QMS 101 Introductory Statistics

Topic II: Data Collection and Presentation

Dr. Anna Fome

Jordan University College

2026-05-01

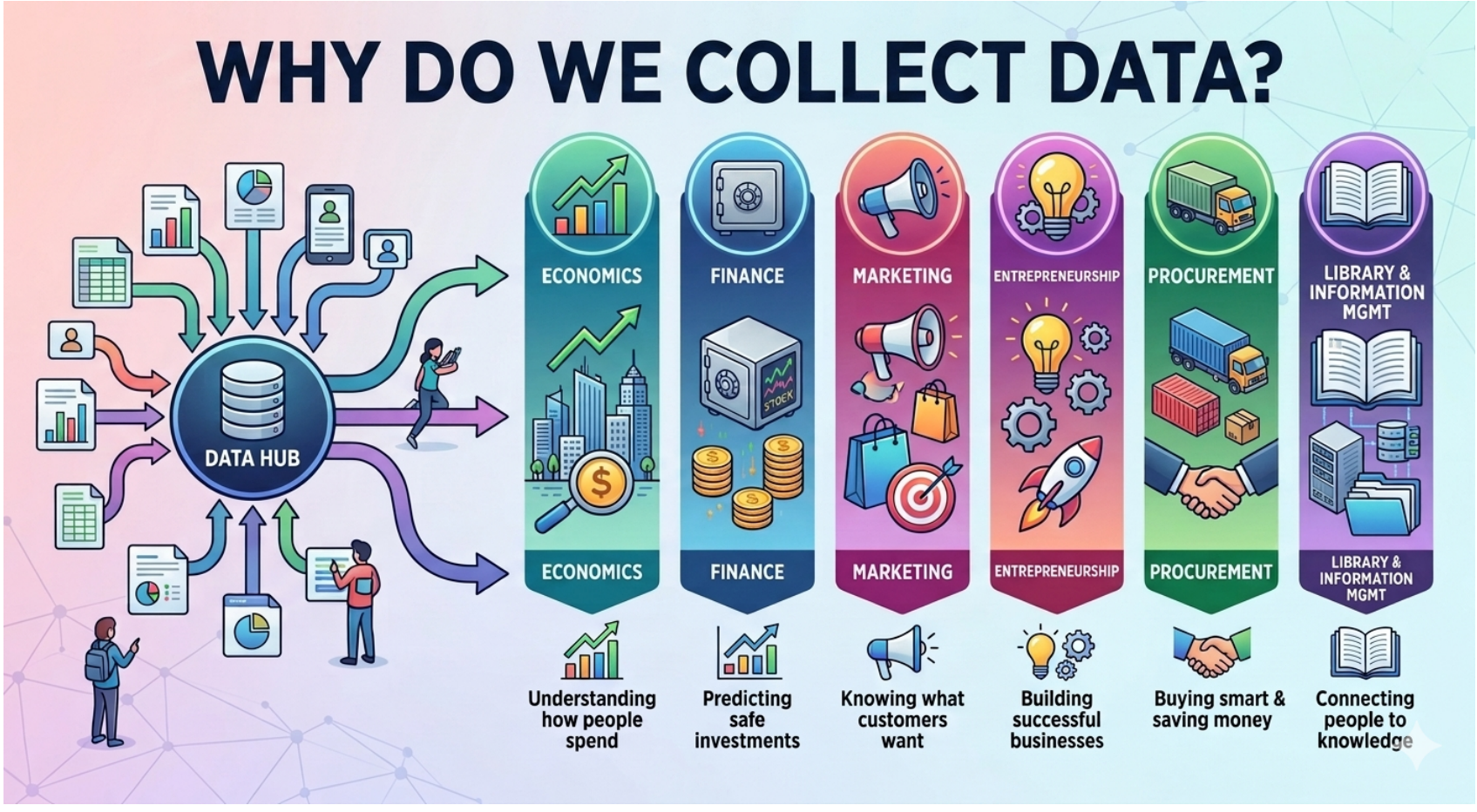
Learning Outcomes

By the end of this topic, students will be able to:

- Explain the concept and motivation behind data collection
- Distinguish between primary and secondary data
- Evaluate different data collection methods
- Construct frequency distributions
- Draw histograms, frequency polygons, and ogives

