STATISTICS FOR DATA SCIENCE

Types of Data and Experiments

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Types of Data and Types of Experiments

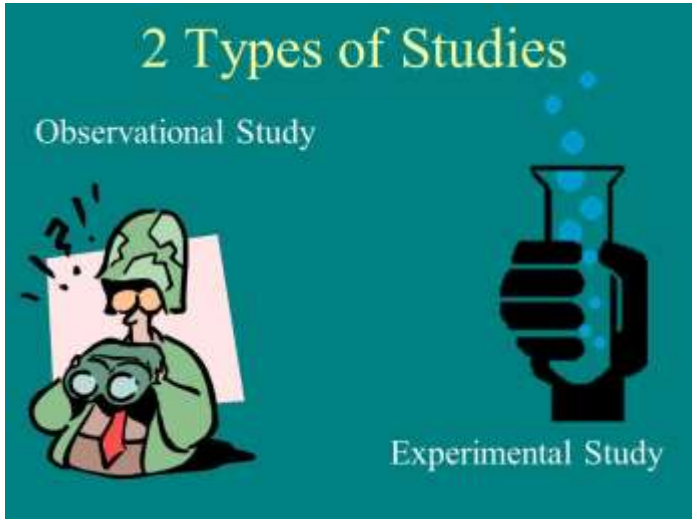
Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- **TYPES OF DATA**

- **TYPES OF STUDY**



Source: www.slideshare.net

Types of Study

There are various types of scientific studies such as

- Experimental Study
- Observational Study
- Surveys
- Interviews
- Comparative Study etc.

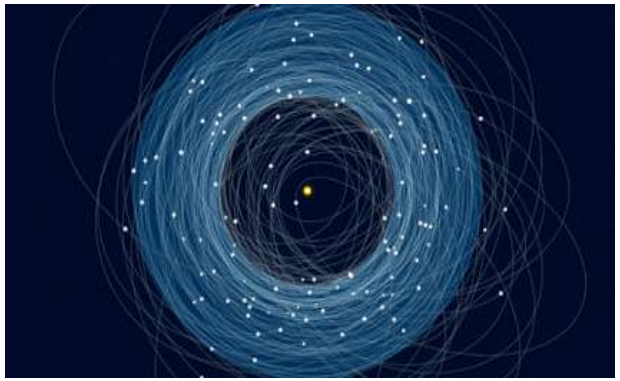
- What is the cause of the condition?
- What is the natural cause of disease if left untreated?
- What will change because of the treatment?
- How many people have the same condition?

Source: www.slideshare.net

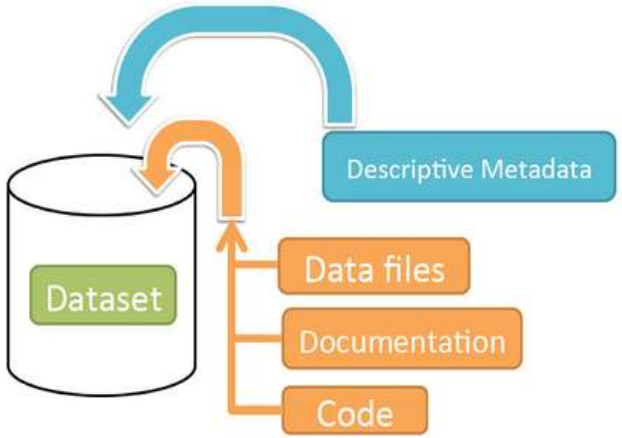
STATISTICS FOR DATA SCIENCE

Data

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



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



Container for your data, documentation, and code.

Sources: guides.dataverse.org, giphy.com, towardsdatascience.com,dribbble.com

STATISTICS FOR DATA SCIENCE

Variable/Attributes

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.

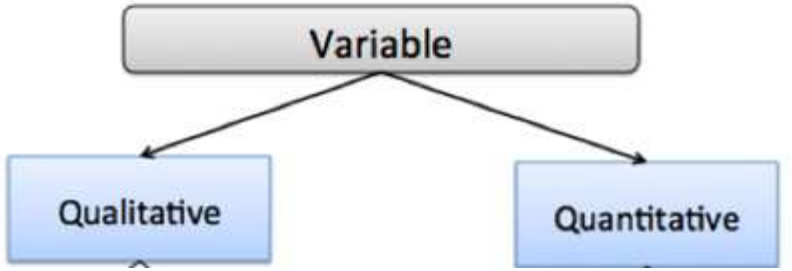
Source: www.slideshare.net

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**.

