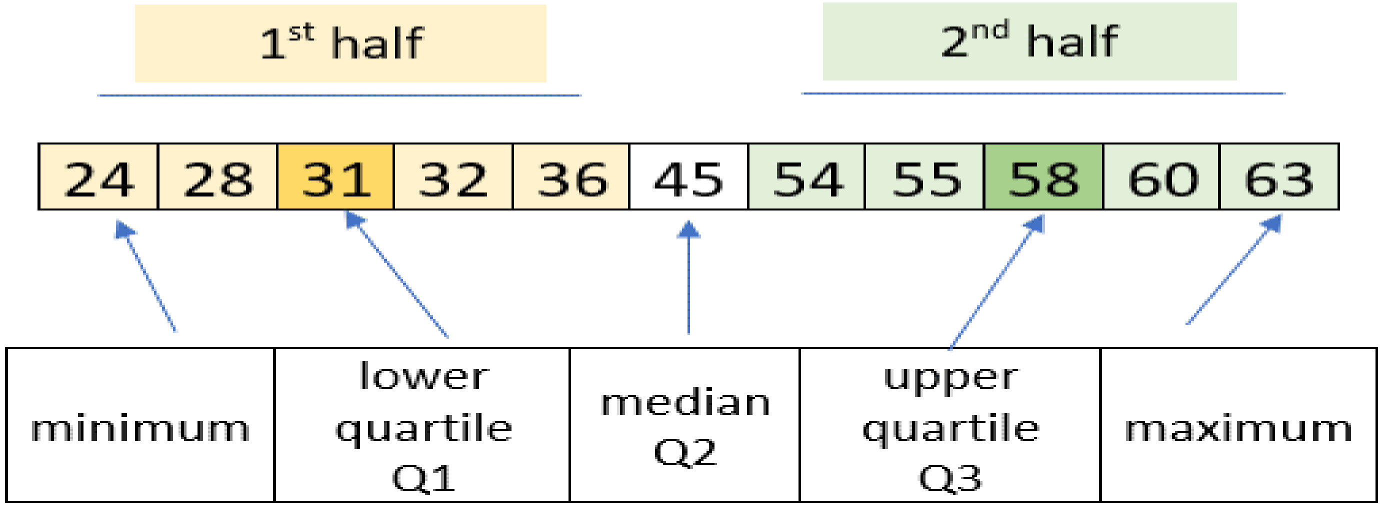
Measures of Dispersion cont.

Interquartile Range.

One way around the range’s disadvantage is to calculate it after removing extreme values. One convention is to cut off the top and bottom one-quarter of the data and calculate the range of the remaining middle 50% of the scores.

Interquartile range = upper quartile − lower quartile,

Where The lower quartile value is the median of the lower half of the data, The upper quartile value is the median of the upper half of the data, and the data should be arranged ascendingly.



1st half

2nd half

24	28	31	32	36	45	54	55	58	60	63

minimum	lower quartile Q1	median Q2	upper quartile Q3	maximum

Interquartile Range cont.

Example.1

Find the interquartile range for the following set of data.

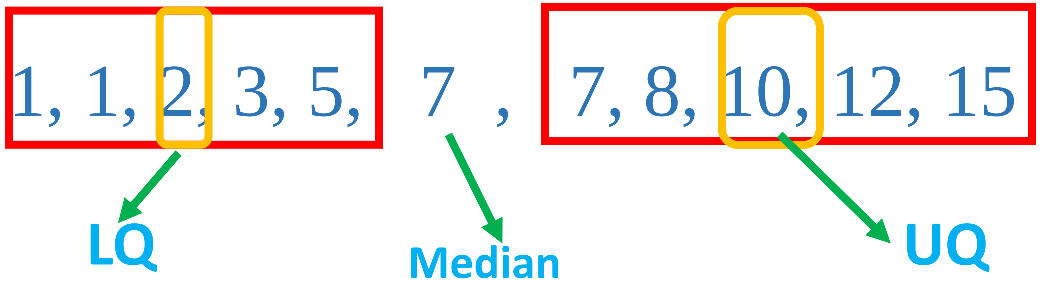
1,10,2,3,15,7,7,8,1,12,5

Solution

Arrange the data ascendingly

1, 1, 2, 3, 5, 7, 7, 8, 10, 12, 15

Divide the data to 2 halves



1, 1, 2, 3, 5, 7 , 7, 8, 10, 12, 15

LQ

Median

UQ

LQ=2
UQ=10
Interquartile Range=10-2=8

Interquartile Range cont.

