



STATISTICS FOR DATA SCIENCE

Types of Data and Experiments

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Types of Data and Types of Experiments

D. Uma

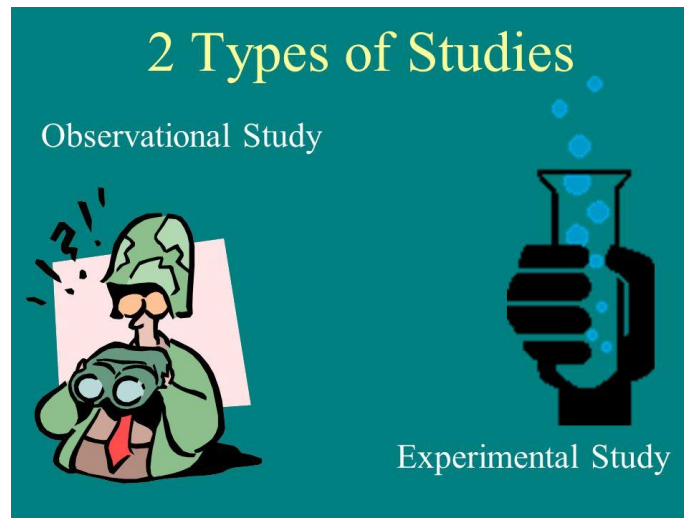
Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- TYPES OF DATA

- TYPES OF STUDY



There are various types of scientific studies such as

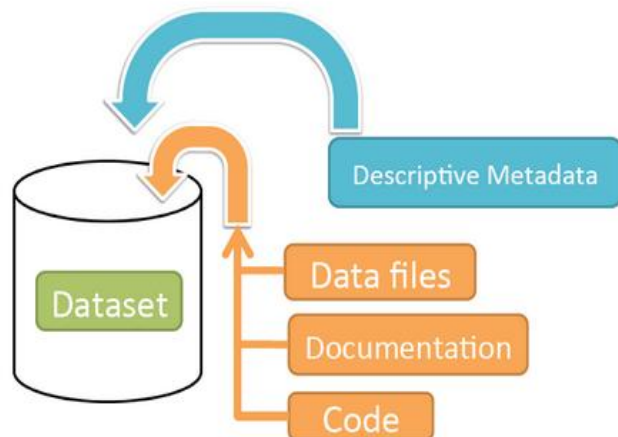
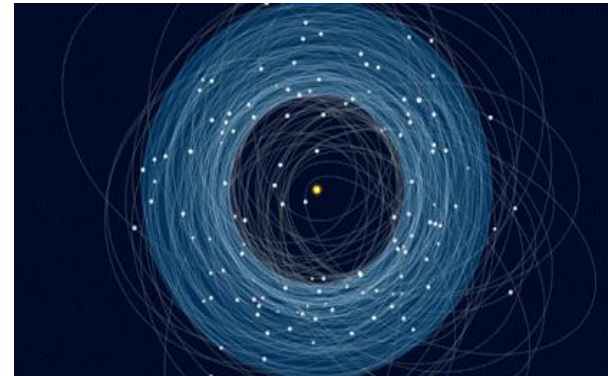
- Experimental Study
 - Observational Study
 - Surveys
 - Interviews
 - Comparative Study etc.
-
- What is the cause of the condition?
 - What is the natural cause of disease if left untreated?
 - What will change because of the treatment?
 - How many people have the same condition?



Data are the **facts and figures** collected, summarized, analyzed and interpreted.

DATA

The **data collected** in a **particular study** are referred to as the **data set**.



Container for your data, documentation, and code.

Attribute (or **variables**, **features**, **dimensions**) is a data field, representing a **characteristic** or **feature** of a **data object**.

Example : Name, Age, Student-ID, address, Marks, Gender.

- **Types:**
Nominal, Binary, Interval-scaled, Ratio-scaled.



STATISTICS FOR DATA SCIENCE

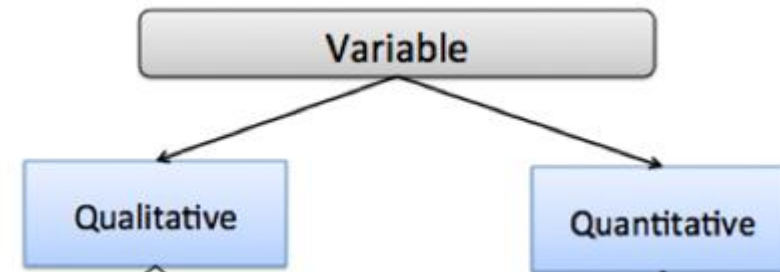
Quantitate vs. Qualitative

A **variable** that **can be measured numerically** is called a **quantitative variable**.

The **data collected** on a quantitative variable are called **quantitative data**.

A **variable** that **can't assume a numerical value** but can be classified into two or more **nonnumeric categories** is called a qualitative or **categorical variable**.

The **data collected** on such a variable are called **qualitative data**.



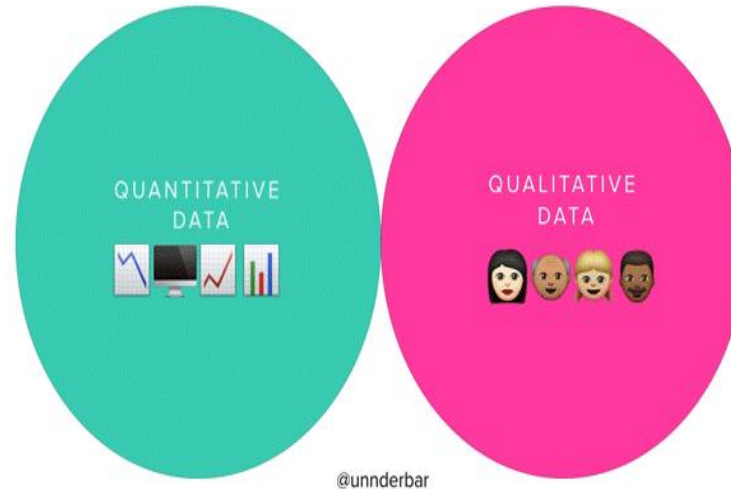
STATISTICS FOR DATA SCIENCE

Quantitate vs. Qualitative

Quantitative Data are measurements that are **recorded** on a naturally occurring **numerical scale**.

Example:

- Age
- GPA
- Salary
- Cost of books



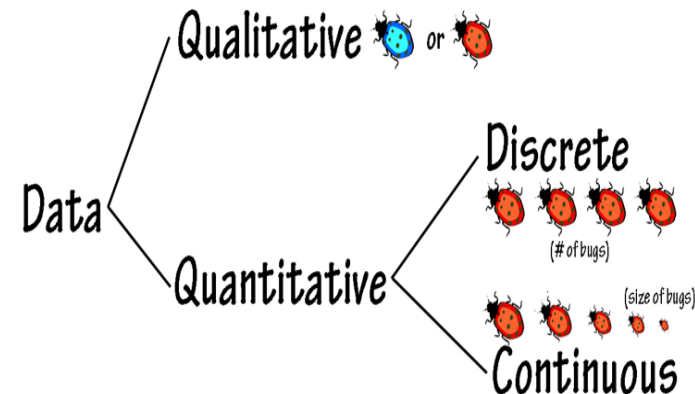
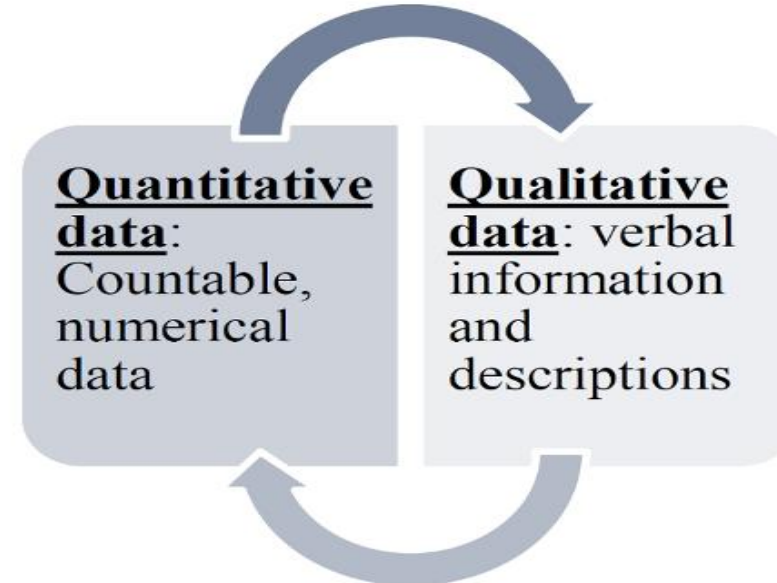
Qualitative Data are measurements that **cannot be recorded** on a natural numerical scale, but are **recorded in categories**.

