Statistics Essentials for Data Science

Understanding the Data

Learning Objectives

By the end of this lesson, you will be able to:

- Explore different types of data
- Categorize data from a statistical perspective
- Distinguish among raw data, processed data, and primary and secondary data

Learning Objectives

By the end of this lesson, you will be able to:

- Identify the difference between structured and unstructured data
- Understand the importance of data quality
- Differentiate among cross-sectional, time series, pooled, and panel data

Business Scenario

ABC is a healthcare organization managing hospitals, colleges, clinics, and medicines. It generates a vast amount of data daily, encompassing patients, medicines, and students.

However, ABC encounters issues due to improper data management and ineffective use of information. The organization aims to address those issues to better understand their customers and leverage this knowledge to improve services and boost revenue. For instance, timely delivery of a patient's medicines will not only elevate the customer experience but also increase the organization's revenue.

To achieve this goal, ABC will explore various data types and their storage methods.

Types of Data in Business Contexts

Discussion

Data Used in Businesses

Duration: 15 minutes

How can data analysis assist in growing a business?

- What is data?
- What are the types of data?

Data Used in Businesses

Different types of data are gathered and used in a business.

Identifying the right type of data is crucial for choosing the optimal statistical analysis that will yield the best results.

Data

Data consists of facts and figures collected, analyzed, and summarized for presentation and interpretation.

- It is the foundational information used to produce statistics.
- The data collected or compiled for a statistical investigation is referred to as a dataset.

Dataset

Example: A five-star hotel analyzes its business by incorporating relevant data from its operations at numerous locations.

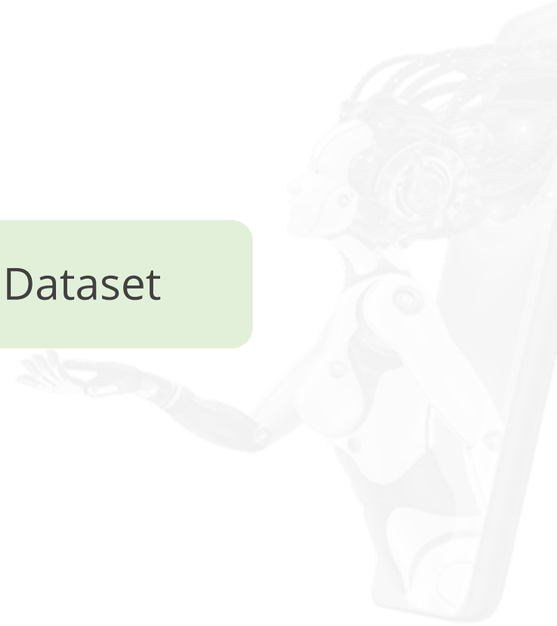
Dataset

The information collected across the locations constitutes the dataset.

Data for analysis can be:

- Income from room occupancy
- Income from the coffee shop
- Unused rooms during the offseason
- Turned down customer requests due to reaching capacity

Dataset



Elements

The entities on which the data is collected are referred to as elements.

Element

Hotel location

Observations

Observations are a set of measurements for an element.

Observations

Room occupancy

Observations are obtained by measurements or a physical count.

Data Categorization and Types of Data

Types of Data

Different types of data can be collected and analyzed in different situations.

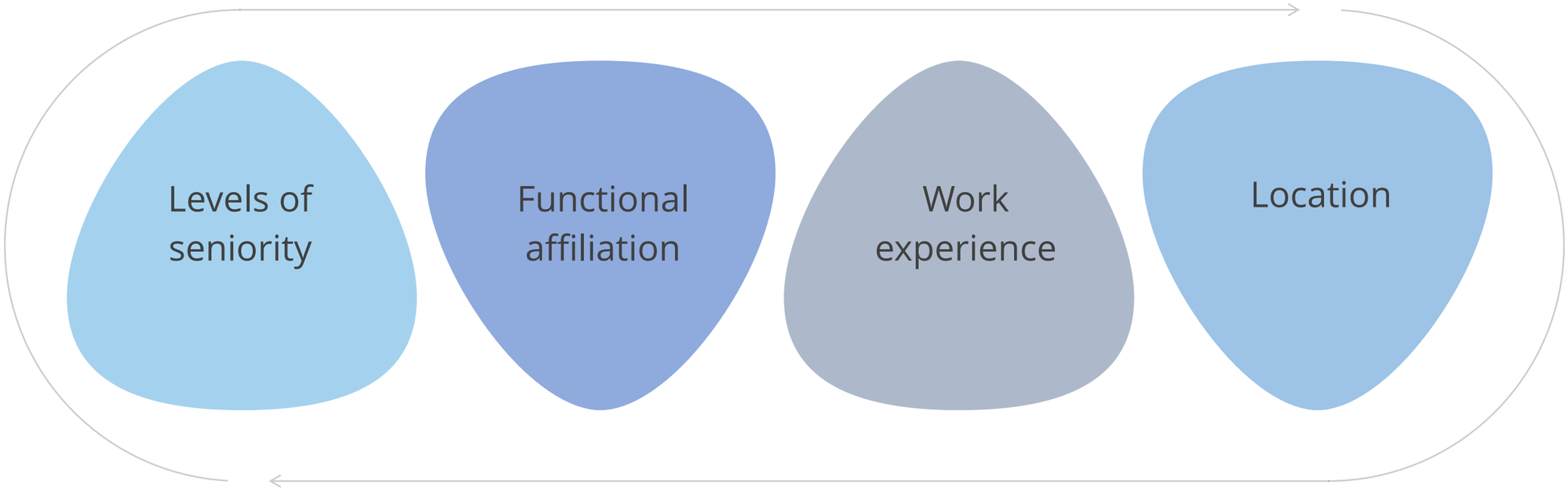
- Data needs to be studied and analyzed statistically to gain useful information.
- There are several ways to categorize data.

