INTRODUCTION TO DATA ANALYSIS

by

GRACE ENEH OGWUCHE

(Data Analytics Instructor – TechyJaunt BootCamp)

CONTENTS

- Overview of Data Analysis
- The Data Analysis Phases
- Data Fundamentals
- Assignment

OVERVIEW OF DATA ANALYSIS

Introduction to Data Analysis

Types of Analytics
(Descriptive, Diagnostic, Predictive, Prescriptive)

OVERVIEW OF DATA ANALYSIS

- **DATA ANALYSIS** is the collection, transformation, and organisation of data to draw conclusions, make predictions, and drive informed decision-making.
- **DATA ANALYTICS** in simplest terms is the "***science of data***".
- A **DATA ANALYST** finds data, analyses it, and uses it to uncover trends, patterns, and relationships. Sometimes, the data driven strategy will build on what has worked in the past. Other times, it can guide a business to branch out in a completely new direction.

OVERVIEW OF DATA ANALYSIS

- Data analysts play a critical role in ensuring that every business strategy incorporates data.
- **Data-driven Decision Making** involves using facts to guide business strategy. ***Data alone*** will never be as powerful as data combined with human experience, observation, and intuition.

DATA + BUSINESS KNOWLEDGE = MYSTERY SOLVED

TYPES OF DATA ANALYSIS

- Data Analysis can be broadly categorized into four main types: **Descriptive**, **Diagnostic**, **Predictive**, and **Prescriptive Analytics**

- These types represent different stages in the process of extracting knowledge and insights from data.

➢ **Descriptive Analysis**: This type focuses on summarizing and describing the basic features of a dataset. It answers the question **"what happened?"** by providing insights into past performance and trends.

- **Examples include** calculating averages, creating charts and graphs, and identifying patterns in historical data.

TYPES OF DATA ANALYSIS

- **Diagnostic Analysis**: Diagnostic analysis delves deeper to understand why something happened. It explores the reasons behind observed trends or patterns.

- This type of analysis often involves identifying correlations, conducting root cause analysis, and discovering relationships between different variables.

- **Predictive Analysis**: Predictive analytics uses historical data and statistical techniques to forecast future outcomes or trends. This involves building predictive models that can anticipate what might happen based on past behaviour.

- **Examples include** predicting customer churn, forecasting sales, and identifying potential risks.

TYPES OF DATA ANALYSIS

➢ **Prescriptive Analysis**:

- Prescriptive analytics takes predictive insights a step further by recommending actions to optimize outcomes. It answers the question **"what should be done?"**.
- This type of analysis often involves using optimization algorithms, simulation techniques, and machine learning to suggest the best course of action.