Example.2

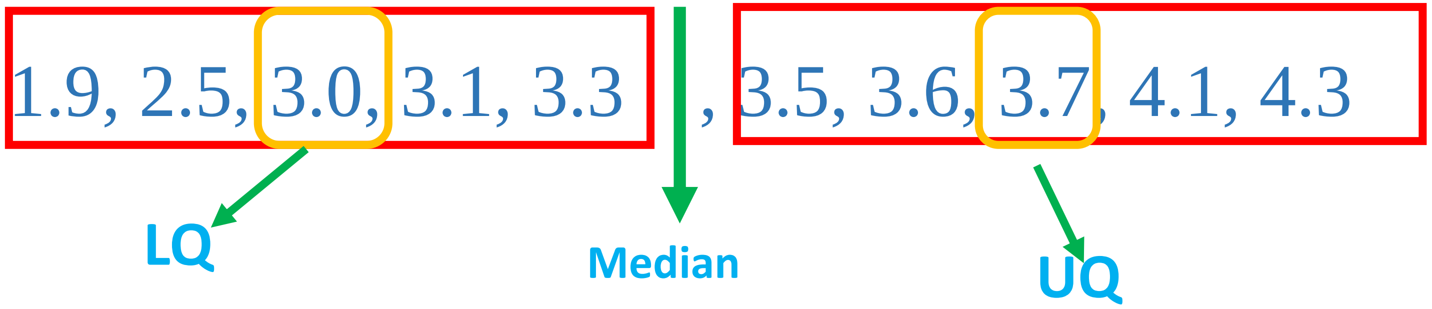
Find the interquartile range for the following set of data.
3.3, 3.7, 2.5, 3.5, 3.0, 4.3, 3.1, 4.1, 1.9, 3.6

Solution

Arrange the data ascendingly

1.9, 2.5, 3.0, 3.1, 3.3, 3.5, 3.6, 3.7, 4.1, 4.3

Divide the data to 2 halves



1.9, 2.5, 3.0, 3.1, 3.3 , 3.5, 3.6, 3.7, 4.1, 4.3

LQ

Median

UQ

LQ=3.0
UQ=3.7
Interquartile Range= 3.7-3.0= 0.7

Measures of Dispersion cont.

Variance.

The variance is a measure used to indicate how spread out the data points are. To measure the variance, the common method is to pick a center of the distribution, typically the mean, then measure how far each data point is from the center.

The **variance of the population** is defined by the following formula:

$$\sigma^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n}$$

Where $\boldsymbol{\bar{x}}$is the population mean, x_i is the ith element from the population, and n is the number of elements in the population.

The **variance of a sample** is defined by a slightly different formula:

$$S^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n-1}$$

Where $\boldsymbol{\bar{x}}$ is the sample mean, x_i is the ith element from the sample, and n is the number of elements in the sample.

Measures of Dispersion cont.

Standard Deviation.

There is one issue with the variance as a measure. It gives us the measure of spread in units squared. So, for example, if we measure the variance of age (measured in years) of all the students in a class, the measure we will get will be in years squared. However, practically, it would make more sense if we got the measure in years (not years squared). For this reason, we often take the square root of the variance, This measure is known as the standard deviation.

The formula to compute the standard deviation of the population is:

$$\sigma = \sqrt{\frac{\sum(x_i - \overline{\boldsymbol{x}})^2}{n}}$$

The formula to compute the standard deviation of the sample is:

$$S = \sqrt{\frac{\sum(x_i - \overline{\boldsymbol{x}})^2}{n-1}}$$

Assignments # 1

1- Give an example of data, information and knowledge?

2- Mention the difference between a data model and a schema.