Definitions

Data = *Information* ┄┄┄▶ **Numbers**, Text, Images, Videos, Signals, etc..

Variable – *characteristic being measured*
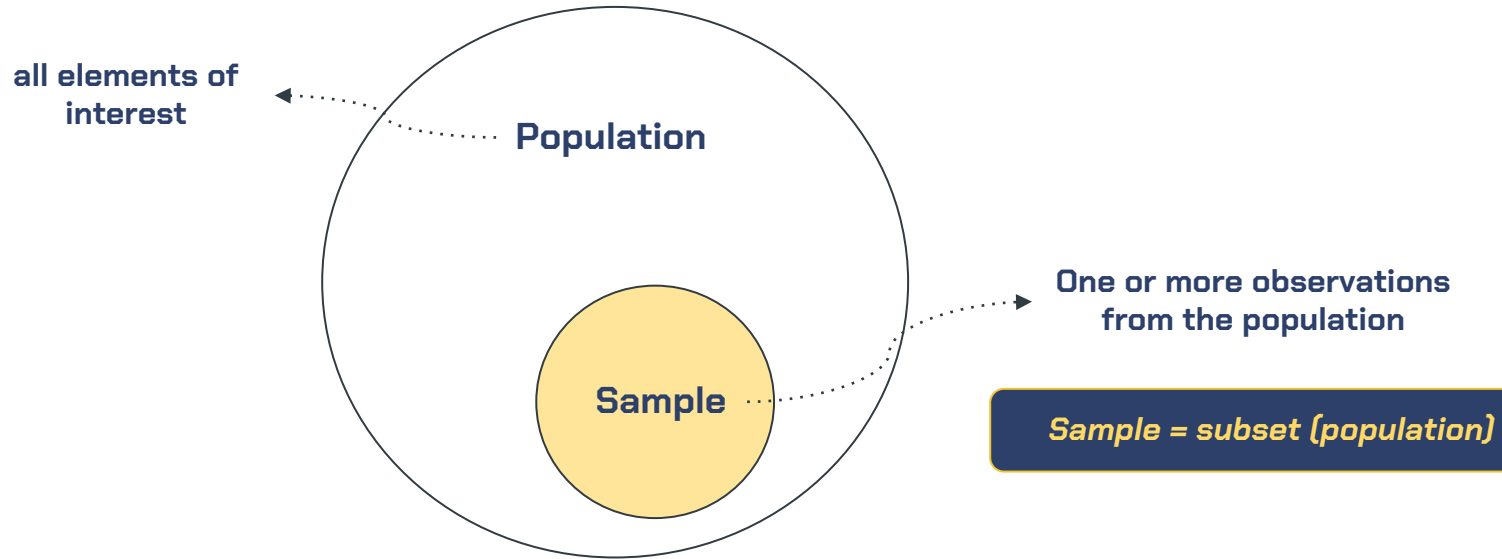- ▶ **Height –** *6 ft*
- ▶ **"varies" in a population**

Observation – *set of values related to a variable*

Population and sample

all elements of interest

Population

One or more observations from the population

Sample

Sample = subset (population)

Sampling

"**Random sampling**" is the method to collect a sample representative of the population.

The Literary Digest
NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Other methods:

- Available sample
- Volunteer sample
- Quota sample
- Referral sample
- Stratified sample

What an error by Literary Digest in 1936 teaches us about Amazon and AI bias today

4

Types of data – based on the information type

Data

Qualitative

Quantitative

Nominal

Ordinal

Discrete

Continuous

names

Names with order/levels

Does not include decimals

Includes decimals
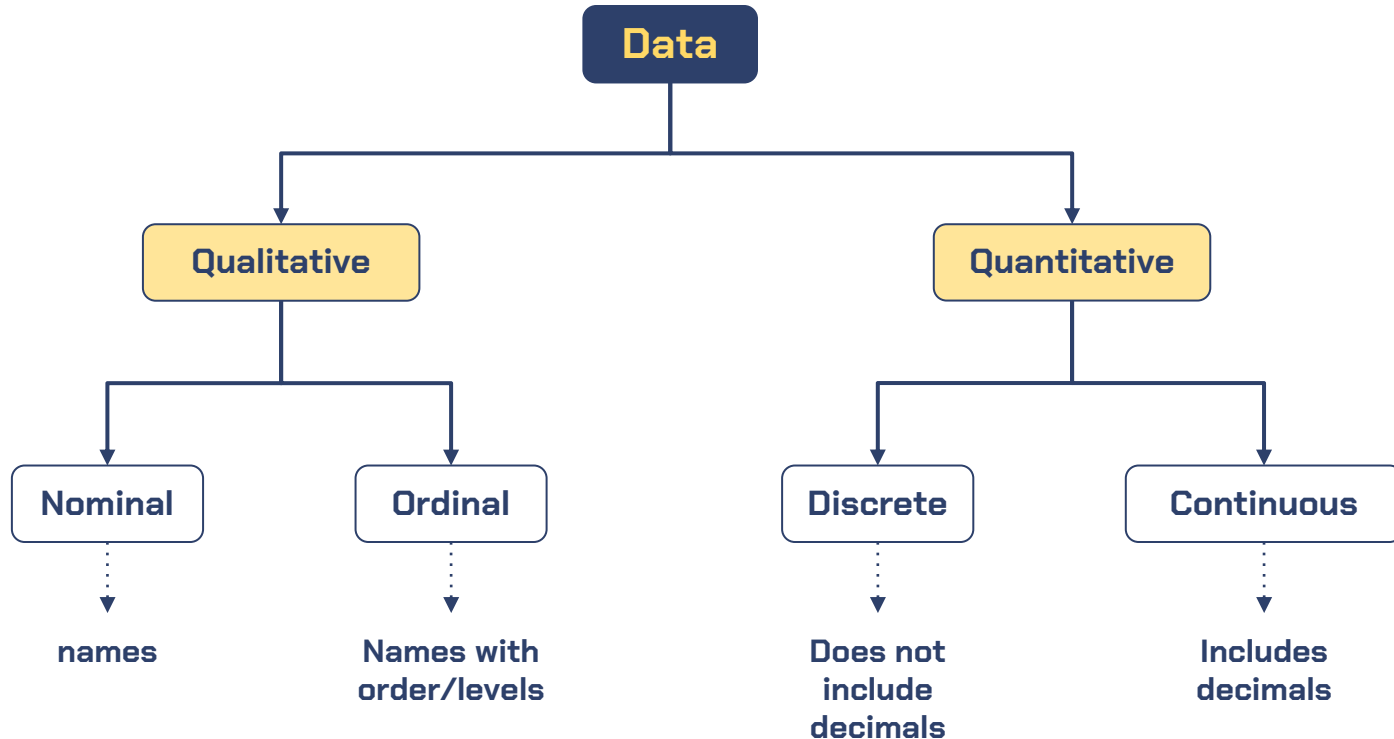
Types of data – based on collection

	2015	2016	2017	2018	2019	2020
Employees	50	55	70	90	160	500
Profit	$1000	$6000	$15,000	$30,000	$80,000	$100,000
Sales	400	800	1000	1500	4000	10000

Time series data ····▶ Data collected over **several time periods**

	2015	2016	2017	2018	2019	2020
Employees	50	55	70	90	160	500

Cross-sectional data

Data collected from **several entities** at the **same time**

	2015
Employees	50
Profit	$1000
Sales	400

Types of data – based on source

Experimental

Introduce a **stimulus** and study the **effects**

Data is collected for controlled variables (*stimulus*) and variable of interest (*effects*)

Observational

No control over any variables

Example: **Surveys**