Source: Wikilean

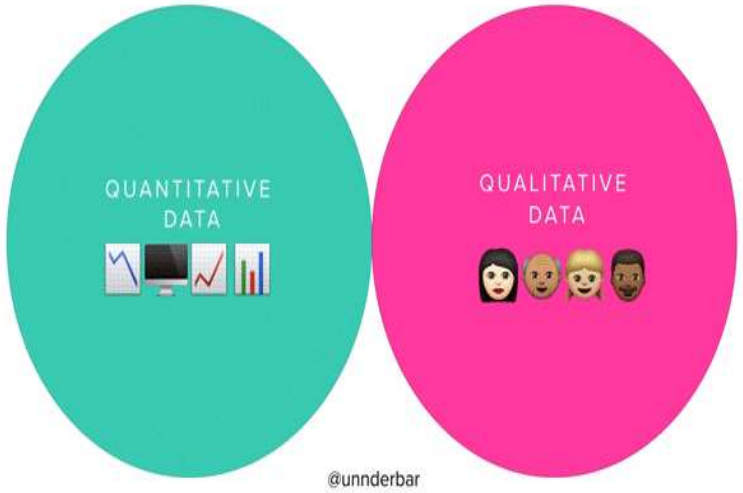
STATISTICS FOR DATA SCIENCE

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

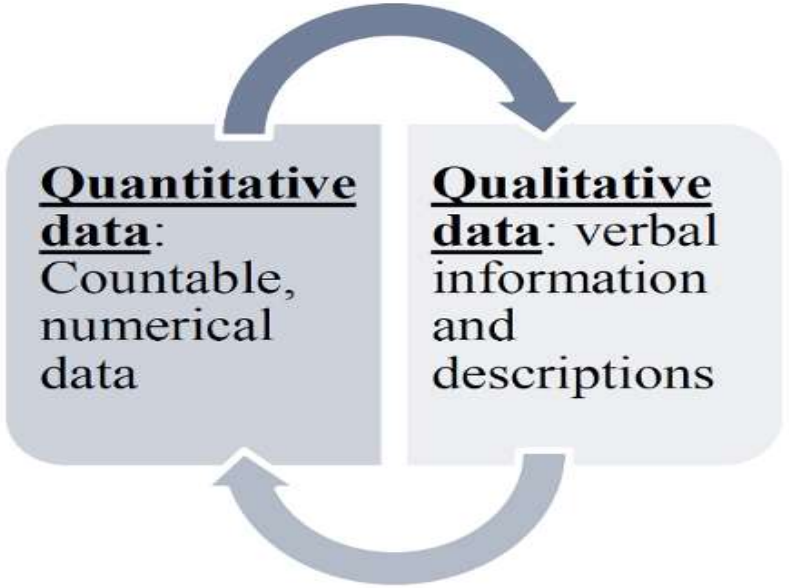
Definition: Text-Book Reference: Section 1.1,Pg.No.11.

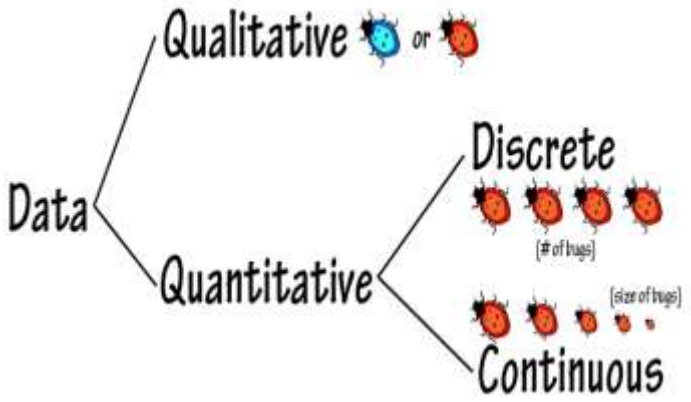
Example: Text-Book Reference: Section 1.1, Example 1.8,Pg.No.11.