Example:

- Major
- Gender
- Live on/off campus
- Moving Ratings

STATISTICS FOR DATA SCIENCE

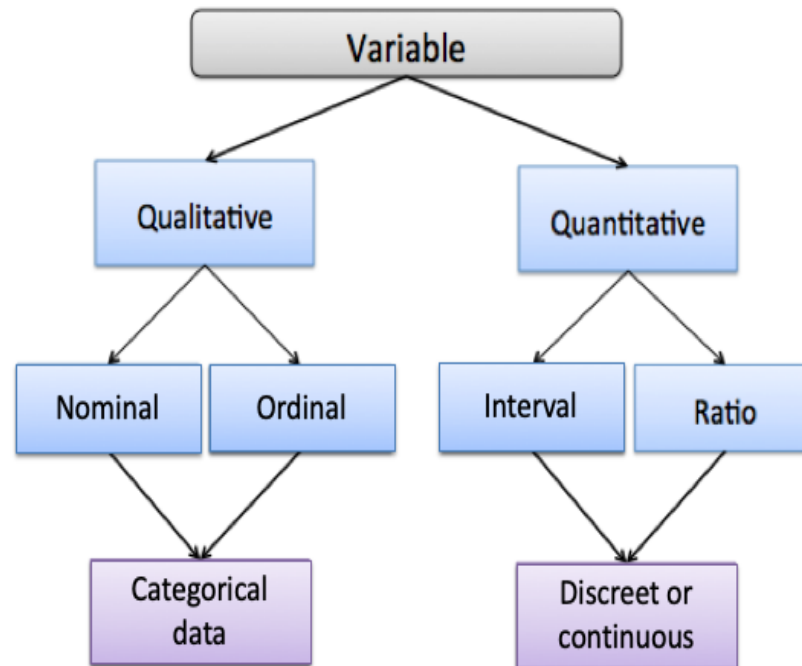
Discrete vs. Continuous



Based on their **mathematical properties**, data are divided into four groups :

NOIR

- Nominal
- Ordinal
- Interval
- Ratio



They are **ordered** with their increasing

- Accuracy
- Powerfulness of measurement
- Preciseness
- Wide application of statistical techniques

STATISTICS FOR DATA SCIENCE

Nominal Data

Nominal means **name and count**; data are alphabetic or numerical in name only.

They are **categories without order** or direction.

Their **use is restricted** to keeping track of people, objects and events.

They are **least powerful** in measurement with no arithmetic origin or order.

Hence, nominal data is of restricted or limited use.



Nominal Data

Used to label variables without providing quantitative values. It can't be ordered. It can't be manipulated using mathematical operators. It can be visualized using pie chart.

Nominal data can be both quantitative and qualitative. Quantitative labels lack a relationship.

How to analyze Nominal Data? : Using grouping method. Group them into categories. For each category, frequency or percentage can be calculated.

Hypothesis testing is carried out using nonparametric tests such as Chi-Square test.

To determine whether there is a significant difference between the expected frequency and the observed frequency.



Gender, marital status or any alphabetic / numeric code without intrinsic order or ranking.

Sl. No.	Subject	Code
1	Physics	P
2	Chemistry	C
3	Mathematics	M
4	Biology	B

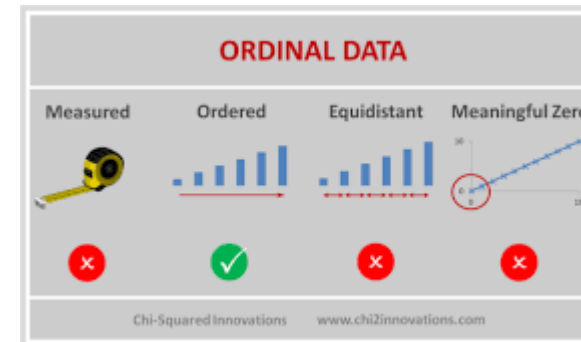
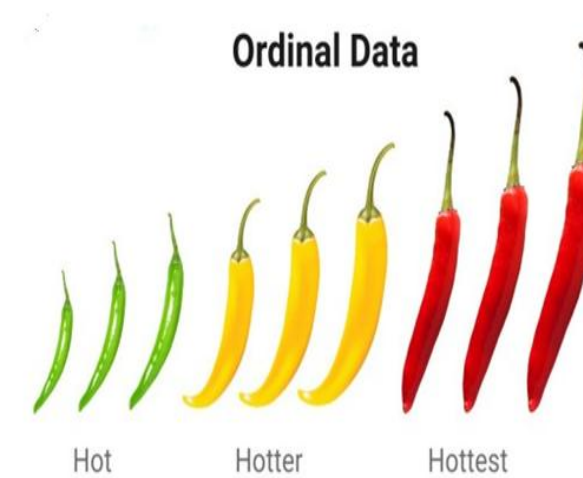
Ordinal means **rank** or **order**.

Ordinal data place events in order.
They are **ordered categories** like rankings or scaling.

Ordinal data **allows** for setting up **inequalities** and nothing much.

Has **no absolute value**(only **relative position in the inequality**)

More precise comparisons are **not possible**.



STATISTICS FOR DATA SCIENCE

Ordinal Data

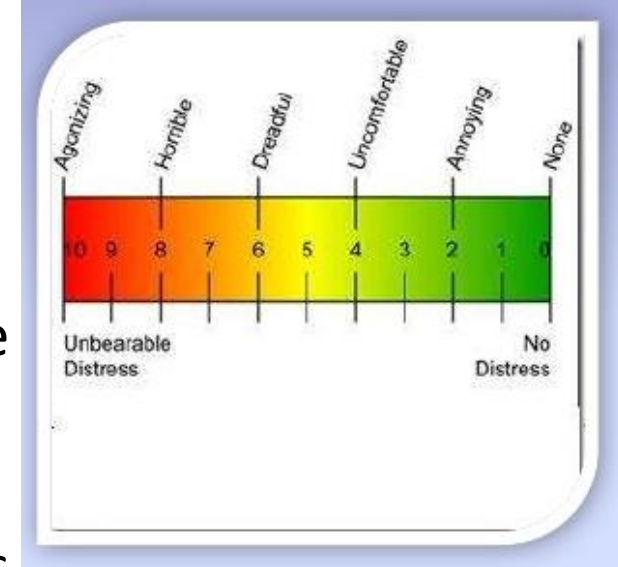
A ordinal variable, is one where the **order matters but not the difference between values.**

Example: **Pain Scales**

Patients are asked to express the amount of pain they are feeling on a scale of 1 to 10.

A score of 7 means more pain than a score of 5, and that is more pain than a score of 3.

But the difference between the 7 and the 5 may not be the same as that between 5 and 3. The values simply express an order.



Ordinal Data - Example

How satisfied are you with our meal tonight?

- ☐ Very satisfied
- ☐ Satisfied
- ☐ Indifferent
- ☐ Dissatisfied
- ☐ Very dissatisfied

In which category do you fall?

- ☐ Child
- ☐ Teenager
- ☐ Youth
- ☐ Middle Age
- ☐ Old

Interval data in addition to ranking(setting up inequalities) further allow for forming **differences**.

For interval data there is **no absolute zero**; **unique origin does not exist**.

Interval data are **more powerful** than ordinal scale due to **equality of intervals**.