The appropriate data must be used based on the context of the study.

Data Classification: An Example

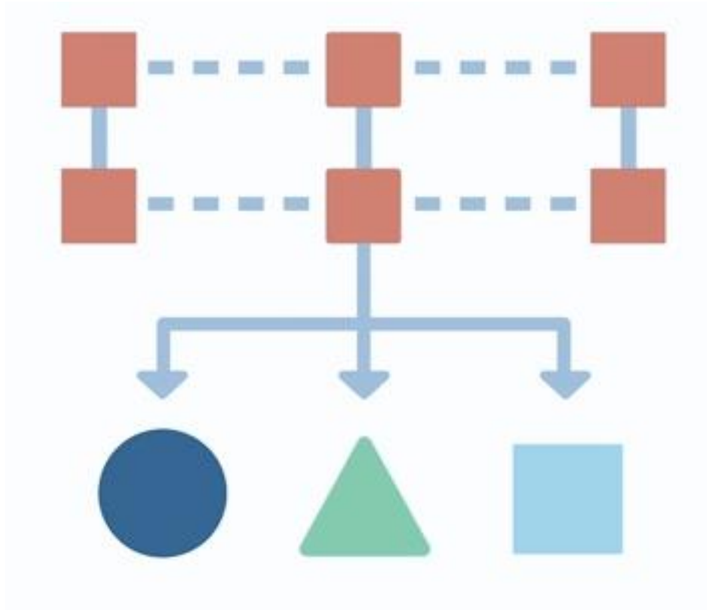
Data is collected during an employee satisfaction survey.

The responses collected from respondents can be analyzed by categorizing them into:



Importance of Data Classification

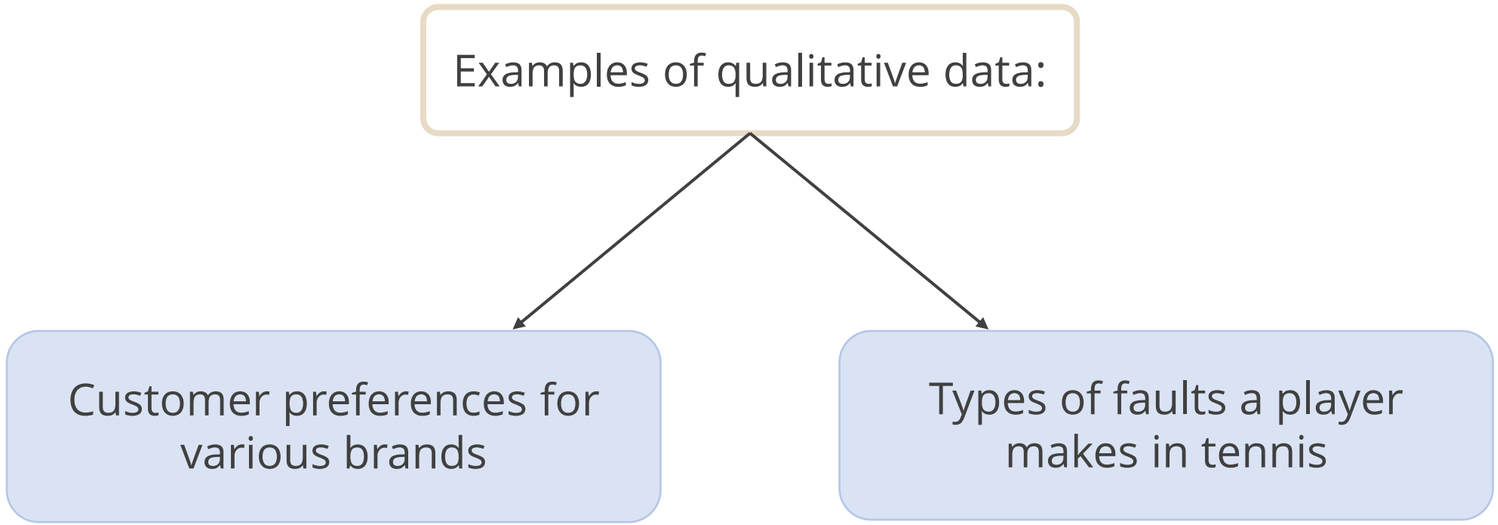
Categorization is invariably important to ascertain possible differences among various categories.



Statistical data can be broadly categorized into qualitative and quantitative data.

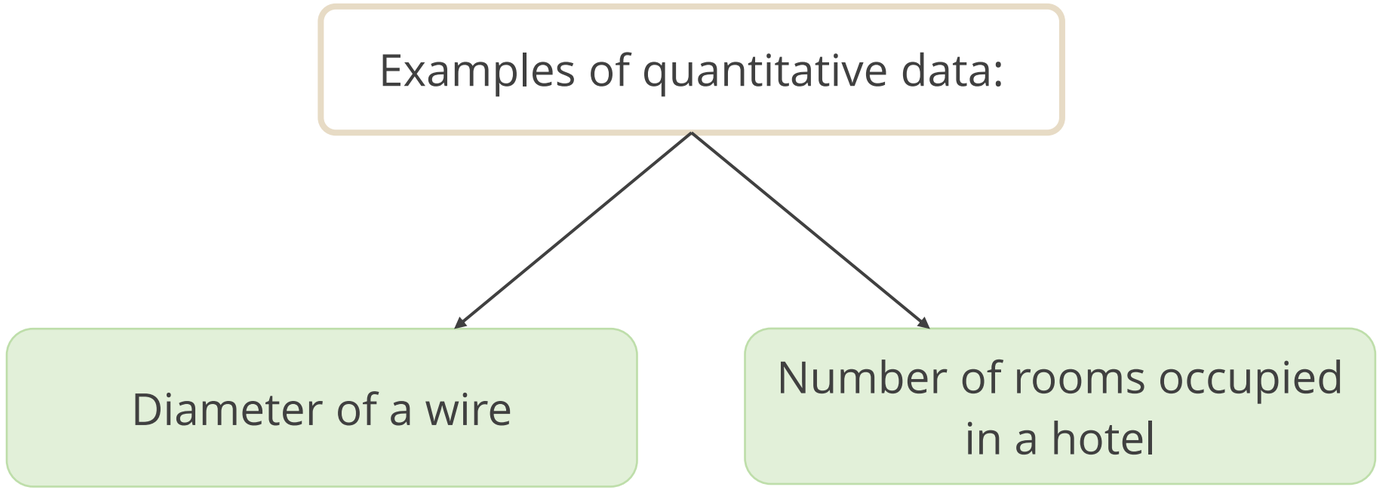
Qualitative, Attributed, or Categorical Data