Source:www.slideshare.net

STATISTICS FOR DATA SCIENCE

Discrete vs. Continuous





Sources: sheilapontis.wordpress.com, toolkit.nefe.org

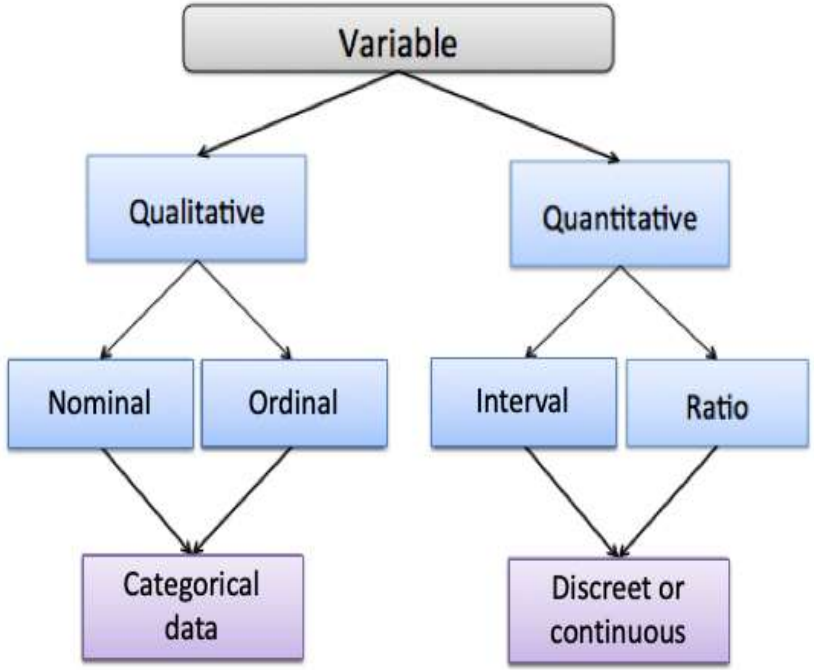
STATISTICS FOR DATA SCIENCE

Types of Data

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

Source: www.slideshare.net, Wikilean

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.

Source: Wikilean, corporatefinanceinstitute.com

STATISTICS FOR DATA SCIENCE

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.

Source: Wikilean, corporatefinanceinstitute.com

STATISTICS FOR DATA SCIENCE

Nominal Data - Examples

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

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE
Ordinal Data

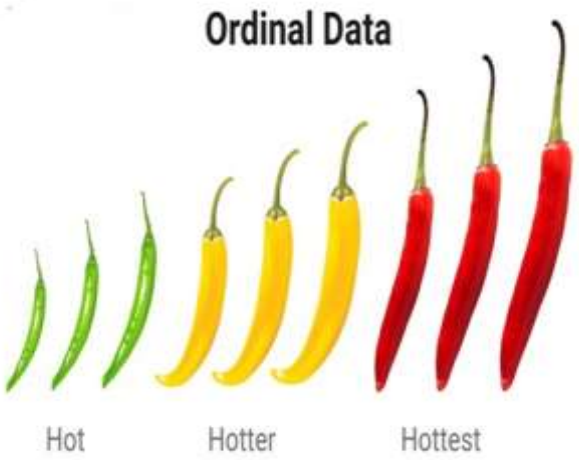
Ordinal means **rank** or **order**.

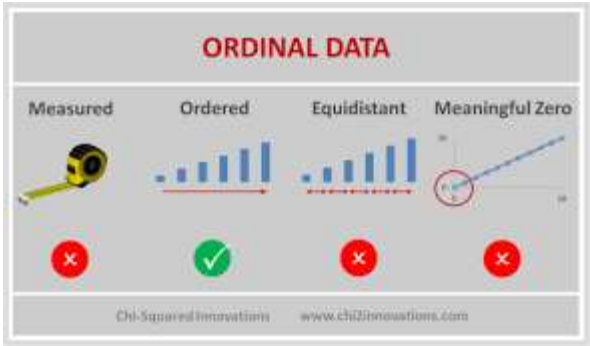
Ordinal data place events in order.
They are ordered categories like rankings or scaling.