PART A: Data Collection

2.1 Motivation:



Golden Rule of Statistics

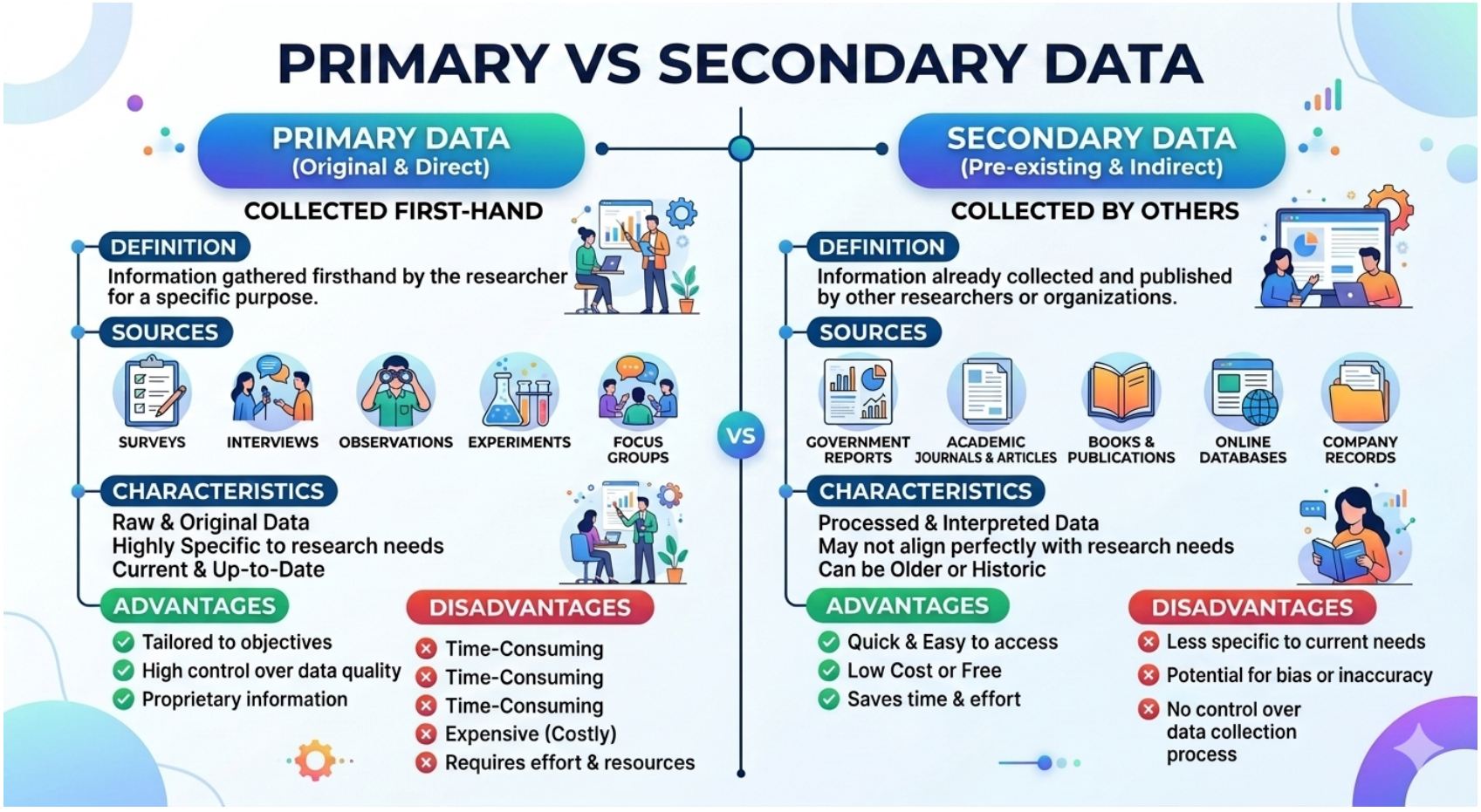
Garbage In = Garbage Out (GIGO)

Poor data $\rightarrow$ *Poor decisions*

Why Data Matters

- Foundation of all statistical analysis
- Used in:
  - Business decisions
  - Government planning
  - Healthcare
  - Research

2.2 Types of Data



When to Use Each?

Use Primary Data When:

- No existing data exists
- You need specific information
- Accuracy is critical

Use Secondary Data When:

- Time is limited
- Budget is low
- Large datasets are needed

2.3 Methods/Approaches of Data Collection

Main methods:

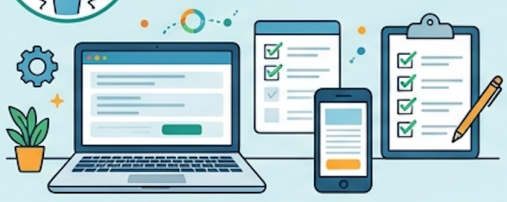
1. Survey
2. Interview
3. Observation
4. Experiment
5. Focus Group

Why Method Choice Matters?

WHY METHOD CHOICE MATTERS: A MOTIVATING EXAMPLE

Three researchers explore the question: "Do Tanzanian university students eat healthily?"

RESEARCHER 1: ONLINE SURVEY



REACHES MANY STUDENTS QUICKLY.

- Data relies on self-reports. People might overreport good habits or underreport bad ones.
- Possible Bias: Social Desirability, Recall Error.

RESULT: High numbers of "yes" answers, but potentially inaccurate data.

RESEARCHER 2: CANTEEN OBSERVATION

OBSERVES REAL BEHAVIOR DIRECTLY.

- Limited to one setting and time (e.g., lunch).
- Cannot see all meals (e.g., dinner, snacks off-campus).
- Misses unobservable choices.

RESULT: Specific data on canteen meals, but but incomplete picture.

RESEARCHER 3: DIRECT INTERVIEW

GATHERS IN-DEPTH, DETAILED INFORMATION

- Small sample size due to time constraints.
- Cannot generalize to the whole population easily.
- Potential interviewer bias.

RESULT: Rich personal stories, but not representative of all students.

HIGH HEALTHY EATING?