When data cannot be quantified, it is expressed descriptively. This is called qualitative data.



Quantitative, Variable, or Measurable Data

Data that is specified numerically through a process of measurement or numerical count is called quantitative data.



Determining the Type of Data to Be Used

Examples:

Diameter of wires

- Inspected using a go or no-go gauge
- Categorized as within or outside specification limits
- Qualitative data is directly collected
- Data collection process is simplified

Scores of students

- Classified into grades, such as Grade A, Grade B, Grade C, and Grade F
- Quantitative data is collected initially
- Data collected is categorized into classes

Data Used in Businesses

Duration: 15 minutes

How can data analysis assist in growing a business?

- What is data?

Answer: Data is facts and figures that are collected, analyzed, and summarized for presentation and interpretation. The data collected or compiled for a statistical investigation is referred to as a dataset.

- What are the types of data?

Answer: The types of data include:

  - Qualitative, attributed, or categorical data
  - Quantitative, variable, or measurable data

Types of Data Collection

Discussion

Types of Data Collection

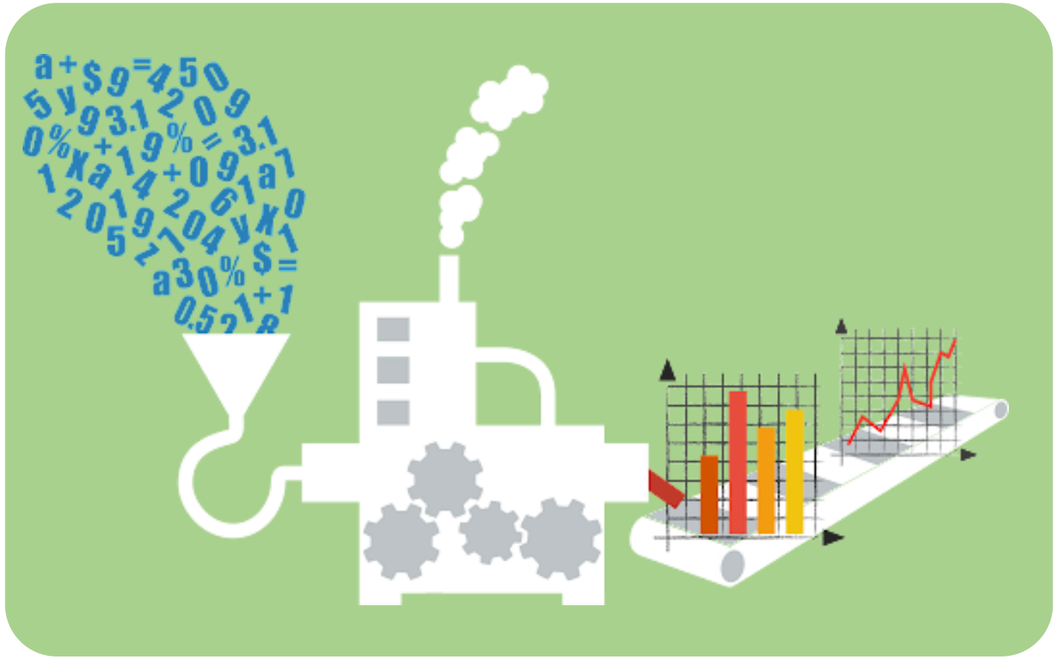
Duration: 15 minutes

You are working in a social media company where the users post photos, videos, and text. You have been asked to collect and store this data in a database.

What type of data is being collected here?

Raw Data and Processed Data

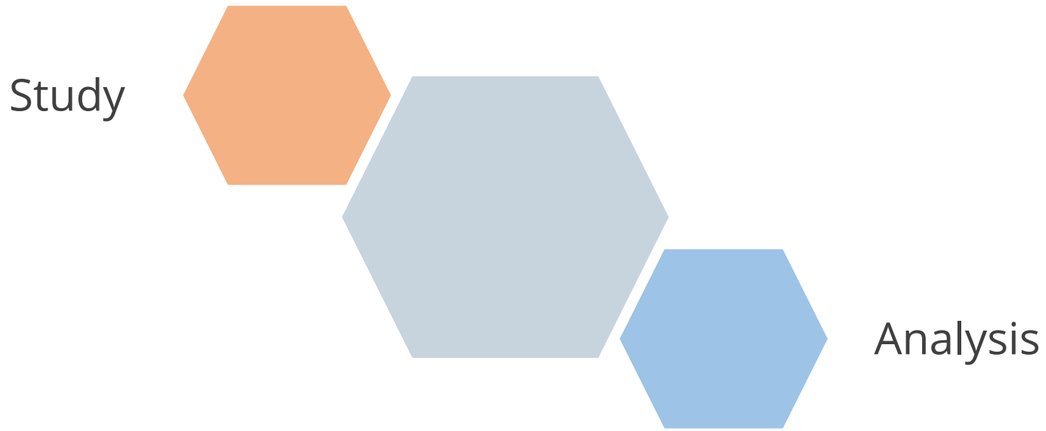
Raw data and processed data are commonly used every day.