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.**



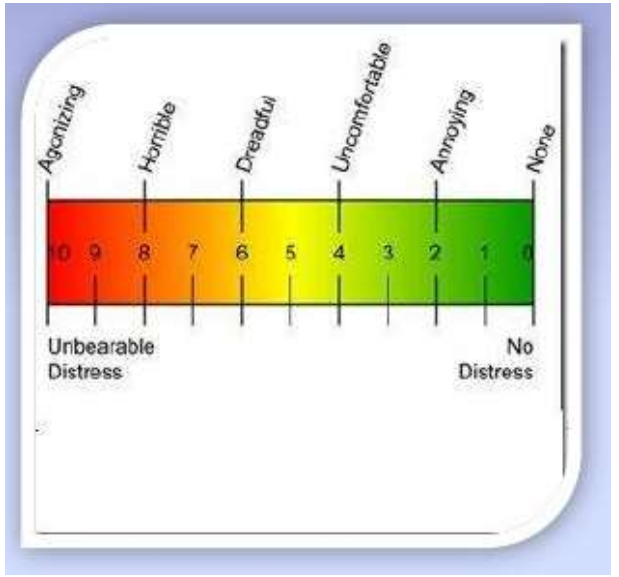


Source: Questionpro, slideshare.net

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 that 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.

Source: Questionpro, slideshare.net

STATISTICS FOR DATA SCIENCE

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

Source: formpl.us

STATISTICS FOR DATA SCIENCE

Interval Data

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:

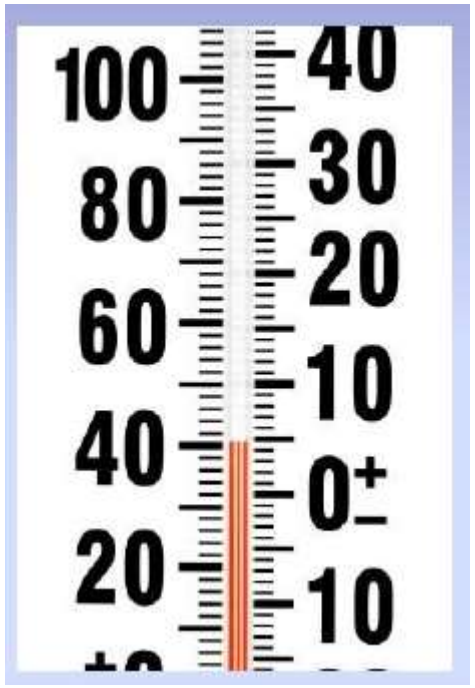
Temperature in Fahrenheit, Standardised scores.

Source: Wikilean, corporatefinanceinstitute.com

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.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Interval Data - Examples

Age is also a variable that can be measured on an interval scale.

For example if A is 15 years old and B is 20 years old, it not only clear than B is older than A, but B is elder to A by 5 years.

One can measure time during the day using a 12-hour clock

- Time in a 12-hour format is a rotational measure that keeps restarting from zero at set periodicity. These numbers are on an interval scale as the distance between them is measurable and comparable.

Source: www.questionpro.com

STATISTICS FOR DATA SCIENCE

Ratio Data

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.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Ratio Data

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.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