PARTIAL VIEW

THE CONSEQUENCES OF METHOD CHOICE: DIFFERENT METHODS CAN LEAD TO CONFLICTING FINDINGS ABOUT THE SAME POPULATION.

PARTIAL VIEW

INDIVIDUAL PERSPECTIVES

CHOOSING THE RIGHT METHOD DEPENDS ON RESEARCH GOALS, RESOURCES, AND THE NATURE OF THE QUESTION.

Why Method Choice Matters?

From the Healthily Example

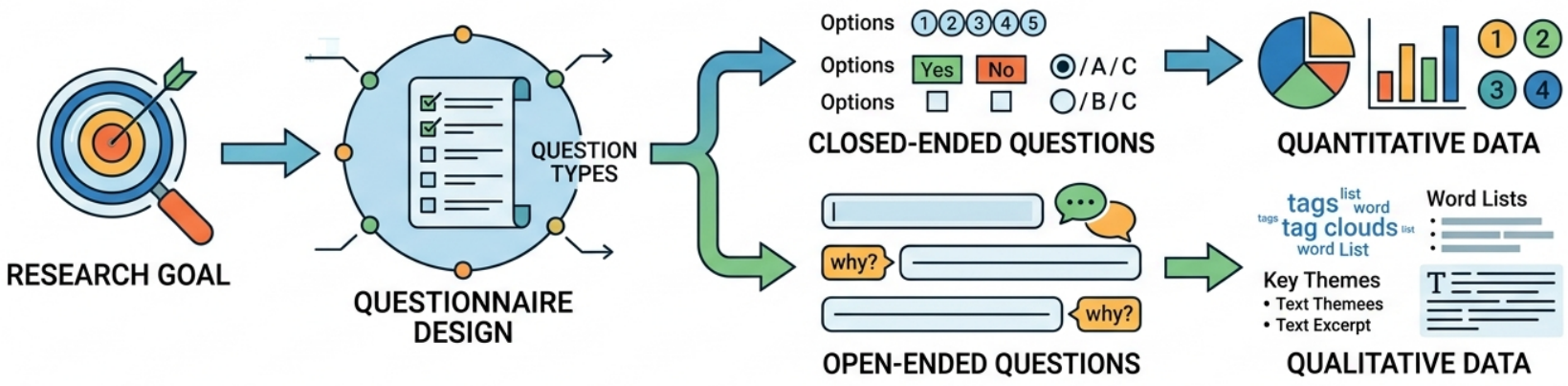
- Methods influencing findings
- No single method gives the complete picture
- A well-designed study would combine more than one methods (this combination is called **triangulation)**

Survey & Questionnaire

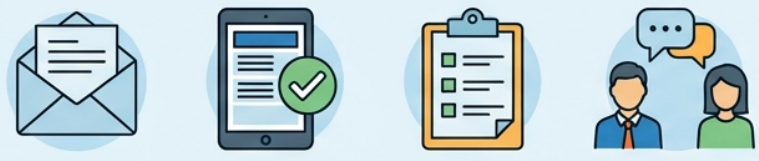
- **Survey** is the overall research **process** of gathering data from a group, while
  - **Questionnaire** is a document (**a tool**) with actual list of structured questions.
  - **The Goal:** To collect self-reported data (opinions, behaviors, and characteristics.
  - **The Method:** Respondents usually read and answer a structured set of questions independently (written or digital).
- **In short:** You use a *questionnaire* to conduct a *survey*.

Feature	Questionnaire	Survey
Definition	A specific set of questions.	A broad research method. It contains a questionnaire
Main Goal	To get information from **one** person.	To find trends in a **group.**
Analysis	Answers might be looked at one-by-one.	It is **aggregated** (added together) and analyzed.
Example	A medical history form (the doctor only cares about *your* health).	A patient satisfaction study (the hospital cares about *everyone's* experience).

QUESTIONNAIRE AS A DATA COLLECTION METHOD



MECHANISM (DISTRIBUTION)