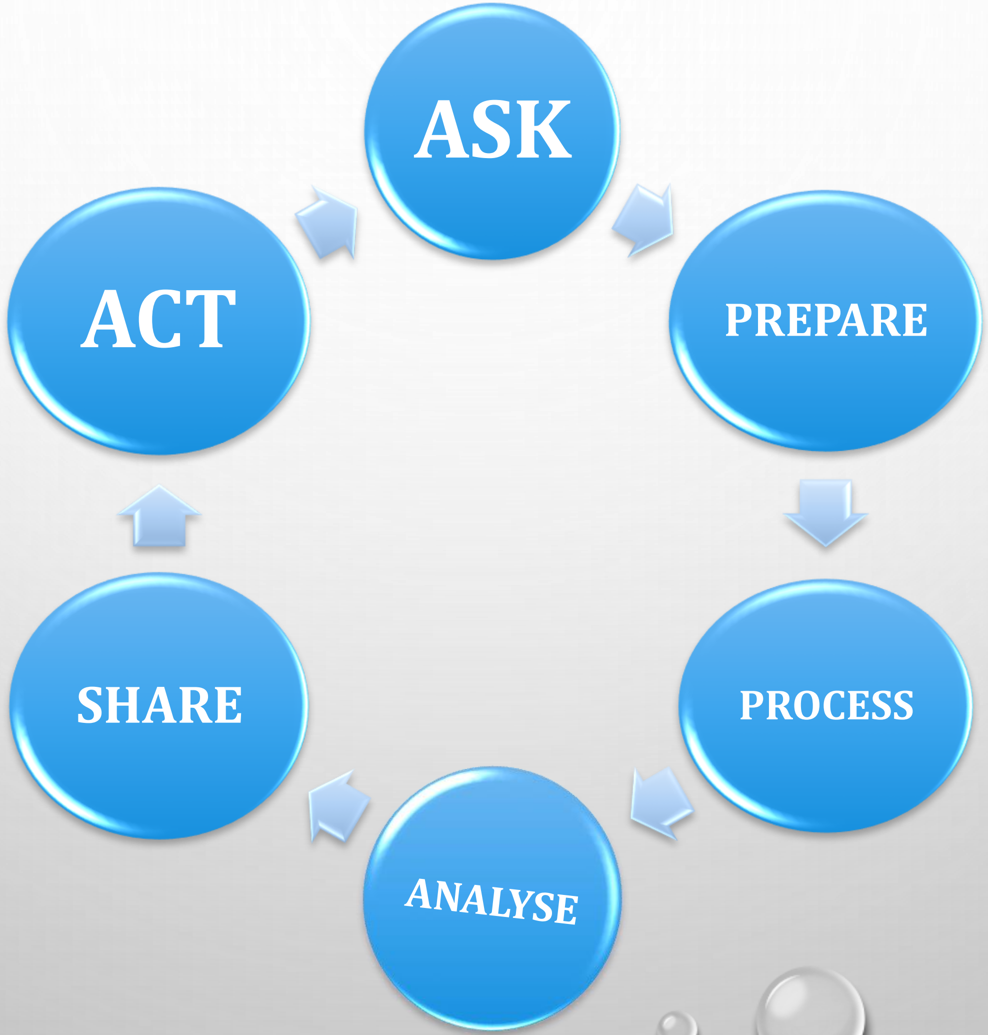
DATA ANALYSIS TOOLS

- Data analysis tools are software programs and applications used to collect, process, analyse, and visualize data to extract meaningful insights and support decision-making.
- These tools range from simple spreadsheet programs to complex statistical software and programming languages.
- Spreadsheets (Microsoft Excel, Google Sheets)
- Programming Languages (Python and R)
- Statistical Software (SPSS)
- Data Visualisation tools (Power BI, Tableau, Google Data Studio)
- Database Query Languages (PostgreSQL, MySQL)