Examples:

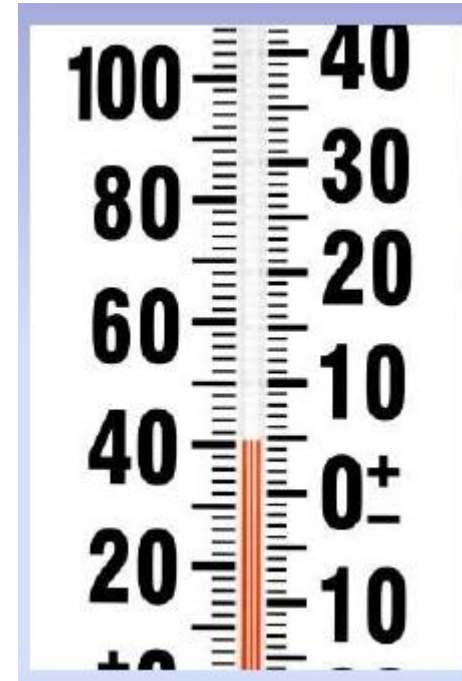
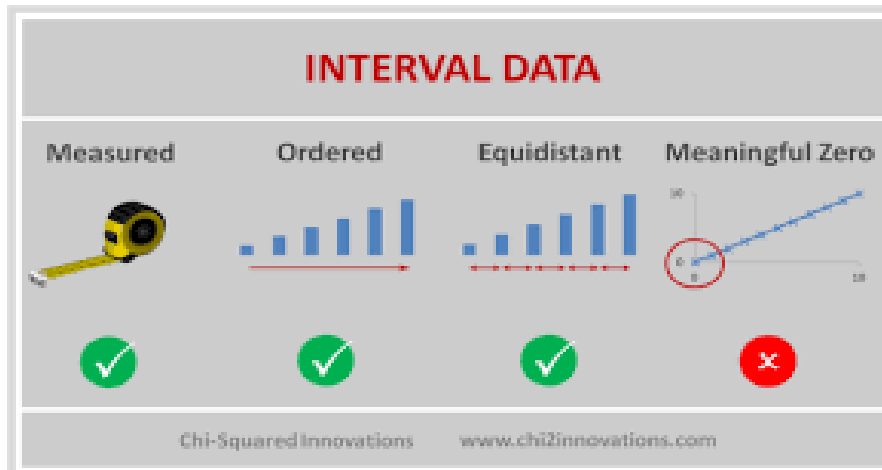
Temperature in Fahrenheit,
Standardised scores.

STATISTICS FOR DATA SCIENCE

Interval Data

An **Interval** variable is a **measurement** where the **difference between two values** is meaningful.

The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.



Ratio data allow for forming quotients in addition to setting up inequalities and forming differences.

All mathematical operations(manipulations with real numbers) are possible on ratio data.

It can have an absolute or true zero and represent the actual amount/ value.

The most precise data and allow for application of all statistical techniques.

Examples: Height, weight, age.

A **ratio** variable, has all the properties of an interval variable, and also has a clear definition of 0.0.

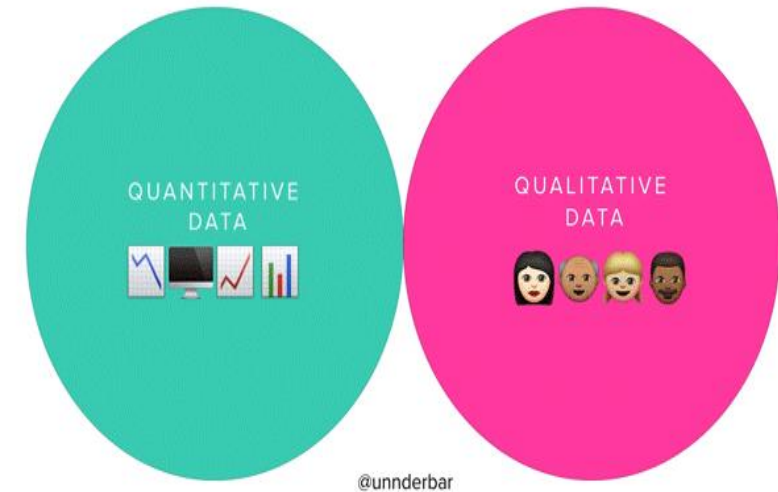
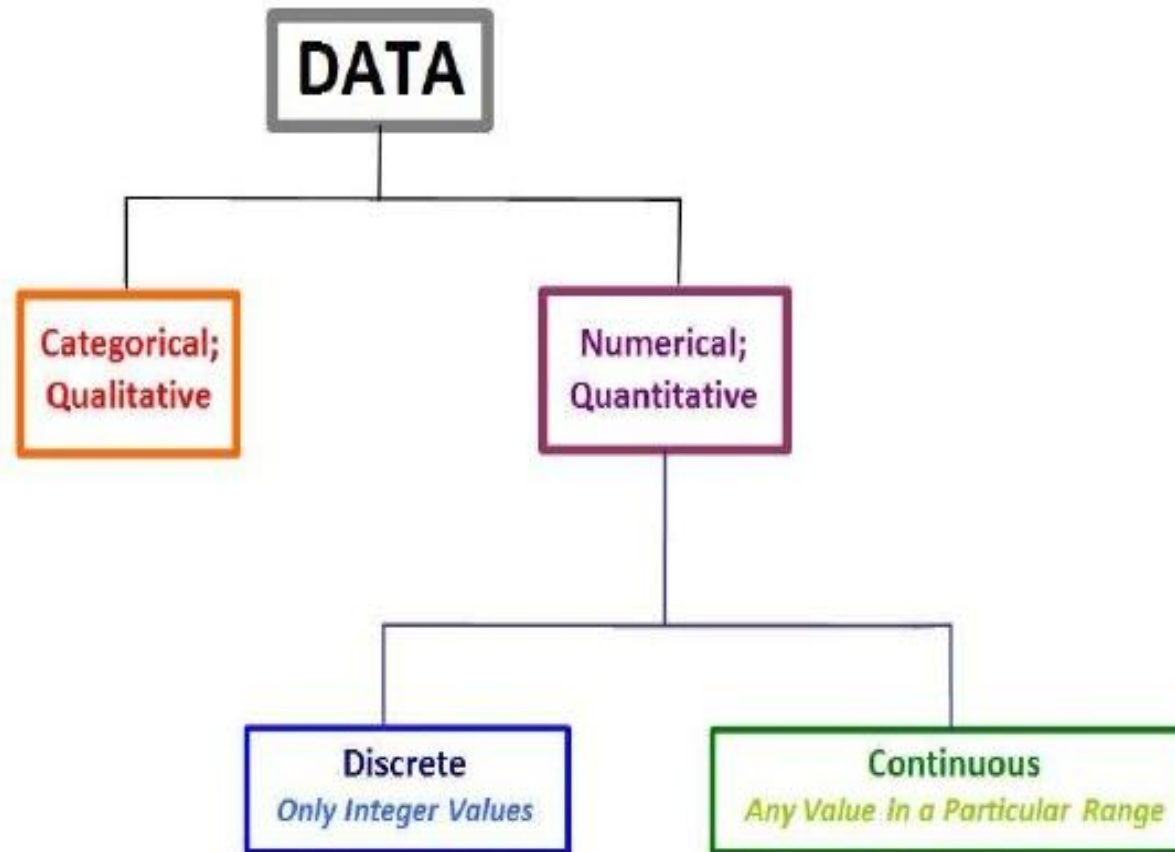
When the variable equals 0.0, there is none of that variable.