PROS & CONS

+PROS

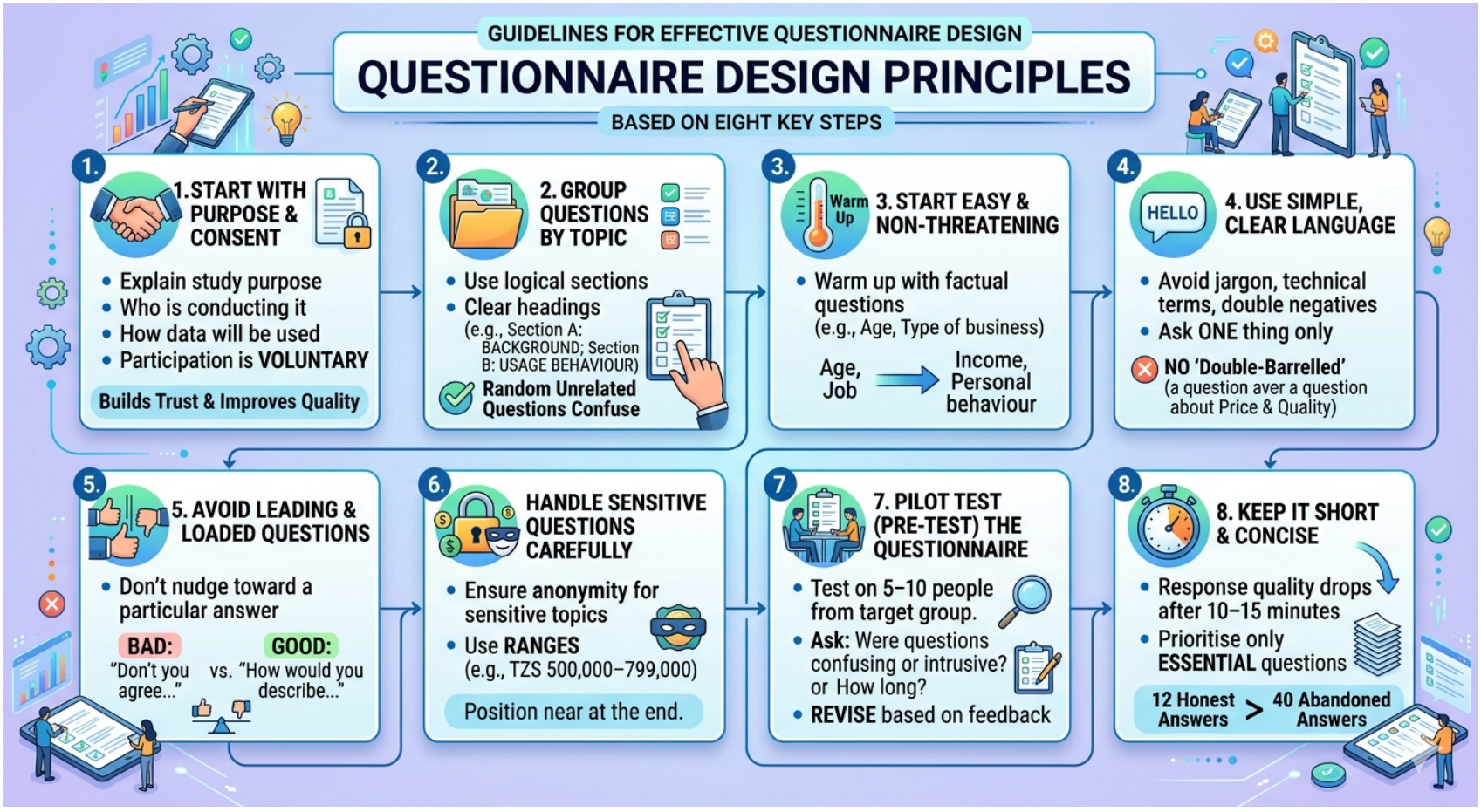
- LOW COST
- HIGH SCALABILITY
- EASY ANALYSIS

- CONS

- LIMITED DEPTH
  - SHORT ANSWERS
  - SHALLOW WATER
- LOW RESPONSE RATE
- POSSIBLE BIAS

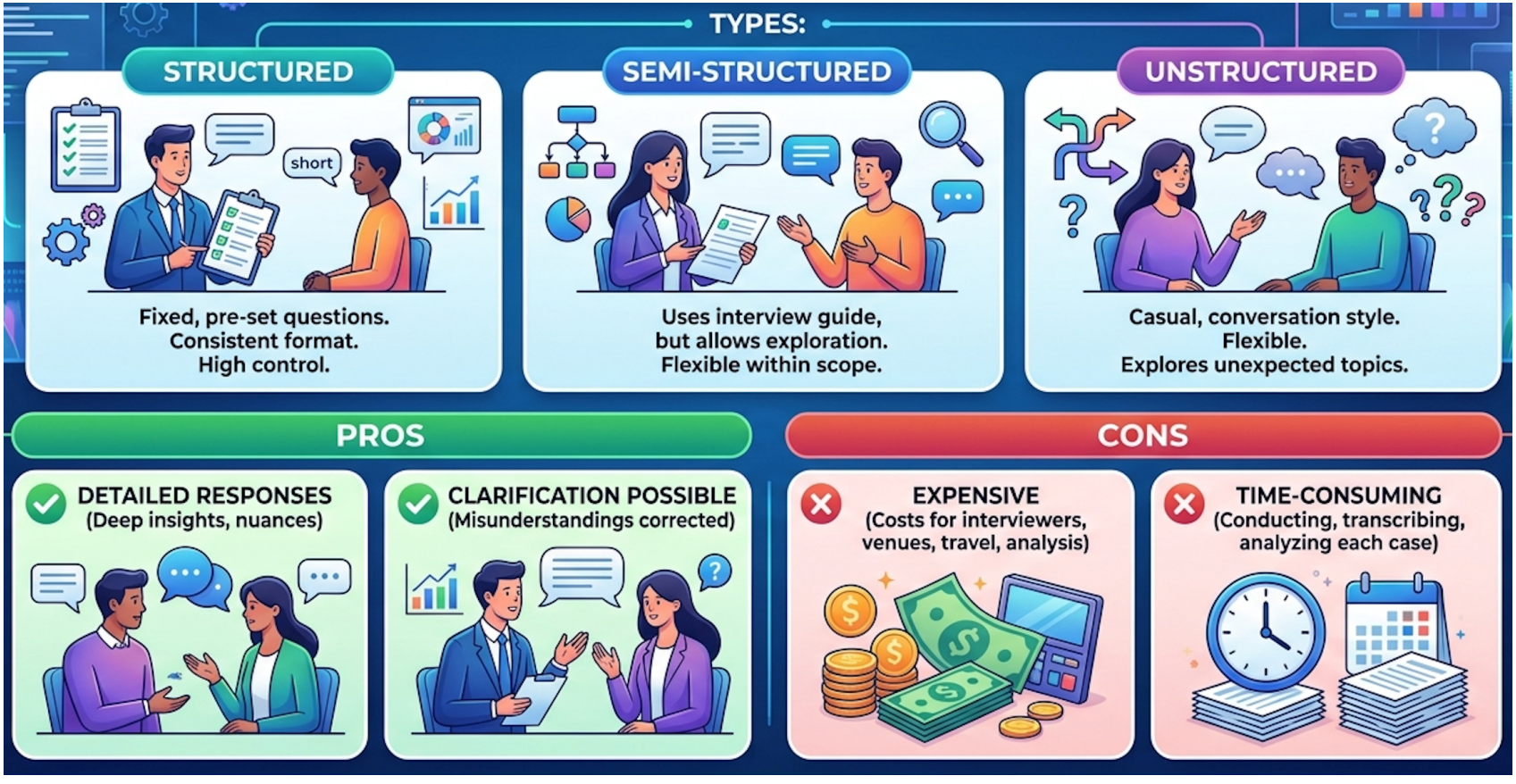
MODERN METHOD OF RESEARCH & DATA COLLECTION