THE DATA ANALYSIS PHASES

Ask, Prepare, Process, Analyse, Share and Act

DATA ANALYSIS PHASES/PROCESS



ASK

PREPARE

PROCESS

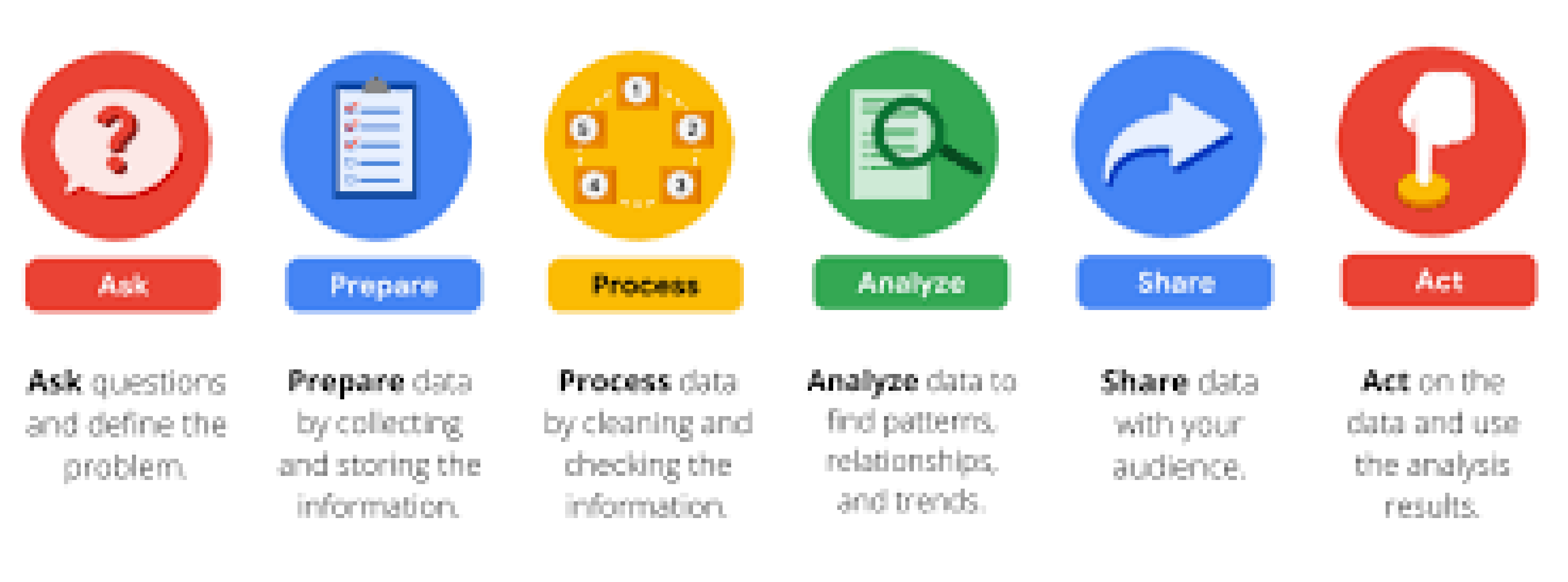
ANALYSE

SHARE

ACT

DATA ANALYSIS PHASES/PROCESS

- Data analysts employ data-driven decision-making and follow a systematic process. There are six (6) steps of the data analysis process.



Ask questions and define the problem.

Prepare data by collecting and storing the information.

Process data by cleaning and checking the information.

Analyze data to find patterns, relationships, and trends.

Share data with your audience.

Act on the data and use the analysis results.

ASK

It is impossible to solve a problem if you do not know what it is. There are some things to consider:

- Define the problem you are trying to solve.
- Make sure you fully understand the stakeholders’ expectations.
- Focus on the problem to avoid any distractions
- Collaborate with stakeholders and keep an open line of communication.
- Take a step back and see the whole situation in context.

Questions to ask yourself in this step:

1. What are my stakeholders saying their problems are?
2. Now that I have identified the issues, how can I help the stakeholders resolve their questions?

PREPARE

You will decide what data you need to collect in order to answer your questions and how to organise it so that it is useful. You might use your business task to;

- Decide what metrics to measure
- Locate data in your database
- Create security measures to protect that data

Questions to ask yourself in this step:

1. What do I need to figure out how to solve this problem?
2. What research do I need to do?

PROCESS

You will need to clean up your data to get rid of any possible errors, inaccuracies, or inconsistencies. This might mean:

- Using spreadsheet functions to find incorrectly entered data.
- Using SQL functions to check for extra spaces.
- Removing repeated entries
- Checking as much as possible for bias in the data.

Questions to ask yourself in this step:

1. What data errors or inaccuracies might get in my way of getting the best possible answer to the problem I am trying to solve?
2. How can I clean the data so that the information I have is more consistent?

ANALYSE

You will want to think analytically about your data. At this stage, you might sort and format your data to make it easier to;

- Perform calculations
- Combine data from multiple sources
- Create tables with your results

Questions to ask yourself in this step:

1. What story is my data telling me?
2. How will my data help me solve this problem?
3. Who needs my company's product or service?
4. What type of person is most likely to use it?

SHARE

Summarise your results with clear and enticing visuals of your analysis using data with graphs or dashboards. This is your chance to show the stakeholders you have solved their problem and how you got there.

Sharing will certainly help your team;

- Make better decisions
- Make more informed decisions
- Lead to stronger outcomes
- Successfully communicate your findings

Questions to ask yourself in this step:

1. How can I make what I present to the stakeholders engaging and easy to understand?
2. What would help me understand this if I were the listener?

ACT

You will take everything you have learned from your data analysis and put it to use. This could mean providing your stakeholders with recommendations based on your findings so they can make data-driven decisions.

Questions to ask yourself in this step:

- How can I use the feedback I received during the share phase (step 5) to meet the stakeholders' needs and expectations?

These six steps can help you break the data analysis process into smaller manageable parts known as **structured thinking.**

DATA FUNDAMENTALS

Database and DBMS

Data Formats

Structure of Data

Data Collection & Data Sources

DATABASES

- **DATA** (singular - datum) are raw or unprocessed facts and are often stored in a database, hard drive, etc.
- A **DATABASE** is an organised collection of data stored and accessed electronically. It enables analysts to manipulate, store, and process data.
- **DATABASES** are essential for storing and managing data in today's digital world. They can be stored as structured, semi-structured, or unstructured data, making them indispensable for modern applications.
- They are managed using Database Management Systems(DBMS), which provide tools for creating, retrieving, and modifying data.