Raw Data and Processed Data

Data is recorded in its raw form, compiled, and presented in a way that enables analysts to derive insights.

Raw data refers to data collected and noted for reference in a record. Such data is typically not amenable to direct a:



Raw Data and Processed Data

When data is systematically presented in a form that enables one to draw insights, it is referred to as processed data.

Correct data processing is required to avoid having a negative impact on the final data output.

Example for Raw Data

Consider an example of a record detailing students' scores in various subjects, along with their total scores and grades:

	Math	Science	English	Total score	Grade
Claire	77	82	89	248/300	B
Matthew	87	91	95	273/300	A
Ryan	70	74	68	212/300	C
Sarah	90	94	97	281/300	A
Will	81	82	75	238/300	B

Such data constitutes raw data.

Example for Processed Data

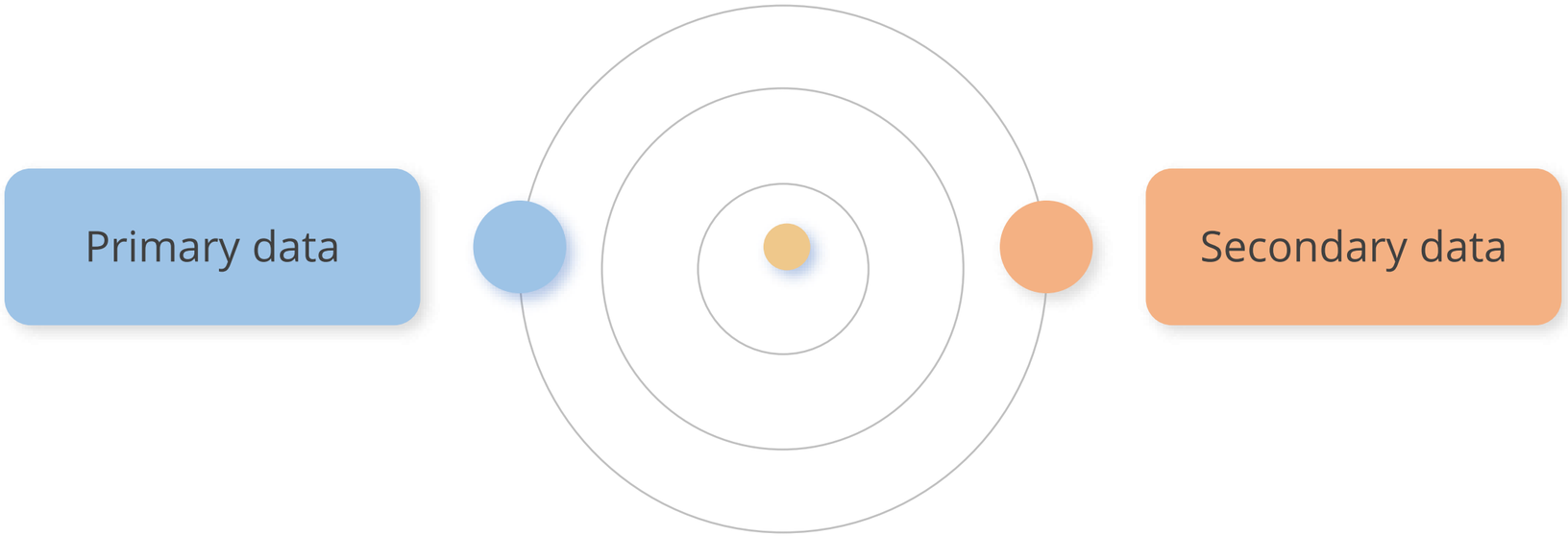
A table that indicates the number of students securing Grade A, Grade B, and Grade C can be obtained, and this constitutes processed data.

Grade A	Grade B	Grade C
2	2	1

The record briefly provides an idea about the overall performance of the class.

Data Collection

Sources of statistical data can be broadly categorized into:



Data collection plays an important role in statistical analysis.

Primary Data

Primary data refers specifically to the data related to the problem under investigation.

The United States Census Bureau collects, analyzes, and provides data about the country’s people and economy. This example illustrates the use of primary data.

Secondary Data

Data that someone else or another company has already collected and analyzed is referred to as secondary data.

Secondary data is often used to supplement primary data.

It can be found in a variety of sources, such as government publications, academic journals, and commercial databases.

Raw Data and Processed Data

Duration: 15 minutes

You are working in a social media company, where the users post photos, videos, and text. You have been asked to collect and store this data in a database.

What type of data is being collected here?

Answer: Raw data is being collected here. Data is usually recorded in raw form. Such data is then compiled and presented in a way that enables viewers to draw insights.