Designing an Effective Questionnaire



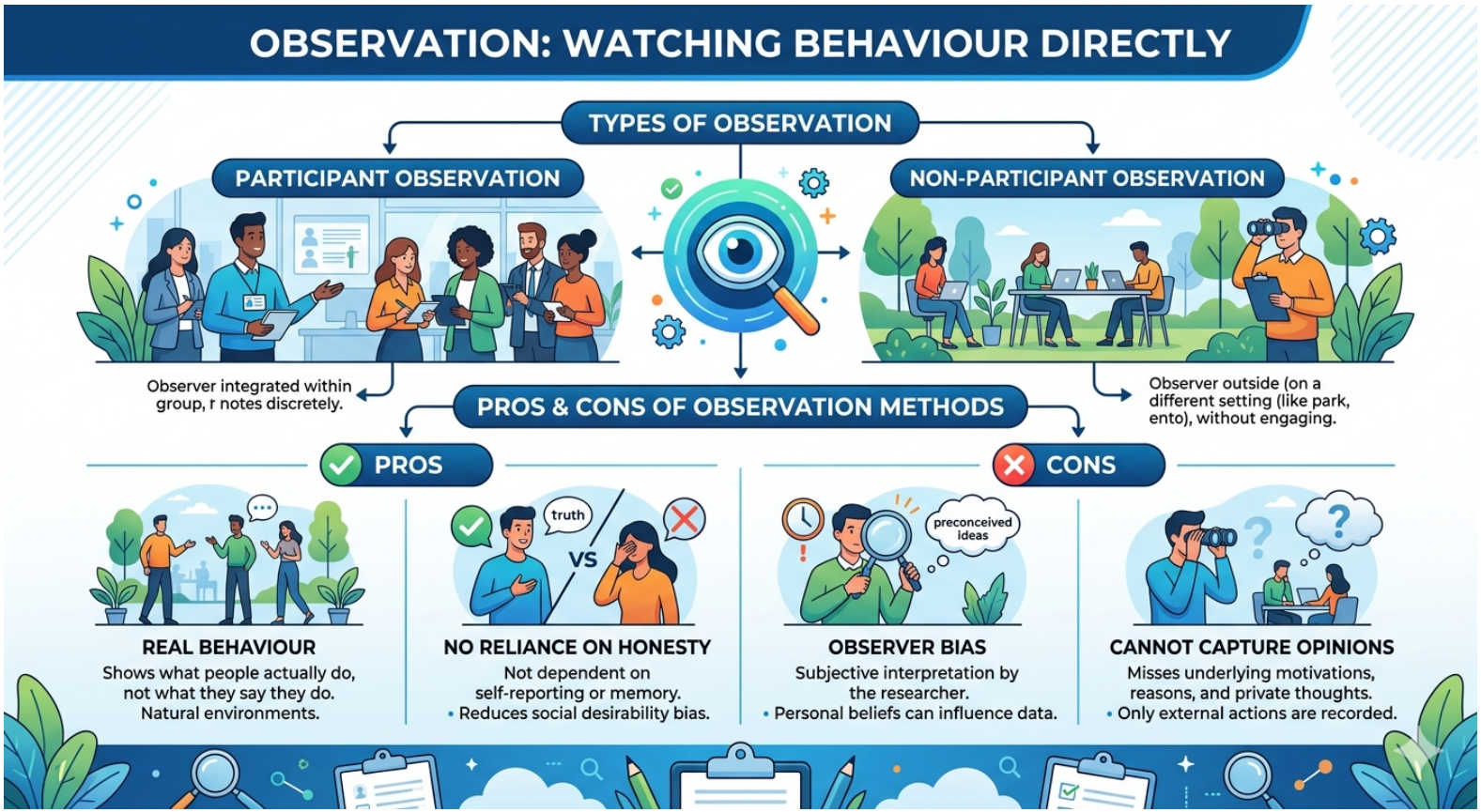
Interview

- Direct questioning/verbal communication. Three types by Structure:



Observation

- Watching behaviour directly



Experiment

- Manipulate variables

Key Concepts:

- Control group
- Experimental group
- Random assignment

Pros:

- Establish causation

Cons:

- Expensive
- Ethical issues

Focus Groups

- Small group discussion (6–12 people)

Pros:

- Rich insights
- Multiple perspectives

Cons:

- Not generalisable
- Group bias

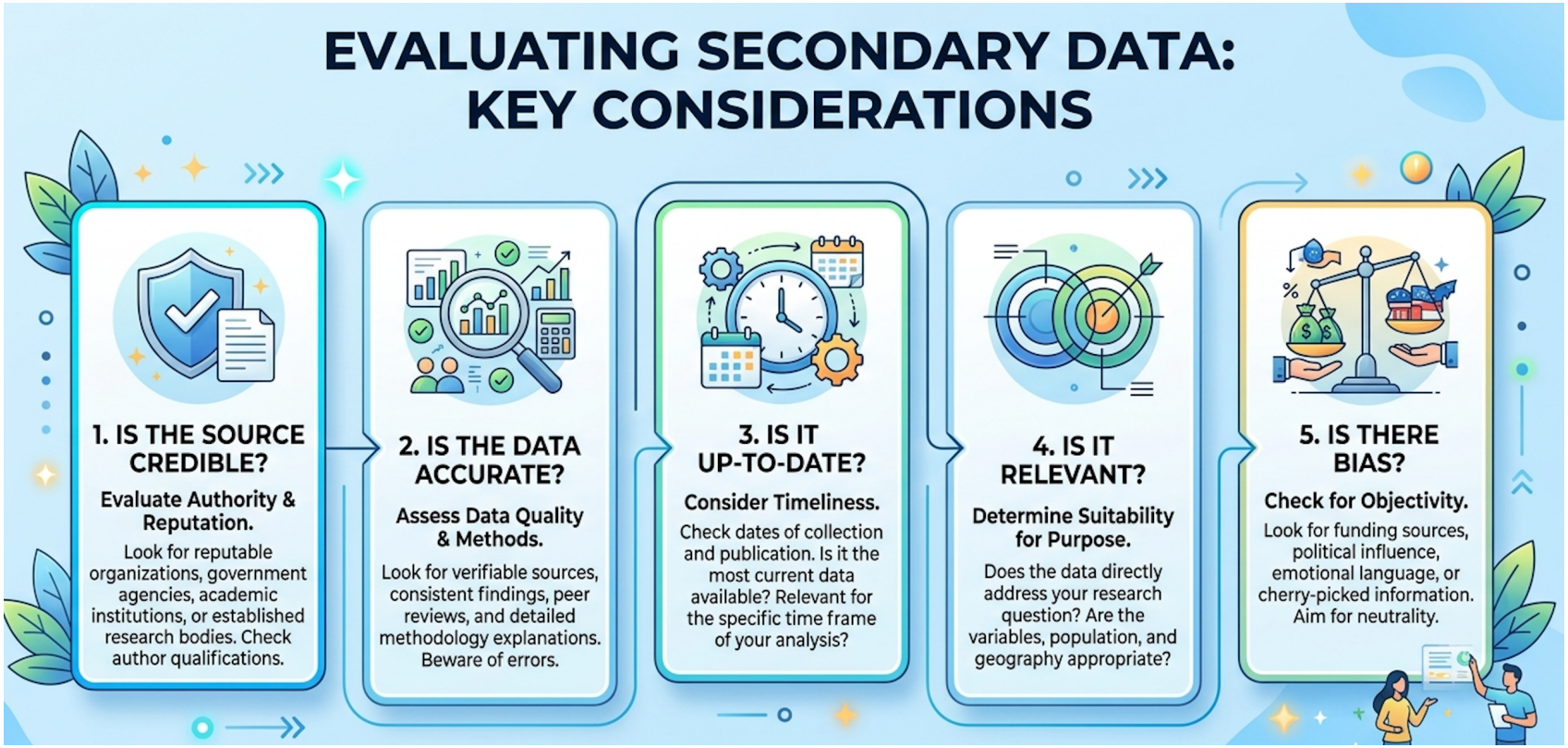
2.4 Secondary Data Sources

Common sources:

- Government statistics
- World Bank / IMF
- Journals
- Company records
- Online databases

Evaluating Secondary Data

Ask:



2.5 Choosing Methods

Factors:

- Research objective
- Population size
- Budget
- Time
- Depth vs breadth
- Need for causation

Data Entry and Cleaning

- Data cleaning involves correcting errors in the dataset
- It includes handling missing values appropriately
- It also involves standardizing responses for consistency
- Data preparation organizes the cleaned data into a structured format
- The structured data is then made ready for analysis

Data Entry and Cleaning (Examples)

- Correcting errors (e.g., Age of the child$=150 \rightarrow$ invalid, corrected or removed)
- Handling missing data (blank responses marked or treated appropriately)
- Removing duplicates (duplicate entries kept once)
- Coding responses (Yes/No $\rightarrow 1/0$)
- Standardizing categories (Male, male, M $\rightarrow$ Male)
- Organizing cleaned data into tables for analysis

PART B: Data Presentation

Why Data Presentation Matters

- After collecting data, it should be presented clearly for easy understanding
- Good presentation helps others quickly identify patterns
- It also makes it easier to compare values/variables and draw insights
- Makes reports and research more engaging and professional

What is a Frequency Distribution?