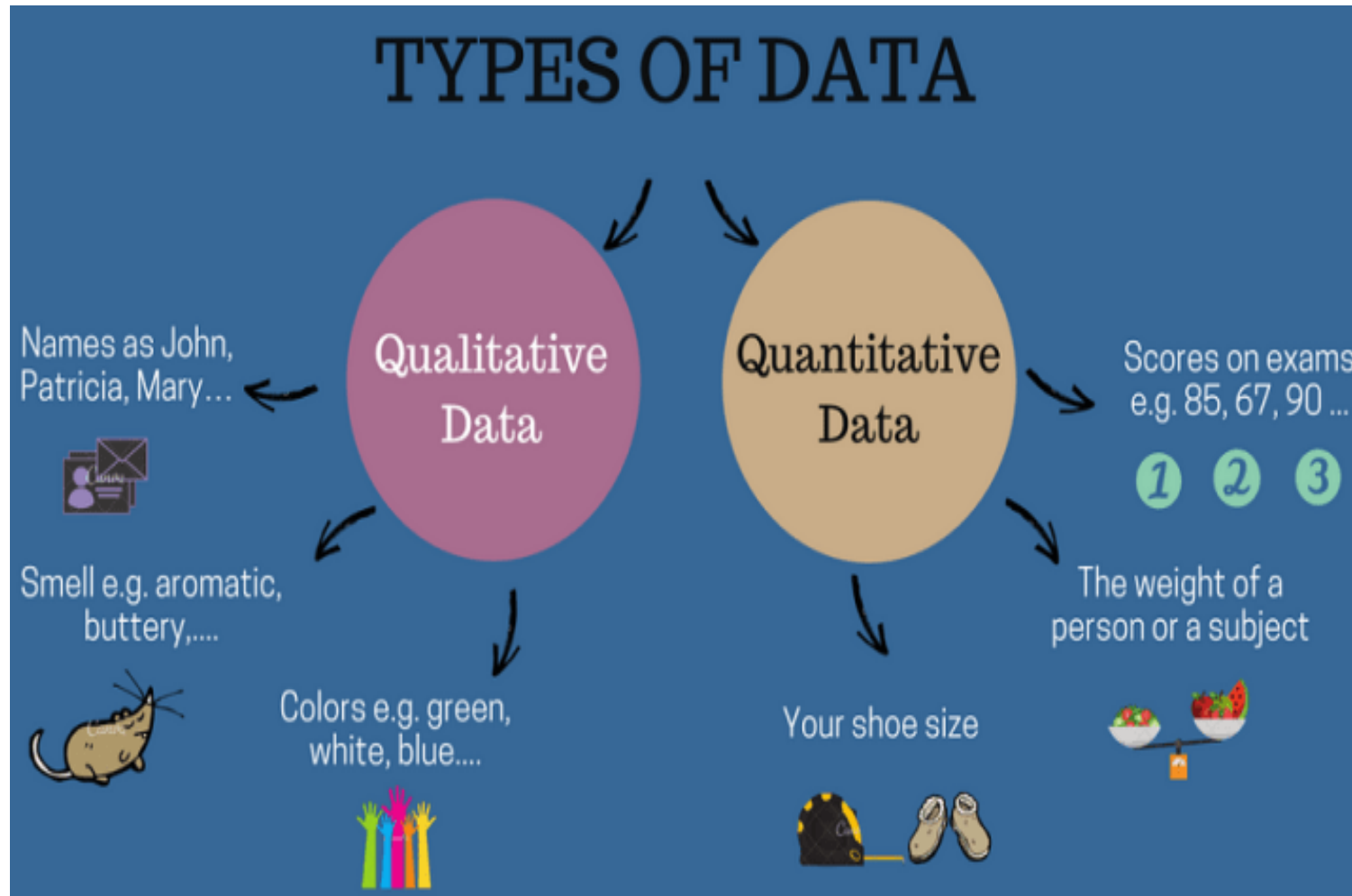


STATISTICS FOR DATA SCIENCE

Ratio Data - Example

Roll No.	Name	Gender	Rank	Height	Weight In Kgs
1	Amar	M	9	4' 8"	51
2	Asha	F	1	3' 10"	39
3	Bhaskar	M	5	4' 5"	48
4	Chandru	M	3	4' 3"	41





Numerical data could be either **discrete or continuous**.

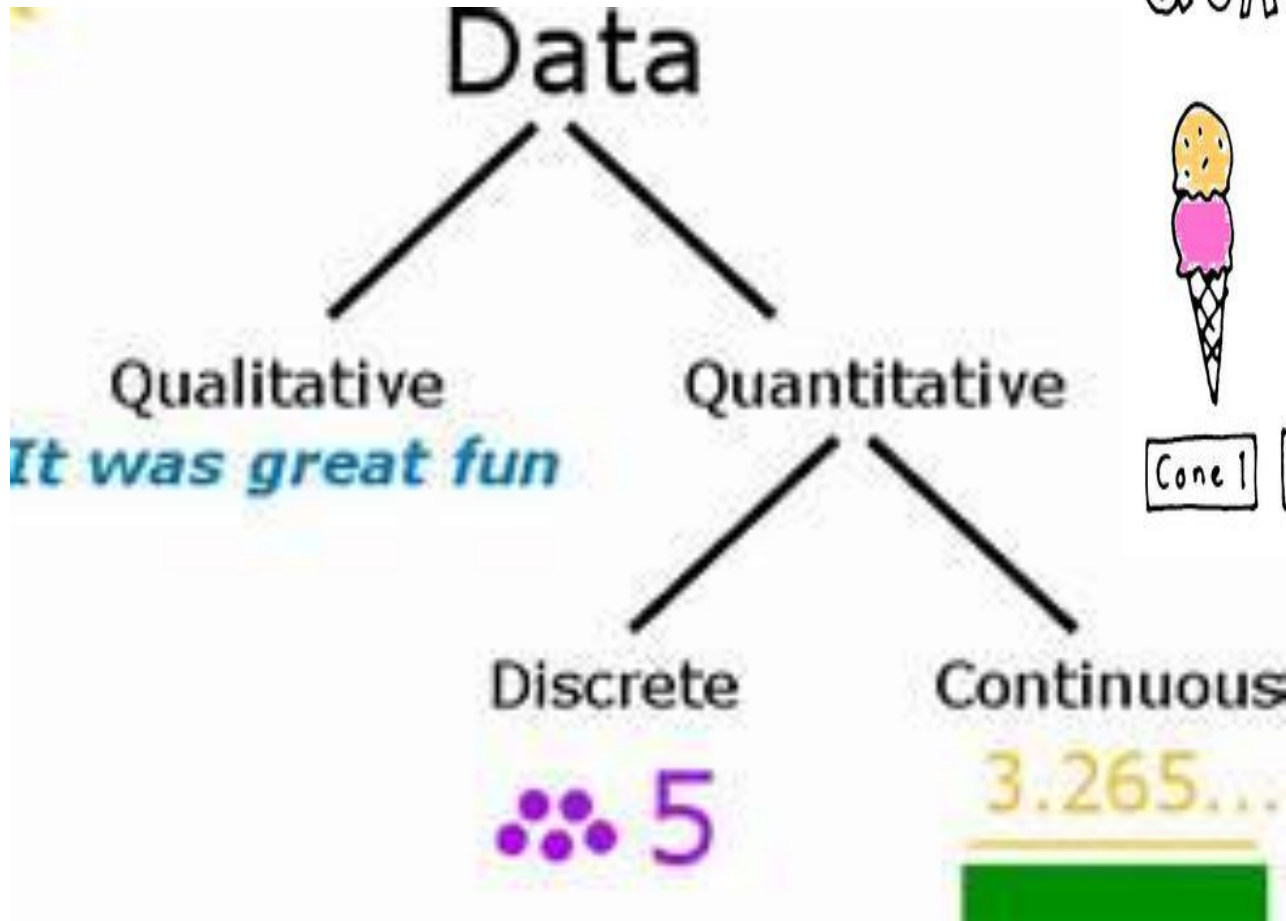
Continuous data can take any **numerical value**(**within a range**).

Examples: Height, weight, age.

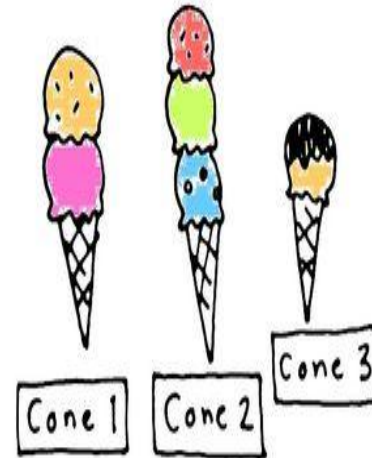
There can be an **infinite number of possible values** in continuous data.

Discrete data can take **only certain values** by a **finite 'jumps'** i.e. It 'jumps' from one value to another **but does not take any intermediate value** between them.

Example: Number of students, Number of accidents.



QUANTITATIVE DATA:

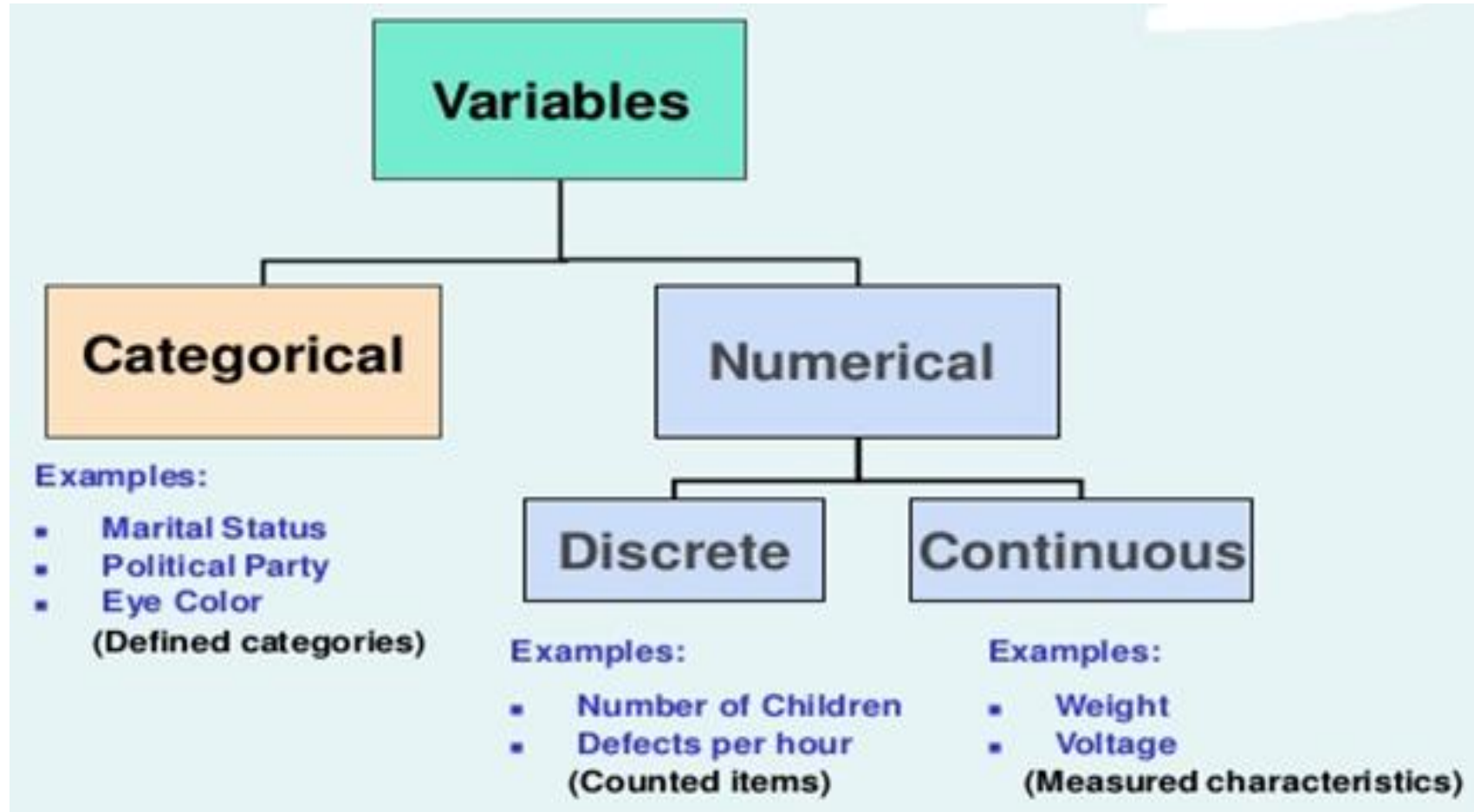


Discrete data:

- There are 3 cones
- Cone 1 has 2 scoops

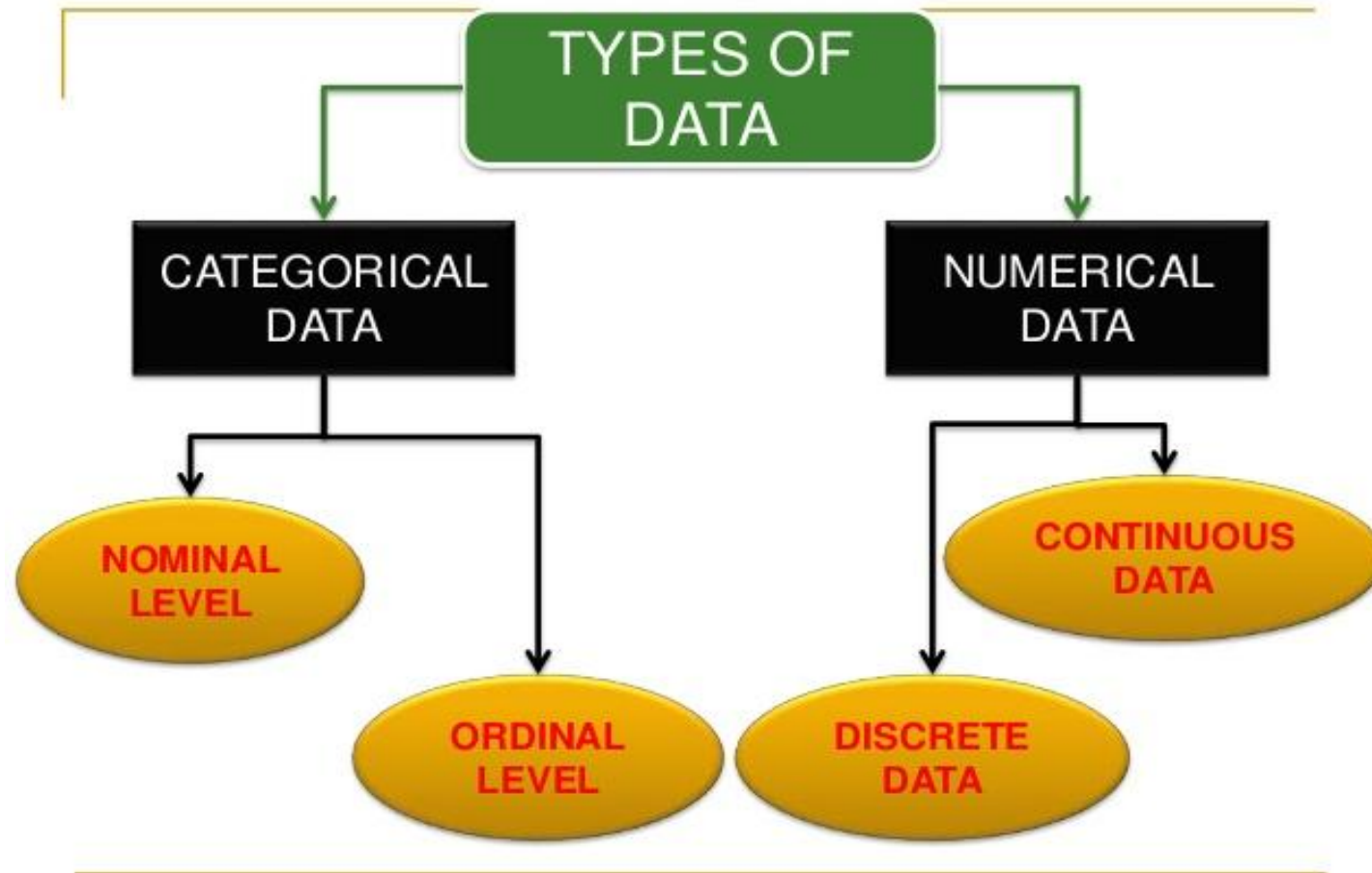
Continuous data:

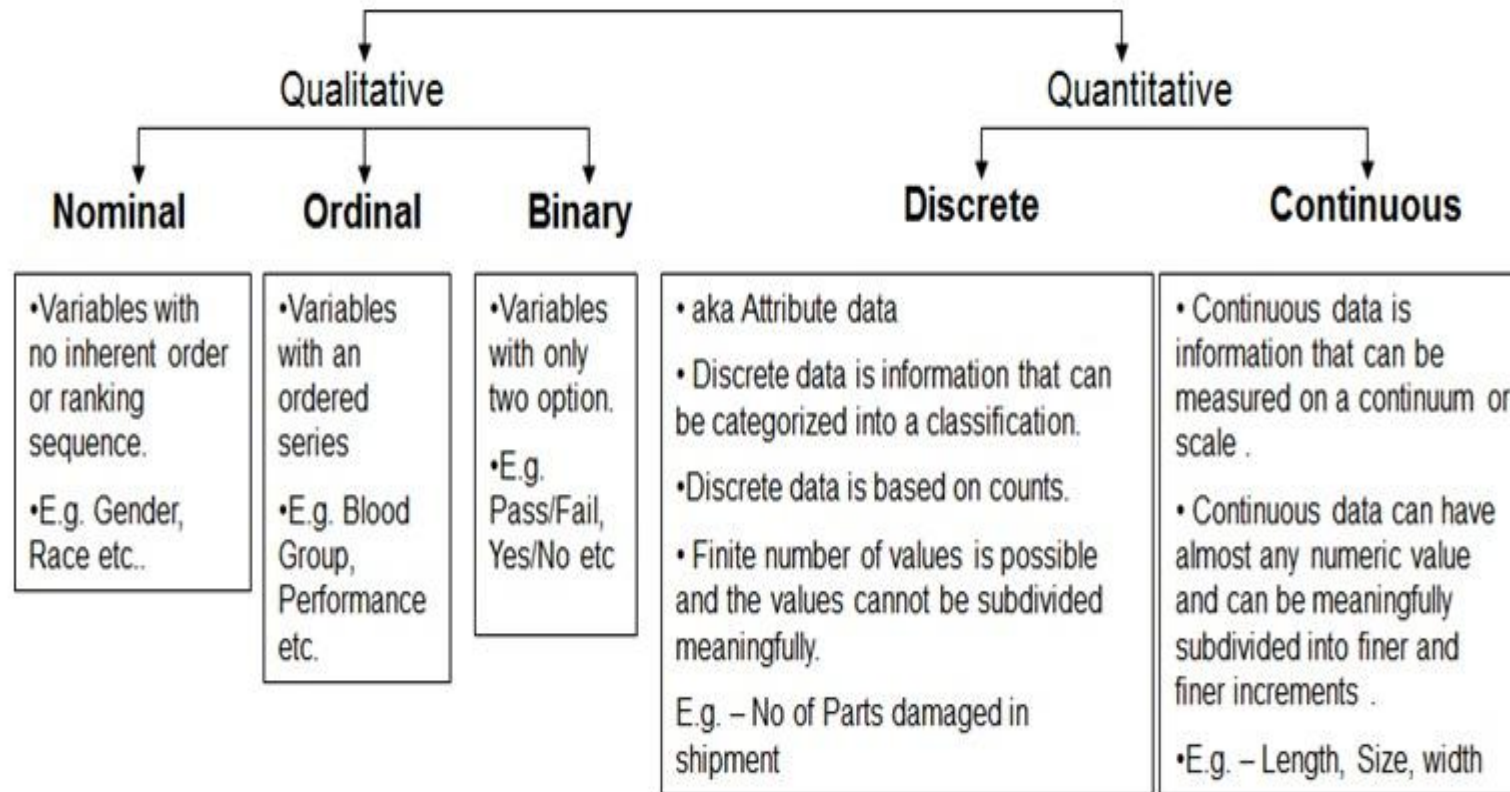
- Cone 3 weighs 79.4 grams
- cone 2 ice cream is at 8.3°F