Types of Data

Discussion

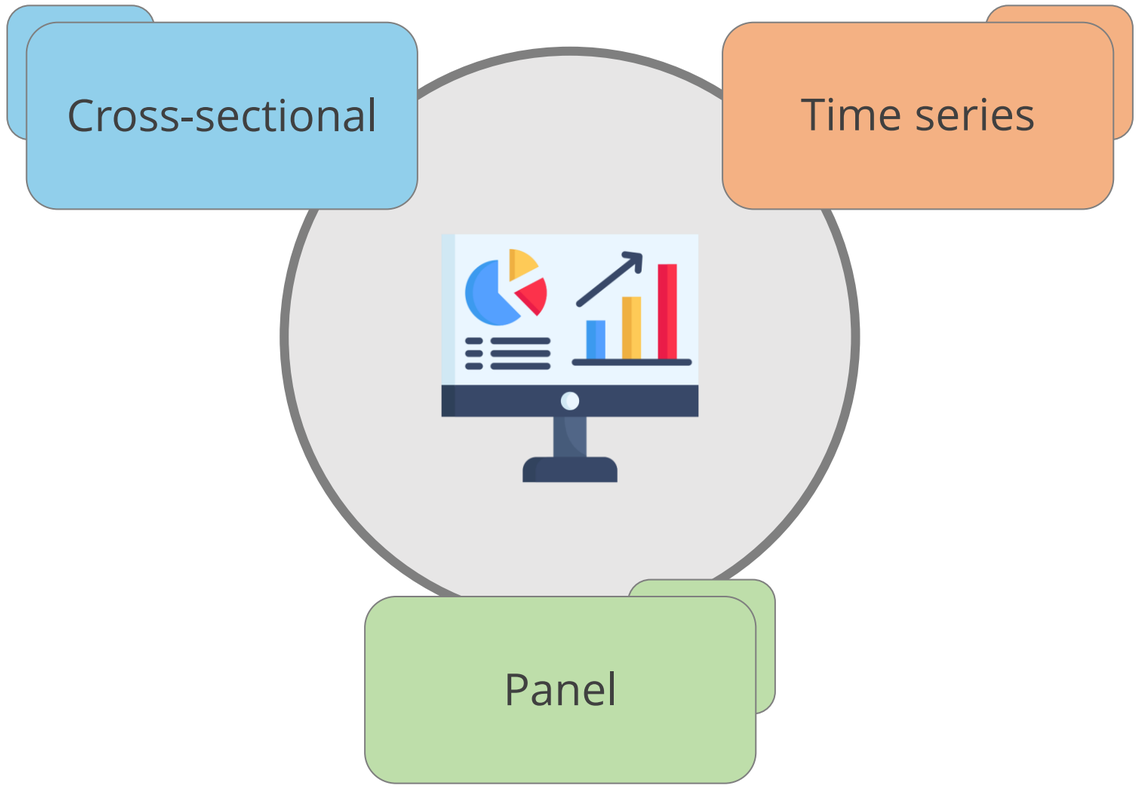
Types of Data

Duration: 15 minutes

You are assessing data from a specific location containing personal and financial information, like the spending habits of families in a large city.

What type of data is needed for data analysis?

Types of Data



Cross-Sectional Data

Cross-sectional data pertains to data collected by observing numerous subjects at a specific point in time, such as:

Individuals

Families

Localities in a city

Machines in a factory

Example: Studying Expenditure Patterns

Data from a sample of families in a large city can be collected to study the expenditure patterns of families in that city.

The sample is a cross-section of the population from several localities.

In this sample, the age profiles and incomes of the different units will vary.

Time Series Data

Time series data refers to data collected over a period of time.

Example: Data on the annual sales of a specific commodity over several years

Panel or Longitudinal Data

Panel data, also known as longitudinal data, encompasses data collected over time from cross-sectional units.

When the yearly sales of a commodity over several years are collected from numerous retail outlets, the data would constitute cross-sectional time series data or panel data.

Different Types of Data

Duration: 15 minutes

You are assessing data from a specific location containing personal and financial information, like the spending habits of families in a large city.

What type of data are you analyzing?

Answer: This is the cross-sectional data for a certain time period. Cross-sectional data refers to data collected by observing numerous subjects such as individuals, families, factory machines, or city localities over a specific time period.

Structured vs. Unstructured Data

Discussion

Structured vs. Unstructured Data

Duration: 15 minutes

What are the other types of data?

- Give an example of structured data
- Give an example of unstructured data

Data Classification

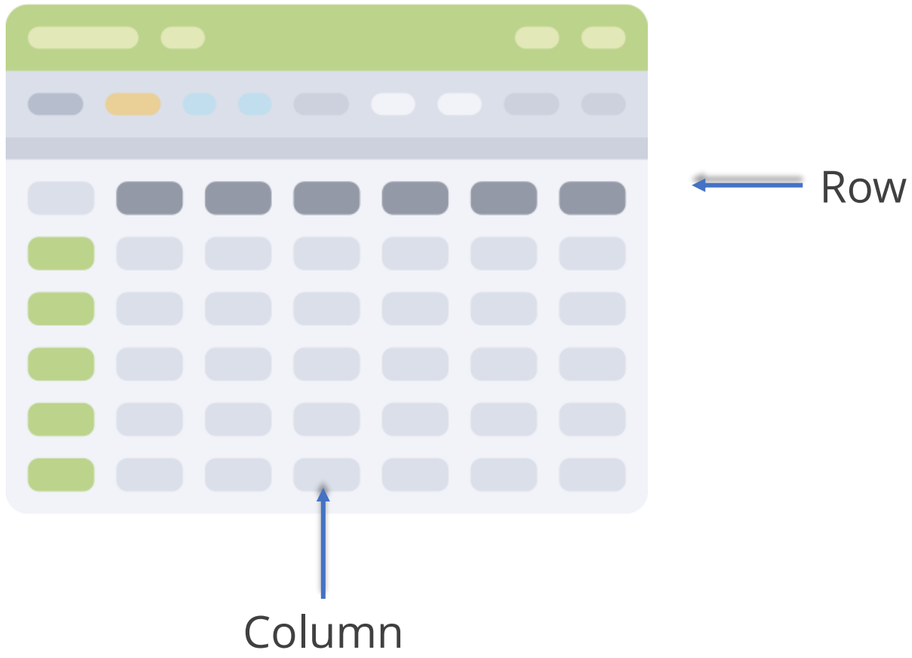
Data can also be categorized as:

Structured data

Unstructured data

Structured Data

Data recorded or stored in a predefined format for easy search and use during analysis is referred to as structured data.