Definition: A table that organises raw data by showing each value (or class interval) and the number of times it occurs.

Why do we need it?

- Raw data is hard to read and interpret
- It is the foundation for all graphical displays

Example: Raw Data

Daily spending (TZS '000) for 20 students:

45, 32, 55, 38, 47, 60, 33, 50, 42, 38, 55, 47, 30, 65, 42, 50, 38, 55, 47, 44

This is hard to interpret as a list.

A frequency table makes patterns immediately visible.

Two types

1 . Discrete or Ungrouped — for data with few distinct values

- Displays data where each value is separate and distinct
- Frequencies show how often each exact value occurs
- Used when data is not continuous (discrete values)
- No overlapping or class intervals between values
- Examples: number of children in a family, rooms in a house

How to construct:

- List all possible values from lowest to highest
- Use tally marks to record occurrences
- Group tallies in sets of five for easy counting
- Convert tallies into frequency values

Example: The number of children in 15 families: 2, 3, 1, 2, 4, 2, 1, 3, 2, 5, 1, 2, 3, 4, 2. Construct the frequency distribution table

Number of Children	Tally	Frequency
1		
2		
3		
4		
5		
6		

2. **Continuous or Grouped** — for continuous data with a wide range

- Data is grouped into class intervals (ranges)
- Used for large datasets or continuous data
- Shows how values are distributed within intervals
- Classes are continuous (no gaps between intervals)
- Examples: age, height, weight, grade point, income

How to construct

- **Step 1: Find the Range:**
  Range = Maximum − Minimum

How to construct

- **Step 1: Find the Range:**
  Range = Maximum − Minimum
- **Step 2: Decide Number of Classes (k).** You will be told, otherwise
  Use Sturges' Rule:
  $k \approx 1 + 3.322 \times \log_{10}(n = \text{No. of observed data})$

How to construct

- **Step 1: Find the Range:**
  Range $=$ Maximum $-$ Minimum
- **Step 2: Decide Number of Classes (k).** You will be told, otherwise
  Use Sturges' Rule:
  $k \approx 1 + 3.322 \times \log_{10}(n =$ No. of observed data$)$
- **Step 3: Calculate class width (h):**
  $h = \dfrac{\text{Range}}{k} \rightarrow$ always **round up**

How to construct ...

- **Step 4: Establish class limits:**
  Start from the minimum. Add h to get each next lower limit.

How to construct ...

- **Step 4: Establish class limits:**

  Start from the minimum. Add h to get each next lower limit.

  - Classes must be **mutually exclusive** eg 10-<20, 20-<30, ...

How to construct ...

- **Step 4: Establish class limits:**

  Start from the minimum. Add h to get each next lower limit.
  - Classes must be **mutually exclusive** eg 10-<20, 20-<30, ...
- **Step 5: Tally and count** — record f for each class.

How to construct ...

- **Step 4: Establish class limits:**
  Start from the minimum. Add h to get each next lower limit.
  - Classes must be **mutually exclusive** eg 10-<20, 20-<30, ...
- **Step 5: Tally and count** — record f for each class.
- **Add extra columns:** Relative frequency, cumulative frequency.

How to construct ...

- **Step 4: Establish class limits:**

  Start from the minimum. Add h to get each next lower limit.
  - Classes must be **mutually exclusive** eg 10-<20, 20-<30, ...
- **Step 5: Tally and count** — record f for each class.
- **Add extra columns:** Relative frequency, cumulative frequency.
  - Add these column if required

Example: Raw Data (Marks of 20 students): 12, 18, 25, 30, 22, 27, 35, 40, 15, 28, 33, 21, 19, 24, 29, 31, 26, 38, 20, 23.

Class Interval	Frequency
12-18	2
18-24	6
24-30	6
30-36	4
36-42	2
Total	20

Graphical Displays

Three Essential Graphs

Feature	Histogram	Frequency Polygon	Ogive (Cumulative Curve)
Definition	Bar graph of frequency distribution	Line graph joining class midpoints	Curve showing cumulative frequency
Type of Data	Continuous (grouped data)	Continuous (grouped data)	Cumulative frequency data
Shape	Adjacent bars (no gaps)	Connected straight lines	Smooth increasing curve
Purpose	Shows distribution of data	Shows shape and trends	Shows cumulative totals
X-axis	Class intervals	Class midpoints	Class boundaries
Y-axis	Frequencies	Frequencies	Cumulative frequencies
Key Use	Identify frequency distribution	Compare distributions	Find median, quartiles, percentiles

Histogram vs. Bar Chart

Feature	**Histogram**	**Bar Chart**
Data type	Quantitative, continuous (grouped)	Categorical (qualitative) or discrete
Bars	Touch — no gaps	Separated by gaps
X-axis	Class interval	Category labels
Example	Daily spending in TZS	Preferred mobile platform (M-Pesa, Tigo...)