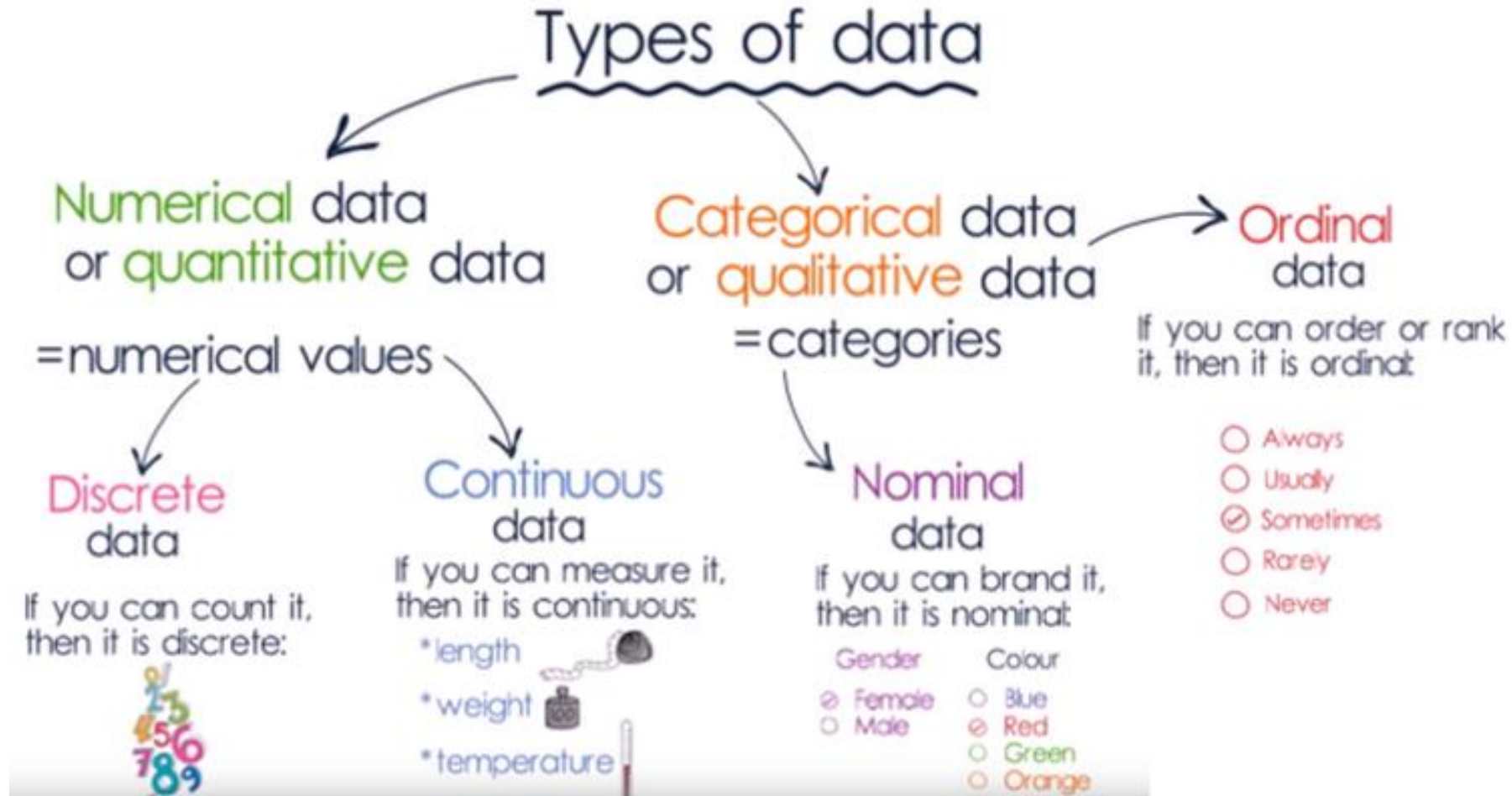


Qualitative Data	Quantitative Data
<p>Overview:</p> <ul style="list-style-type: none">• Deals with descriptions.• Data can be observed but not measured.• Colors, textures, smells, tastes, appearance, beauty, etc.• Qualitative → Quality	<p>Overview:</p> <ul style="list-style-type: none">• Deals with numbers.• Data which can be measured.• Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc.• Quantitative → Quantity







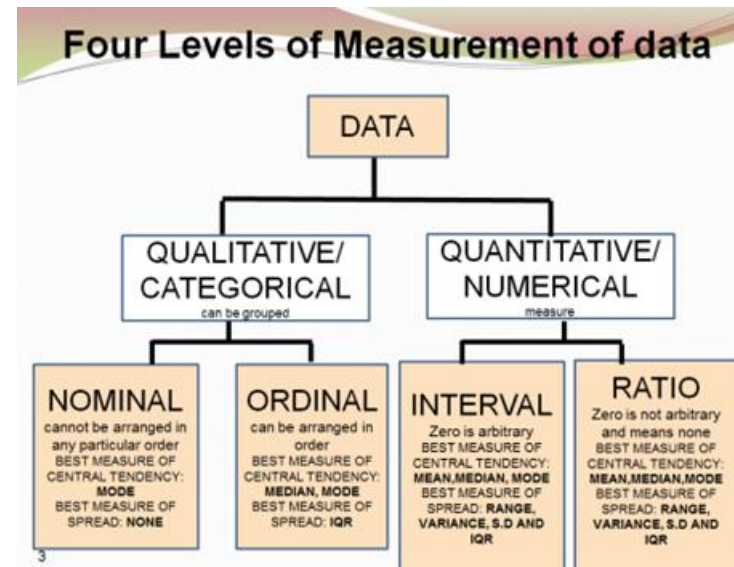
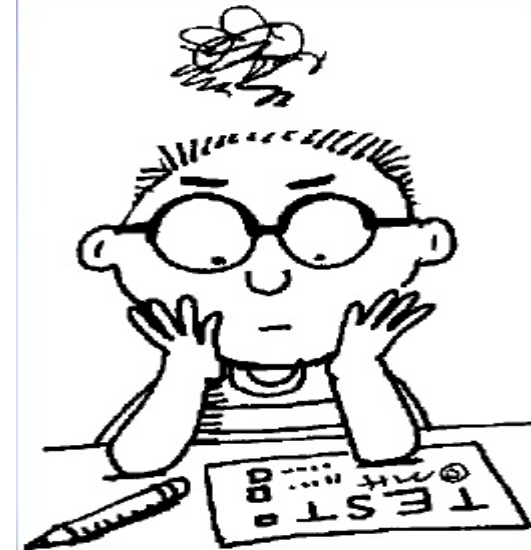
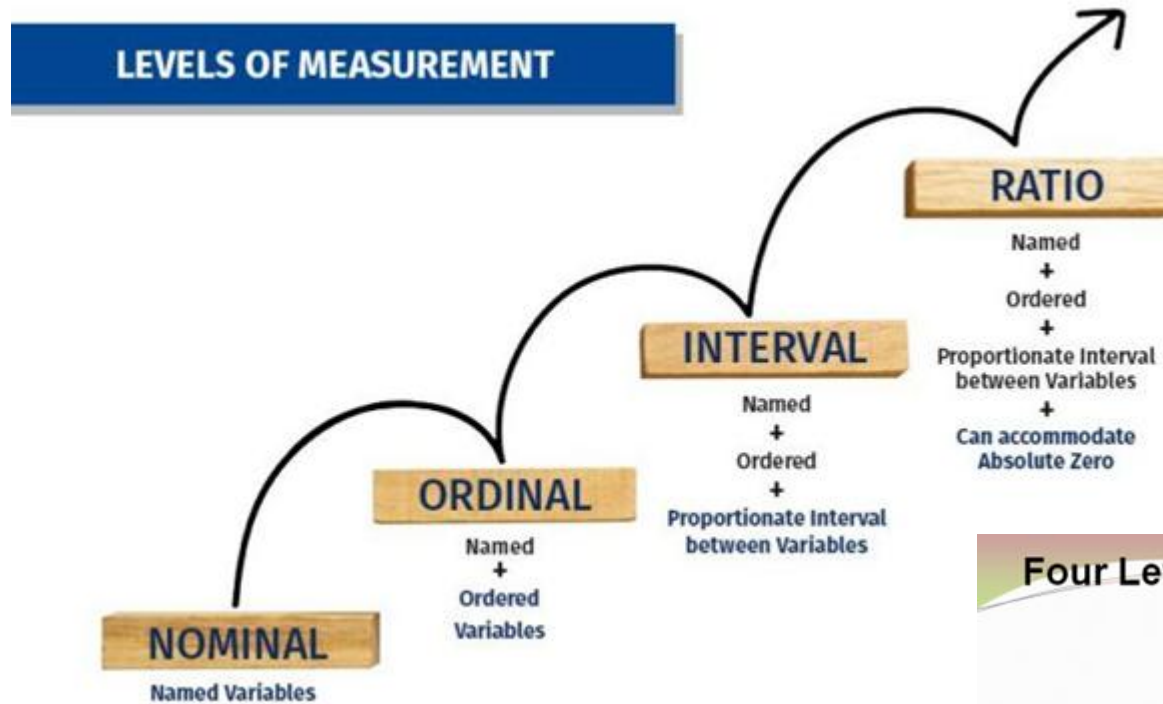