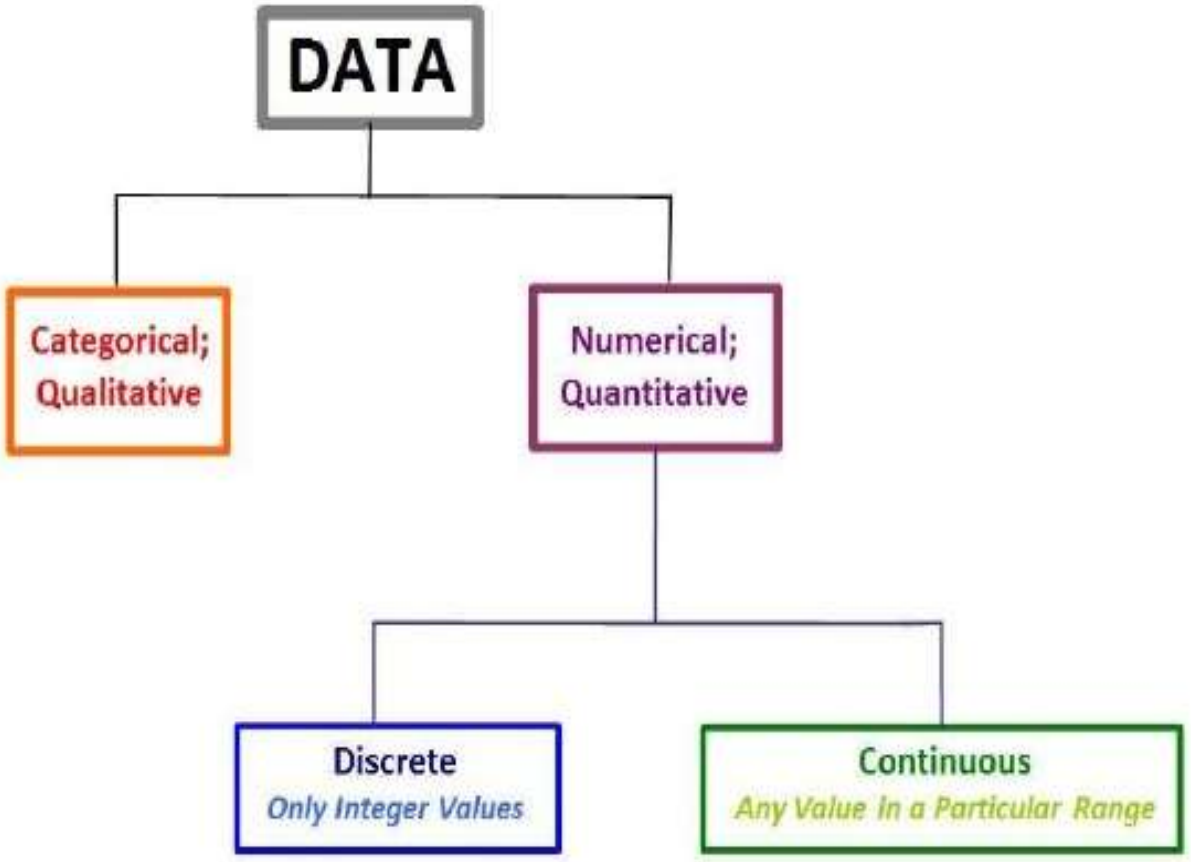
Ratio Data - Example

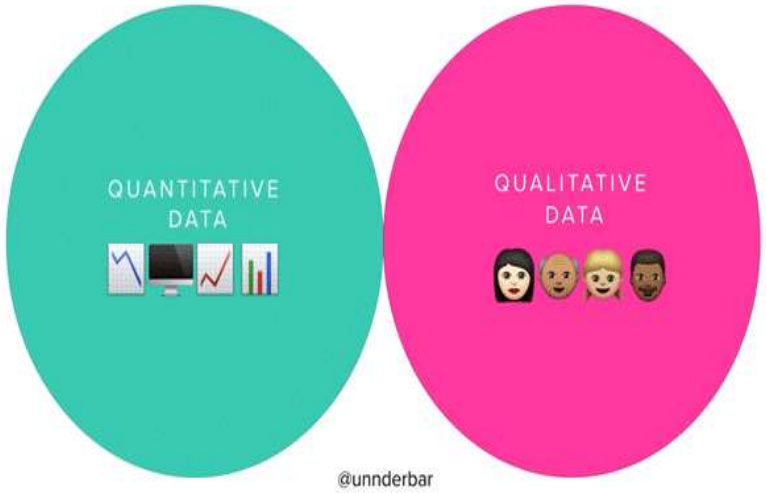
Roll No.	Name	Gender	Rank	Height	Weight In Kgs
1	Amar	M	9	4’ 8”	51
2	Asha	F	1	3’ 10”	39
3	Bhaskar	M	5	4’ 5”	48
4	Chandru	M	3	4’ 3”	41