Warning

This is one of the **most common errors** in student work.
If your data is grouped numerical data $\rightarrow$ **histogram**.
If your data is categories $\rightarrow$ **bar chart**.

Drawing the Histogram — Step by Step

Using the spending data table from before:

1. Draw two axes: x-axis = class interval; y-axis = frequency

Drawing the Histogram — Step by Step

Using the spending data table from before:

1. Draw two axes: x-axis = class interval; y-axis = frequency
2. For each class, draw a bar with height = frequency, width= class

Drawing the Histogram — Step by Step

Using the spending data table from before:

1. Draw two axes: x-axis = class interval; y-axis = frequency
2. For each class, draw a bar with height = frequency, width= class
3. Continuous Bars: must **touch each other** — they share boundaries

Drawing the Histogram — Step by Step

Using the spending data table from before:

1. Draw two axes: x-axis = class interval; y-axis = frequency
2. For each class, draw a bar with height = frequency, width= class
3. Continuous Bars: must **touch each other** — they share boundaries
4. Label both axes with units

Drawing the Histogram — Step by Step

Using the spending data table from before:

1. Draw two axes: x-axis = class interval; y-axis = frequency
2. For each class, draw a bar with height = frequency, width= class
3. Continuous Bars: must **touch each other** — they share boundaries
4. Label both axes with units
5. Add a descriptive title: Eg "Students' Academic Performance"

Example: Raw Data (Marks of 20 students): 12, 18, 25, 30, 22, 27, 35, 40, 15, 28, 33, 21, 19, 24, 29, 31, 26, 38, 20, 23.

Class interval	Frequency
12-18	2
18-24	6
24-30	6
30-36	4
36-42	2
Total	20

Interpreting Histogram (from the example)

- The tallest bar: **modal class** - 18-<24 and 24-<30. Most students scored between **18 and <30 marks** (highest frequency = 6 in each class)
- Very few students scored **below 18** or **above 36** (lowest frequencies = 2)
- The data is **concentrated in the middle intervals**, showing where most performance lies
- The distribution is **approximately symmetric (balanced)** around the center
- There is **no extreme outlier**, as values are fairly spread within the range

Implications of the Data

- Most students are performing at an **average level**
- There is a **need to support low-performing students** (below 18)
- Few high scores suggest **limited top performance/excellence**
- Teaching methods may be **adequate but not highly effective for excellence**
- Additional support or enrichment may be needed to **improve overall performance**

The Frequency Polygon

How to draw:

1. Calculate the **midpoint** of each class:
   $$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$
2. Plot points: (midpoint, frequency)
3. Connect with straight lines
4. **Close the polygon** by extending to midpoints of hypothetical classes before and after the data (frequency $= 0$)

Example: From the students marks example we have:

Mid-point	**Frequency**
15	2
21	6
27	6
33	4
39	2

The Ogive (Cumulative Frequency Curve)

Definition: A smooth S-shaped curve drawn by plotting **cumulative frequency** against the **upper class boundary**. **How to draw:**

1. Use the upper boundary (x-axis) cumulative frequency (y-axis) columns
2. To plot, start at (lower boundary of first class, 0)
3. Connect with a smooth curve
4. The curve must end at (n, total)

Example: From the students marks example we have:

Class limit	Cumulative Freq
12	0
18	2
24	8
30	14
36	18
42	20

Using the Ogive to Find Percentiles

The ogive is the **only graph** that lets you directly estimate medians, quartiles, and percentiles.

Reading the Ogive — Student Marks Example

To find the median (Q2=50% of the data is below this value):
$\rightarrow n/2 = 20/2 = 10$. Draw horizontal line from 10 on y-axis $\rightarrow$ hit curve $\rightarrow$ drop to x-axis.
$\rightarrow$ Estimated median 25.6 (50% scored below 25.6)
To find Q1 (25% of the data is below this value):
$\rightarrow n/4 = 5 \rightarrow$ reading from ogive $\rightarrow$ Q1 **21 (**25% scored below 21)
To find Q3 (75% of the data is below this value):
$\rightarrow 3n/4 = 15 \rightarrow$ reading from ogive $\rightarrow$ Q3 **32 (**75% scored below 32)

Choosing the Right Graphical Tool

Decision Guide

- You want to **show the shape** of a distribution $\rightarrow$ **Histogram**
- You want to **compare two groups** (male vs. female) $\rightarrow$ **Frequency Polygon**
- You want to **find the median, Q1, Q3, or percentiles** $\rightarrow$ **Ogive**
- ***Key Idea***
  *Ogive values are always **estimates**, not exact numbers, because they depend on graph reading accuracy.*
- Your data is **categories** (transport mode: bus, boda-boda, walking) $\rightarrow$ **Bar chart**