STATISTICS FOR DATA SCIENCE

Summary

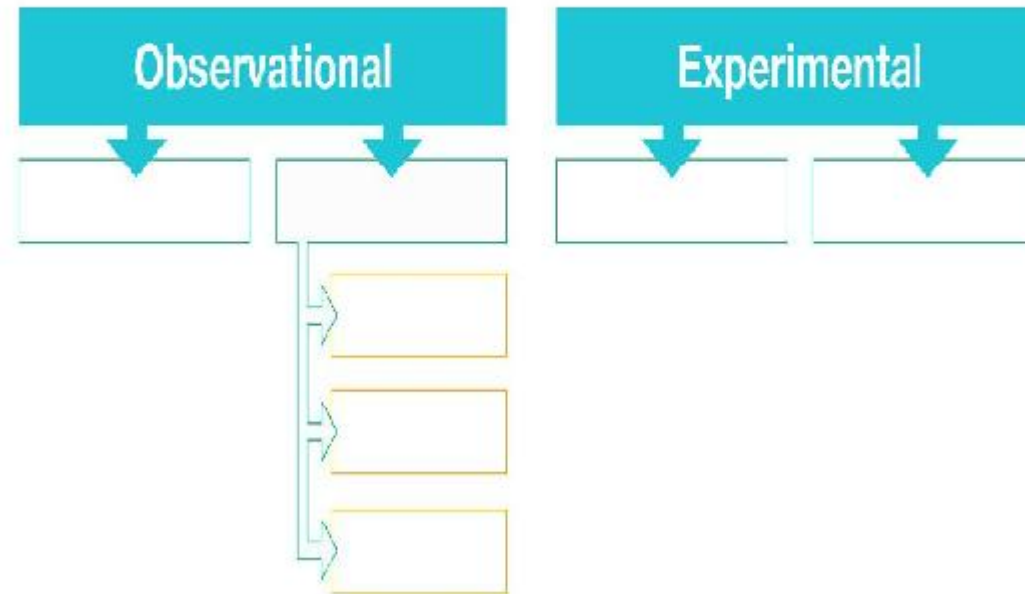


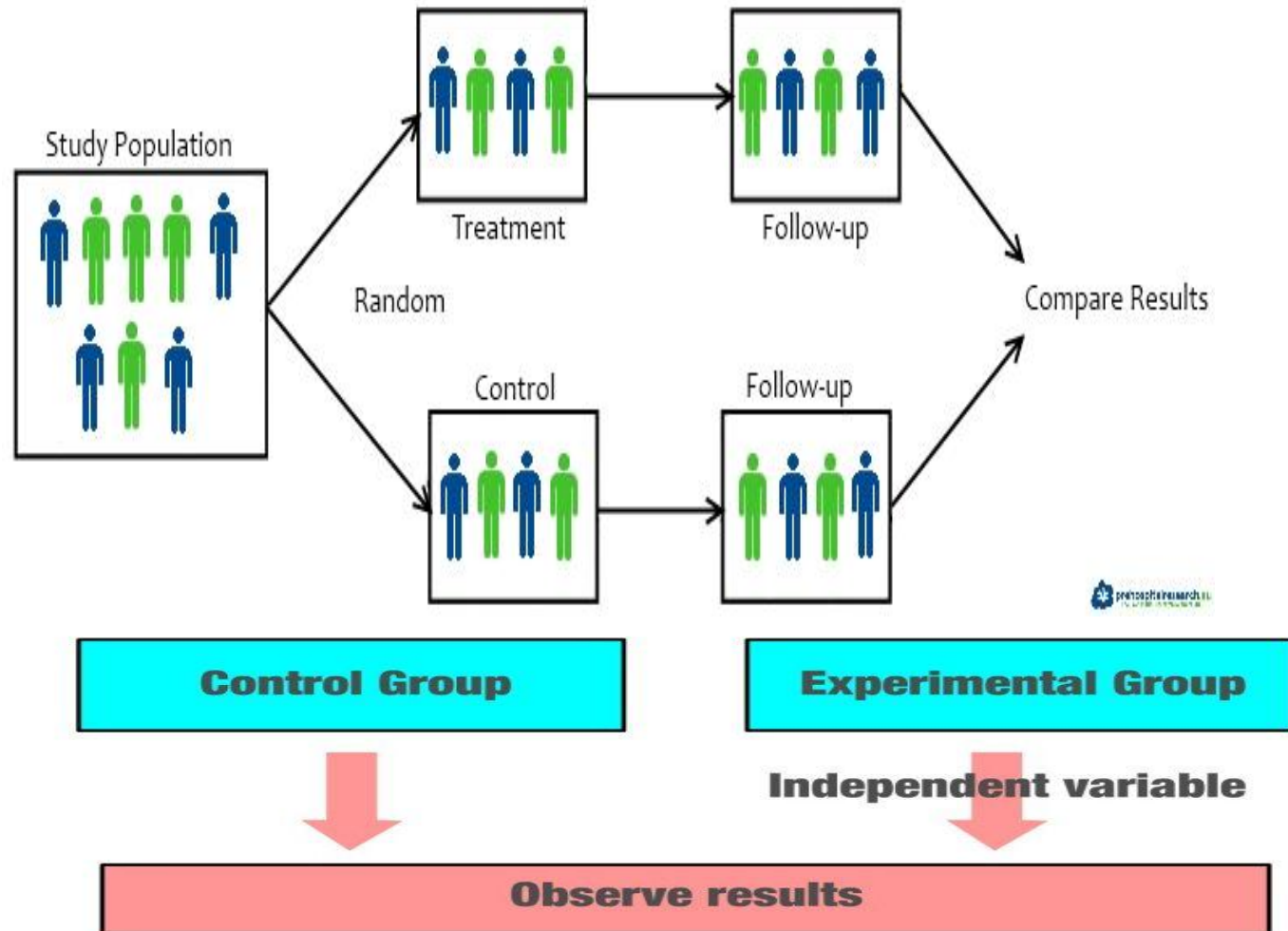
PES
UNIVERSITY
ONLINE

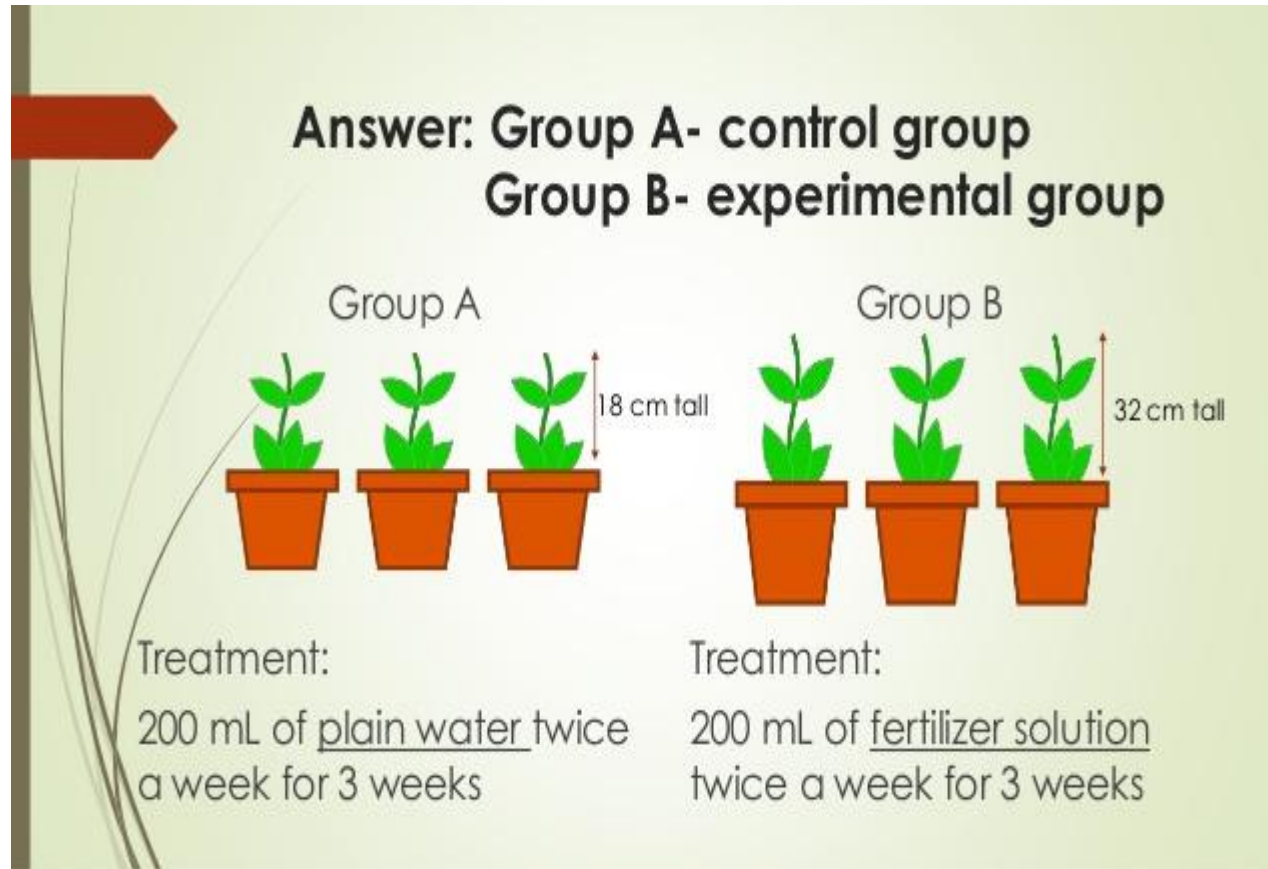


- Number of cartons of milk manufactured each day.
- Temperatures of airplane interiors at a given airport.
- College major of each student in a class.
- Method of payment
- Incomes of college students on work study programs.
- Weights of newborn calfs.
- Gender of each employee at a company.
- Number of tomatoes on each plant in a field.
- Number of defective items in a lot.
- Salaries of CEOs of oil companies.

Types of Studies

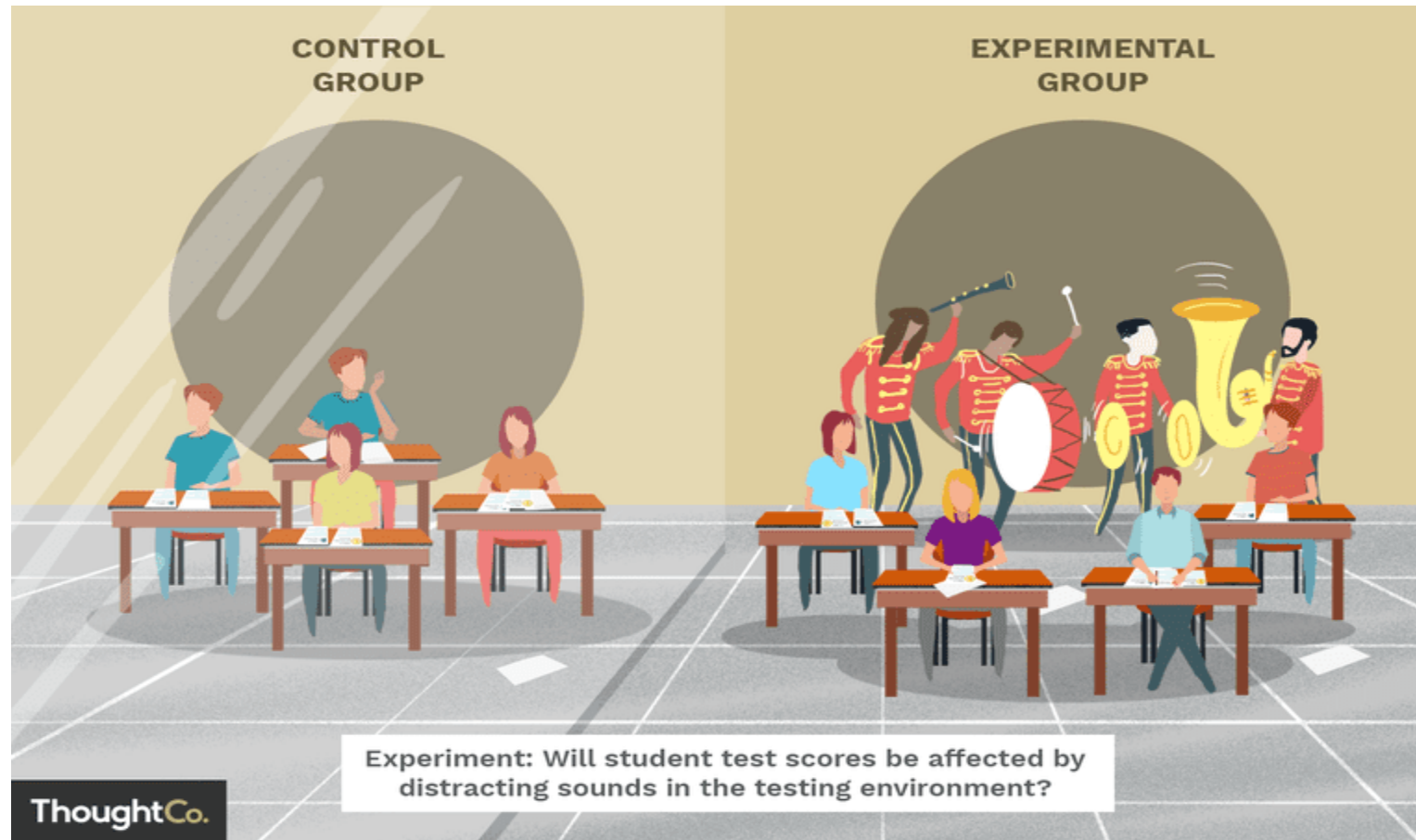






STATISTICS FOR DATA SCIENCE

Types of Experiments



STATISTICS FOR DATA SCIENCE

Controlled Experiments

Unregistered FrontCam


- what is different about the groups
- what you are testing

dependent

- data
- what you measure

control

- used for



1. What is the independent variable?
color of light

2. What is the dependent variable?
rate (speed) of plant growth

3. Which group is the control group?

4. What are the constants?

Observational Study	Experimental Study
Observe only, no “treatment” assigned.	“Treatment” assigned.
Generally a control group is not needed.	Uses control group for comparison.
Reports an association.	Report a cause and effect.
May (or not) use random sample sets.	Randomization of sample group.
May (or not) generalize to population.	Generalize to population.

Example

A study took random sample of adults and asked them about their bedtime habits. The data showed that people who drank a cup of tea before bedtime were more likely to go to sleep earlier than those who didn't drink tea.

Answer : Observation Study

Example

A study took a group of adults and randomly divided them into two groups. One group was told to drink tea every night for a week, while the other group was told not to drink tea that week. Researchers then compared when each group fell asleep.

Answer : Experimental Study

Example



A study randomly assigned volunteers to one of two groups:

One group was directed to use social media sites as they usually do.
One group was blocked from social media sites.

Answer : Experimental Study

Example

A study took a random sample of people and examined their social media habits. Each person was classified as either a light, moderate, or heavy social media user. The researchers looked at which groups tended to be happier.

Answer : Observation Study



THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335