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Types of Data

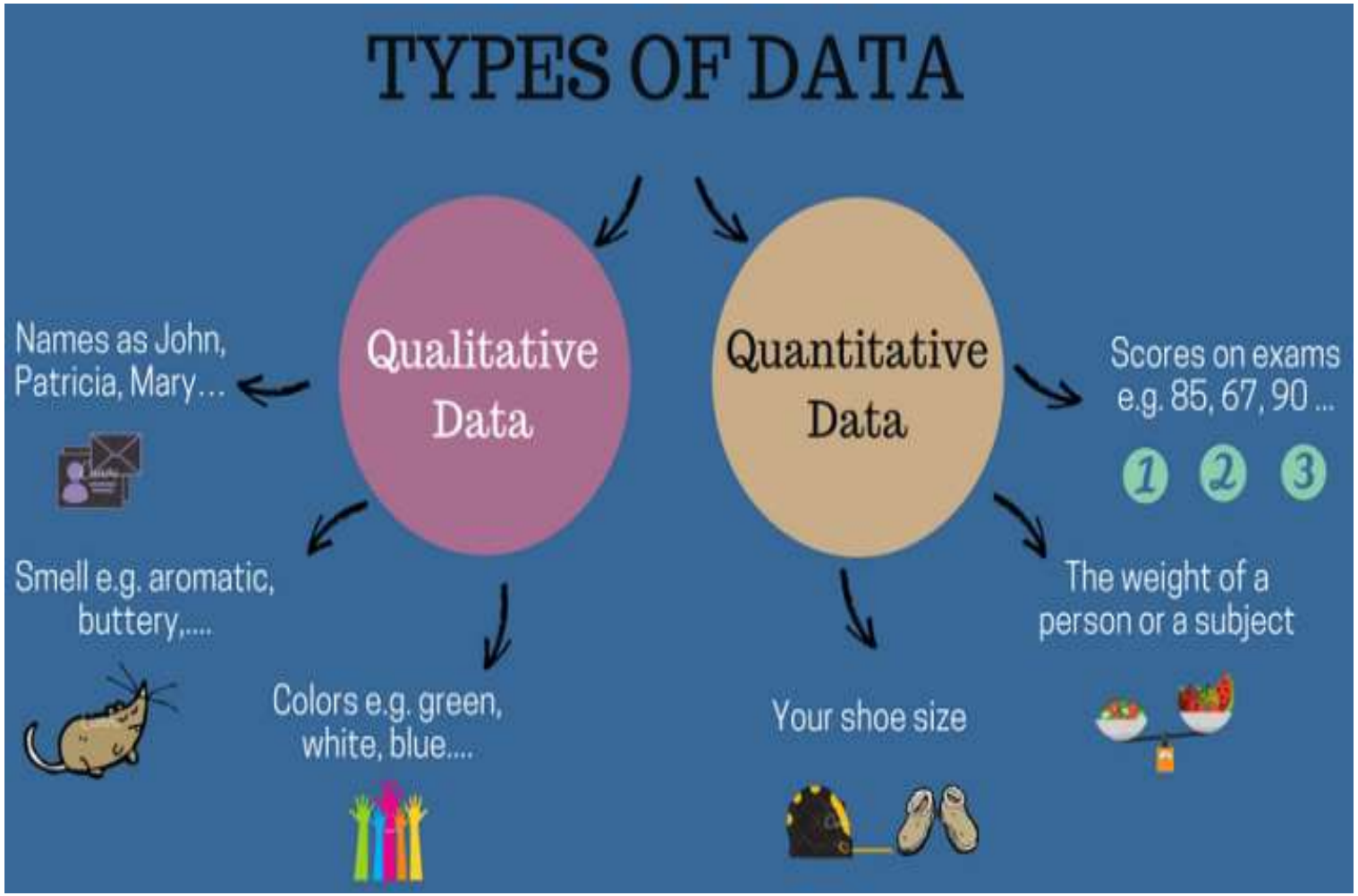




Sources: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Types of Data



Source: www.wordable.io

STATISTICS FOR DATA SCIENCE

Quantitative: Continuous and Discrete

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.