TYPES OF DATABASES

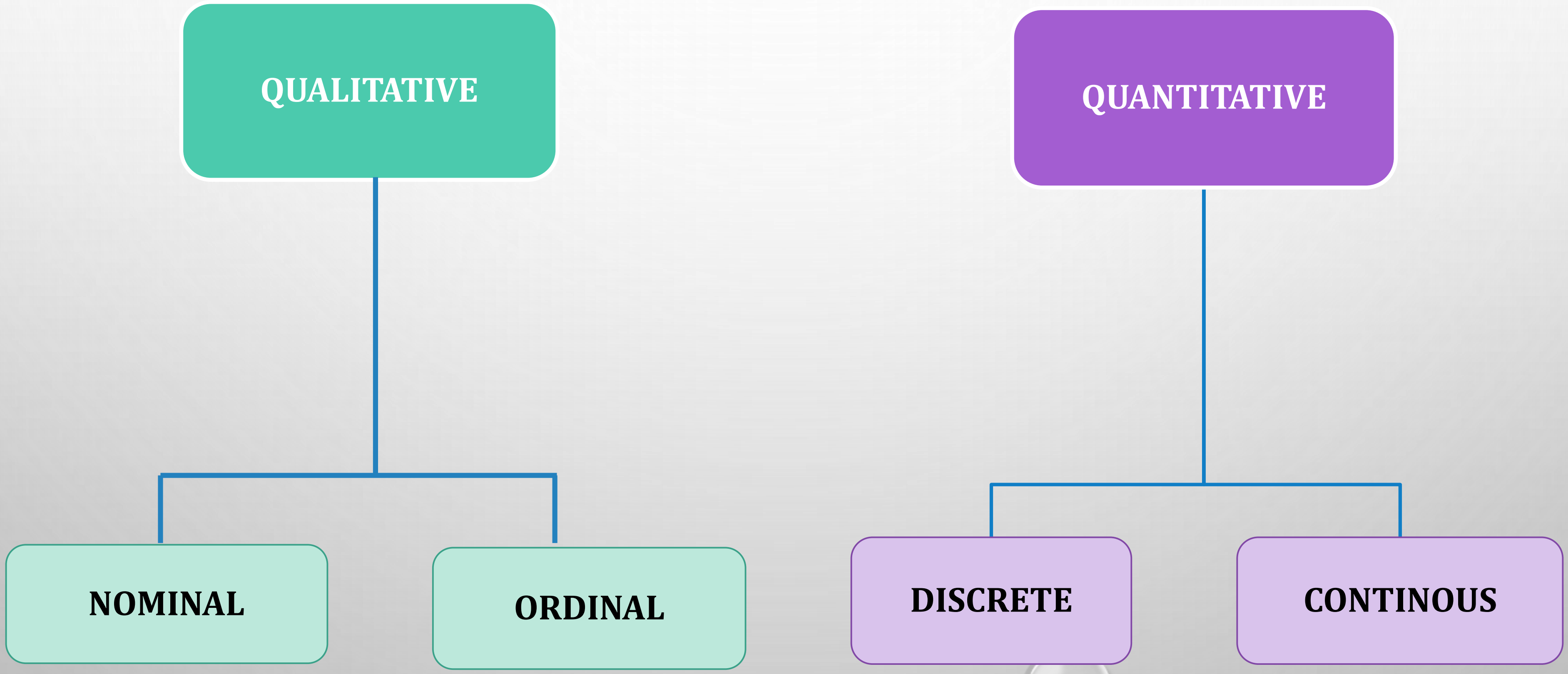
There are different types of databases.

- Relational Database
- NoSQL Database
- Graph Databases
- Time-Series Databases
- Columnar Databases. Etc.

TYPES OF DATABASES

- **RELATIONAL DATABASES**: is a database that contains a series of related tables that can be connected via their relationships. every piece of information has a relationship with every other piece of information.
- **Examples** of relational databases include
- MySQl, Oracle Database, Microsoft SQL Server, PostgreSQL, And IBM DB2.
- In terms of **STRUCTURE**, they organise data into tables with rows and columns, allowing for relationships between different data points through shared values. most relational databases use Structured Query Language (SQL) to access and manipulate data.

DATA FORMATS



QUALITATIVE

NOMINAL

ORDINAL

QUANTITATIVE

DISCRETE

CONTINOUS

DATA FORMATS

QUALITATIVE DATA:

- Describes subjective and explanatory measures of qualities and characteristics **or** things that cannot be measured like your *hair colour*, *genotype*, *race*, *gender, religion, blood group,* etc.
- Qualitative data is great for helping us answer **why** questions. It helps analysts understand their quantitative data better by providing a reason or more thorough explanation.