Example: Data entered in Excel becomes structured when specific rows and columns correlate to certain characteristics.

Example of Structured Data

This is a material procurement plan for a certain component for one year.

Month	Opening Stock	Quantity Procured	Material Consumed	End Stock
1	120	700	80	680
2		0	100	
3		0	100	
4		0	100	
5		500	100	
6		0	100	
7		0	100	
8		0	100	
9		0	100	
10		0	100	
11		0	100	
12		0	100	
Total		1200	1180	

This data is in a structured format, and using Excel, the missing values can be obtained.

Example of Structured Data

In this case, the ending stock for the first month becomes the opening stock for the second month.

Month	Opening Stock	Quantity Procured	Material Consumed	End Stock
1	120	700	80	680
2	680	0	100	
3	580	0	100	
4	480	0	100	
5	880	500	100	
6	780	0	100	
7	680	0	100	
8	580	0	100	
9	480	0	100	
10	380	0	100	
11	280	0	100	
12	180	0	100	
Total		1200	1180	

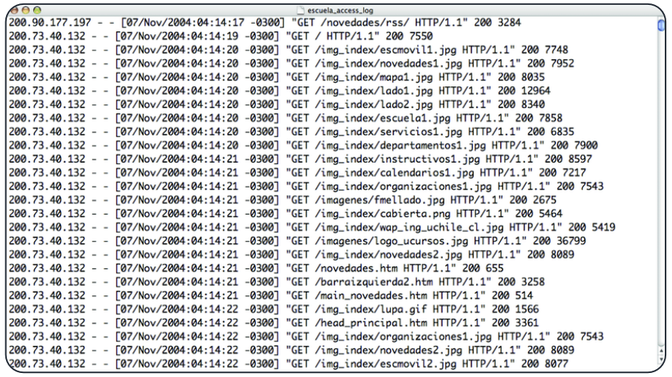
Unstructured Data

When data is not recorded in a predefined format, it is referred to as unstructured data.

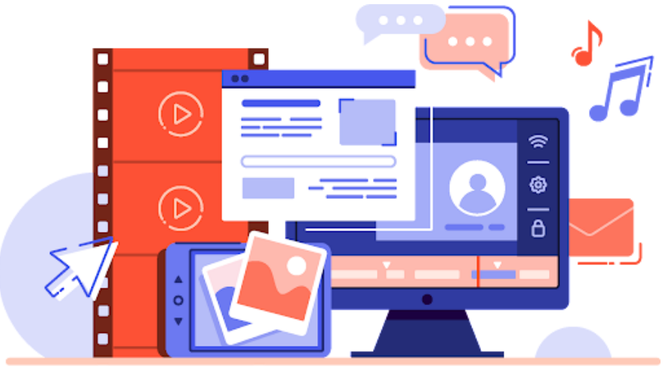
The data does not follow a data model and has no easily identifiable structure, making it difficult for computer programs to use.

Examples of Unstructured Data

The data given below can be stated in an unstructured format:



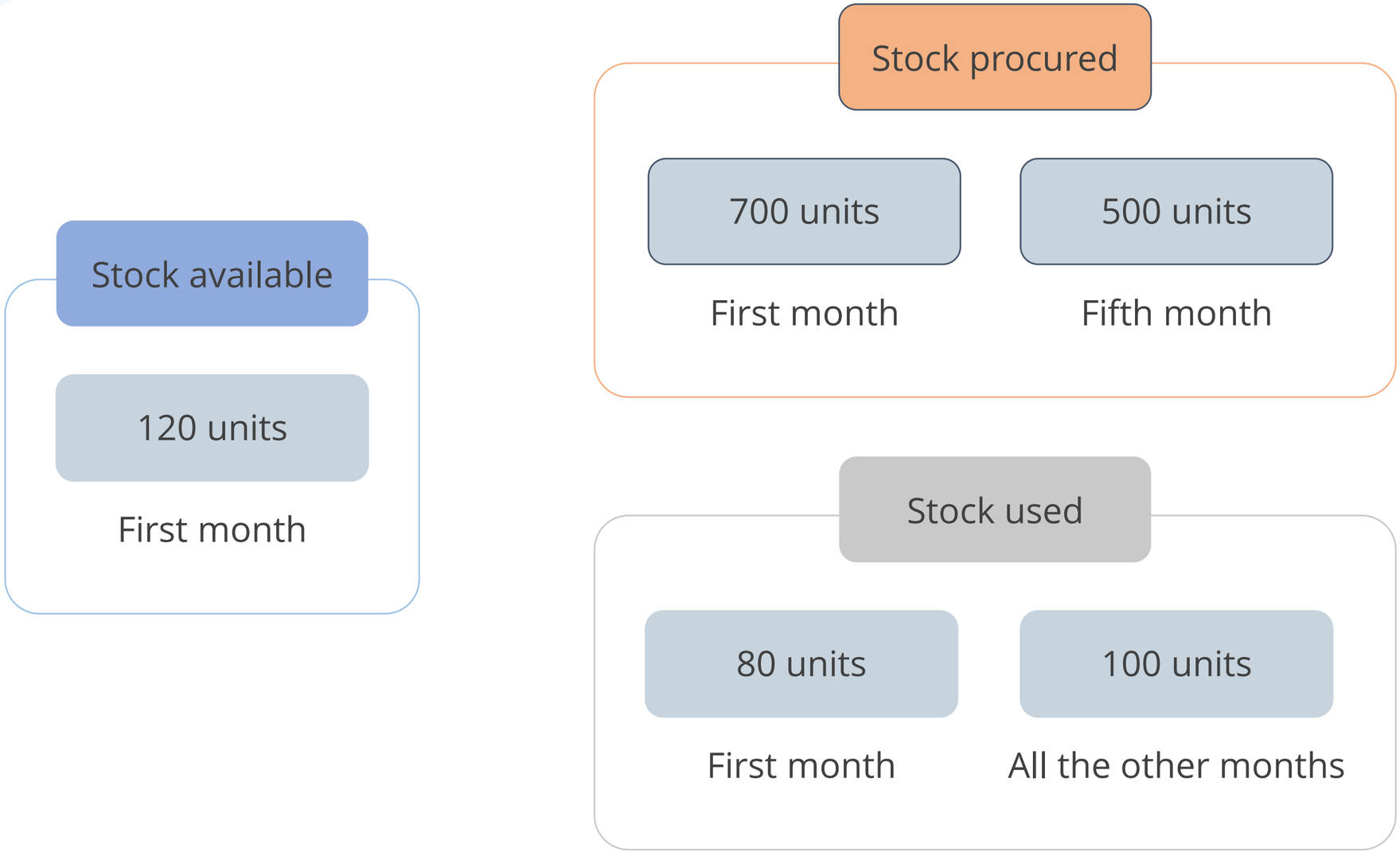
Web logs



Multimedia content

Raw text files

Example of Unstructured Data



Structured vs. Unstructured Data

Structured data

- In structured data, statistical measures such as mean, median, mode, and standard deviation can be directly calculated.
- Structured data facilitates easy searching and objective analysis.

Unstructured data

- In unstructured data, statistical techniques may be applied indirectly, such as analyzing aggregated metrics or patterns.
- Despite its complexity, unstructured data can still yield useful information.

Therefore, both types of data are useful.

Types of Data

Duration: 15 minutes

What are the different types of data?

- Give an example of structured data.

Answer: Data recorded or stored in a predefined format for easy search and use during analysis is referred to as structured data. An example of this is the collection of stock at the end of a month.

- Give an example of unstructured data.

Answer: When data is not recorded in a predefined format, it is referred to as unstructured data. An example of this is document collection through invoices, records, emails, and productivity applications.

Key Sources of Data

Discussion

Key Sources of Data

Duration: 15 minutes

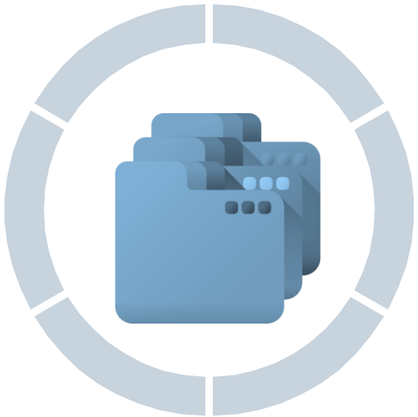
You are planning to gather and analyze data for your study, but you have not developed any specific tools or applications for data collection yet.

What are the methods used to collect data?

Key Sources of Data

The following are the primary sources of data:

- Data collected and published by government, industries, or individual bodies
- Conducting experiments
- Observational studies
- Conducting surveys



Healthcare

The use of electronic health records by hospitals has facilitated the sharing of knowledge on cost and quality measures, in addition to clinical data.

- The Patient-Centered Outcomes Research Institute uses such data for extensive research.
- Statistical analysis of data facilitates the use of evidence-based medicine.

Conduction of Experiments

Agricultural scientists evaluate the relative effectiveness of various seed varieties through carefully planned experiments.

Experimental design is the branch of statistics concerned with the planning and evaluation of experiments.

Surveys

A survey is a kind of observational study that collects data by questioning participants.

For instance, a survey on educational reform might aim to record the opinions of teachers and students regarding the current status and proposed reforms

Observational Studies

Observational studies involve statistical analyses of a population group without any research intervention or treatment.

Example: Researchers observe the behavior of animals, without interacting with them.

The knowledge gathered is used in research studies aimed at addressing behavioral issues and acting accordingly.

Observational Studies

Similarly, in organizations, staff members observe costly machines at random intervals to verify their proper functioning.

This practice assists in identifying factors that contribute to low utilization.

Data collected during inspection for quality control also comes under observational data.

Factors for Selecting Sources of Data

These factors have to be incorporated in selecting the sources of data collection:

Type of study

Availability of data

Time frame

Key Sources of Data

Duration: 15 minutes

You are planning to gather and analyze data for your study, but you have not developed any specific tools or applications for data collection yet.

What are the methods used to collect data?

Answer: The various ways to collect data are:

- Data collected and published by the government
- Data collected by conducting experiments
- Data collected through observational studies
- Data collected by conducting surveys