DATA FORMATS Cont.d

QUANTITATIVE DATA:

- Describes the specific and objective measures of numerical facts **or** things you measure such as *height*, *weight*, *age*, *income*, *population*, etc.
- This can often be the **what**, **how many**, and **how often** about a problem. With quantitative data, we can see numbers visualised as charts or graphs.
- Qualitative data can then give us a more high-level understanding of why the numbers are the way they are. This is important because it helps us add context to a problem.

DATA FORMATS Cont.d

Qualitative Data is further divided into two categories; **Nominal** and **Ordinal** data

Nominal Data does not have a sequence. It is categorised without a set order. Its values are categorised for representation.

For example,

gender of individuals where

male = 1 & female = 0

DATA FORMATS Cont.d

Ordinal Data has a set order or scale. Its values are assigned for coding as well as ranking.

For example,

exam performance where

good = 1 moderate = 2 bad =3

survey opinions where

1 = strongly agree 2 = agree 3 = neither agree nor disagree

4 =disagree 5 = strongly disagree

DATA FORMATS Cont.d

Quantitative Data is further divided into two (2) categories; Discrete and Continuous Data

- **Discrete Data** has a limited number of values.

e.g. *age, population of individuals,* etc.

- **Continuous Data** can be measured using a timer, and its value can be shown as a decimal in several places.

e.g. *height, weight, average, distance, speed,* etc.

THE STRUCTURE OF DATA

Structured Data is data organised in a standardised or consistent format. It is often organised in a table with rows and columns.

E**xample:** *a class performance sheet which consists of students' names, genders, ca scores, exam scores, and total.*

Unstructured Data is data that does not have a pre-defined model or is not organised in a pre-defined manner.

Example: *social media posts, emails, memos,* etc.

	STRUCTURED DATA	UNSTRUCTURED DATA
1	Defined data types	Varied data types
2	Mostly quantitative data	Mostly qualitative data
3	Easy to organise	
4	Easy to search	Difficult to search
5	Easy to analyse	Provides more freedom for analysis
6	Stored in relational databases and data warehouses	Stored in data lakes, data warehouses, and NoSQL databases
7	Stored in rows and columns	Cannot be placed in rows and columns
8	**Examples:** Excel spreadsheets, Google sheets, customer data, phone records, transaction history, etc.	**Examples:** text messages, social media comments, phone call transcriptions, various log files, images, audio, video, etc.

DATA COLLECTION METHODS & TOOLS

QUALITATIVE DATA TOOLS	QUANTITATIVE DATA TOOLS
• Focus Groups	• Structured Interviews
• Social Media Text Analysis	• Surveys
• In-person Interviews	• Polls

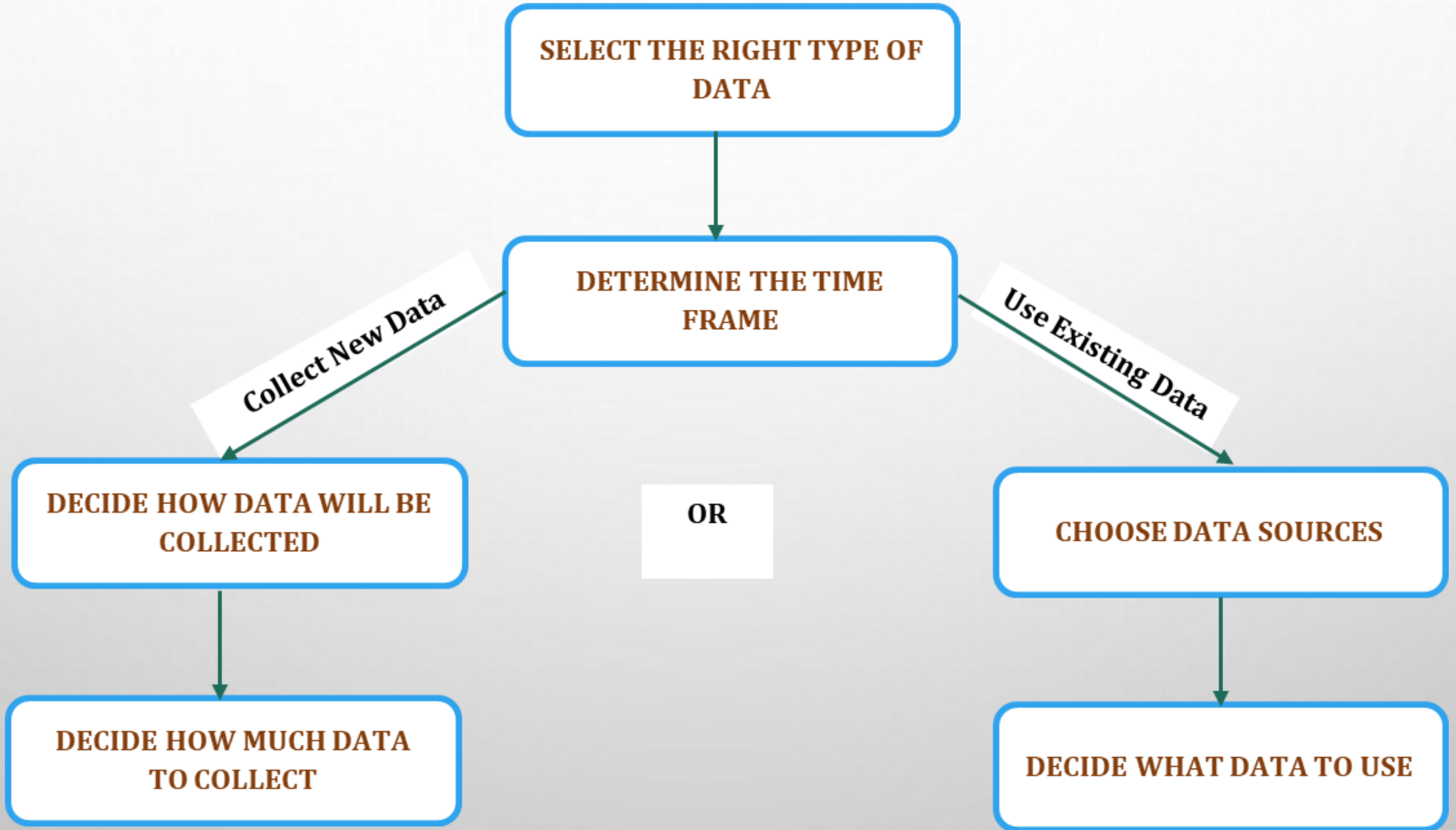
DATA TYPES FROM DATA SOURCES

- **First-party Data:** data collected by an individual or group using their own resources.

  It is typically the preferred method because you are certain of the source.
- **Second-party Data:** data collected by a group directly from its audience and the sold.
- **Third-party Data:** data collected from outside who did not collect it directly.

  It might be generated from multiple sources.

DATA COLLECTION CONSIDERATIONS



SELECT THE RIGHT TYPE OF DATA

DETERMINE THE TIME FRAME

Collect New Data

DECIDE HOW DATA WILL BE COLLECTED

DECIDE HOW MUCH DATA TO COLLECT

OR

Use Existing Data

CHOOSE DATA SOURCES

DECIDE WHAT DATA TO USE

IDENTIFYING GOOD DATA SOURCES

The more quality data we have, the more confidence we have in our decisions.

Acronym for the identification process

R - Reliable

O - ORIGINAL

C - COMPREHENSIVE

C - CURRENT

C – CITED

If you have ORIGINAL data from a RELIABLE organization and it is COMPREHENSIVE, CURRENT, and CITED, it **ROCCCs!**

IDENTIFYING GOOD DATA SOURCES Cont.d

If you are choosing a data source, think about three (3) things.

1. Who created the data set?
2. Is it part of a credible organisation?
3. When last was the data refreshed/updated?

Every good solution is discovered by avoiding bad data. For good data, stick with vetted public datasets, academic papers, financial data, and governmental agency data.

NB: *No matter what data you use, it should be tested for accuracy and trustworthiness.*

IDENTIFYING GOOD DATA SOURCES

- The goal of all data analysts is to use data to draw accurate conclusions and make good recommendations. All that starts with having complete, correct, and relevant data.
- When you clean data, you transform it into a more useful format, create accurate information, and remove outliers.
- When data is ***misinterpreted***, it can lead to ***huge losses***.
- When data is ***used strategically***, businesses can ***transform*** and ***grow*** their revenue.

OPEN DATA

For data to be considered open, it has to:

i. Be available and accessible to the public as a complete dataset

ii. Be provided under terms that allow it to be reused and redistributed

iii. Allow universal participation so that anyone can use, reuse, and redistribute the data

SOME SITES FOR OPEN DATA

- google cloud public datasets
- dataset search
- kaggle datasets

ASSIGNMENT

1. Describe the other types of databases with examples.
2. Describe the different methods of collecting data and tools used
3. Difference between Data Analysis, Data Science, and Data Engineering
4. Differences and similarities between Data Analysis and Business Intelligence Analysis
5. Outline some applications of Data Analysis in different industries.