Data Quality Issues

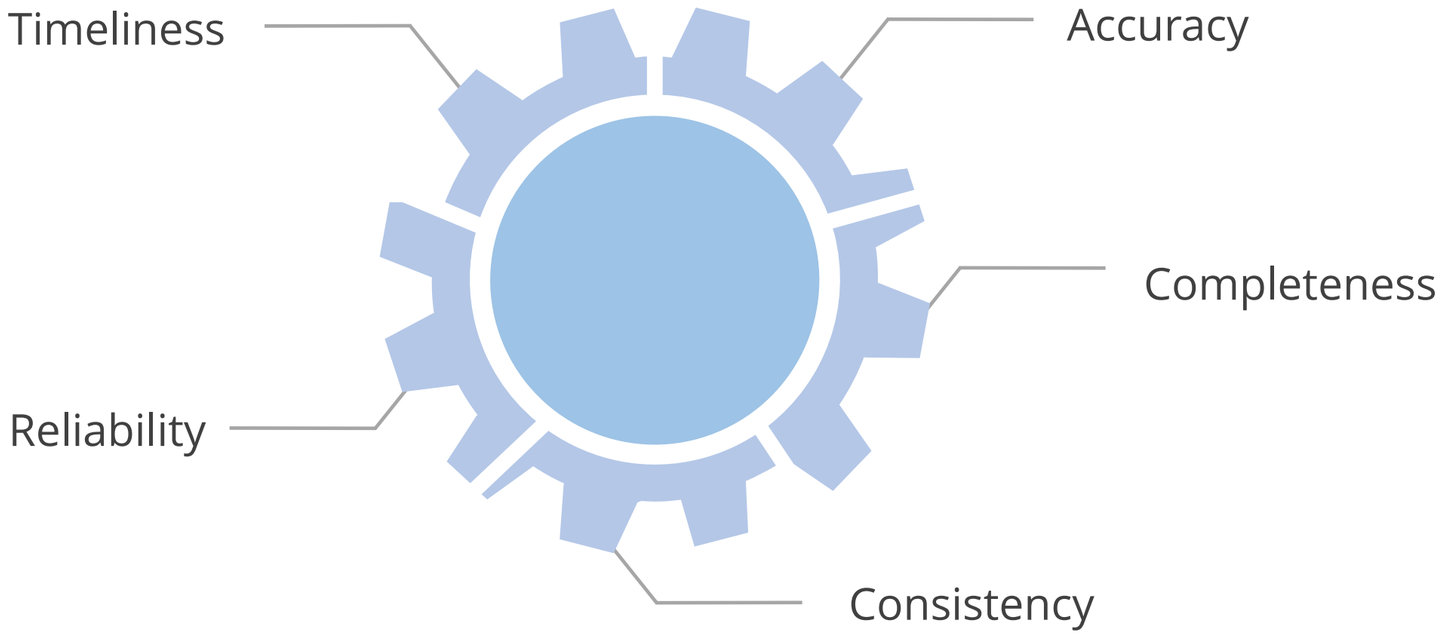
Data Quality

Data quality refers to the degree to which data serves its intended purpose.

Findings from statistical analysis must be used carefully to avoid issues with data quality.

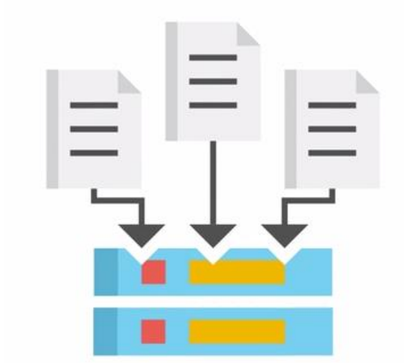
Data Quality

Data quality refers to the degree of:



Factors That Drive Data Quality

There are six factors that drive data quality.

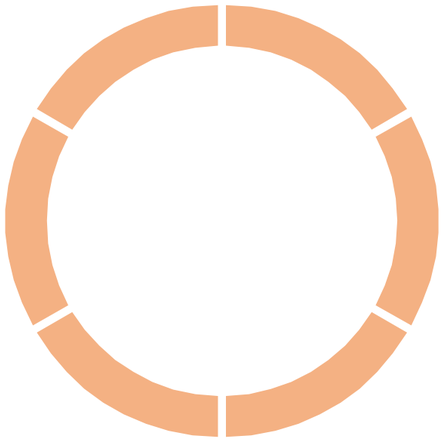


- Quality of measuring devices, questionnaires used, and approaches to data collection
- Clarity of the information needed and its communication to personnel involved in data collection
- Elimination of outliers and nonrepresentative data
- Use of appropriate formats
- The expertise of individuals involved in data collection
- Willingness to provide data to concerned parties

Steps to Minimize Poor Data Collection

Allocate adequate time for data collection

Track and eliminate outliers



Train data collection personnel

Pretest and evaluate questionnaires

Key Takeaways

- Data refers to the facts and figures that researchers collect, analyze, and summarize for presentation and interpretation.
- The two main types of data are qualitative and quantitative.
- Researchers collect statistical data as either primary or secondary data.
- Cross-sectional data refers to data collected by observing numerous subjects.

Key Takeaways

- Data recorded or stored in a predefined format for easy search and use during analysis is referred to as structured data.
- Data quality refers to the degree or extent of accuracy, completeness, consistency, reliability, and timeliness of the data available to users for analysis.

Knowledge Check

Knowledge Check 1

Which of the following refers to the data collected through the observation of numerous subjects?

A. Cross-sectional data

B. Time series data

C. Panel data

D. Pooled data

Knowledge Check 1

Which of the following refers to the data collected through the observation of numerous subjects?

A. Cross-sectional data

B. Time series data

C. Panel data

D. Pooled data

The correct answer is **A**

Cross-sectional data refers to the data collected through the observation of numerous subjects.

Knowledge Check

2

Which of the following refers to data collected over a period of time?

A. Cross-sectional data

B. Time series data

C. Panel data

D. Pooled data

Knowledge Check

2

Which of the following refers to data collected over a period of time?

A. Cross-sectional data

B. Time series data

C. Panel data

D. Pooled data

The correct answer is **B**

Time series data refers to data collected over a period of time.

Knowledge Check 3

The entities on which data is collected are referred to as _____.

A. Data

B. Datasets

C. Elements

D. Variable

Knowledge Check

3

The entities on which data is collected are referred to as _____.

A. Data

B. Datasets

C. Elements

D. Variable

The correct answer is **C**

The entities on which data is collected are referred to as elements.

Thank You