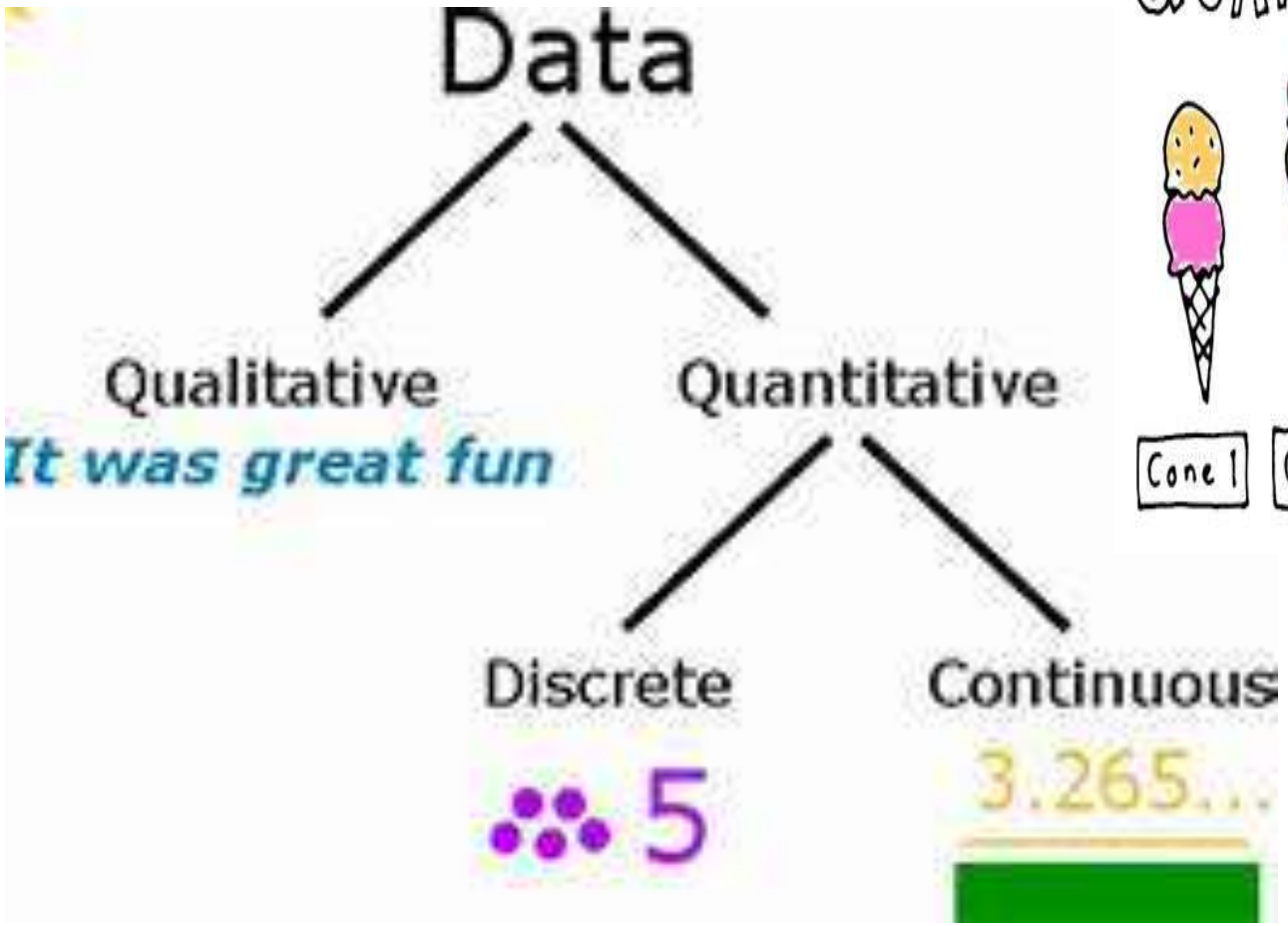
Discrete data can take **only certain values** by a **finite 'jumps'** i.e. It 'jumps' from one value to another **but does not take any intermediate value** between them.

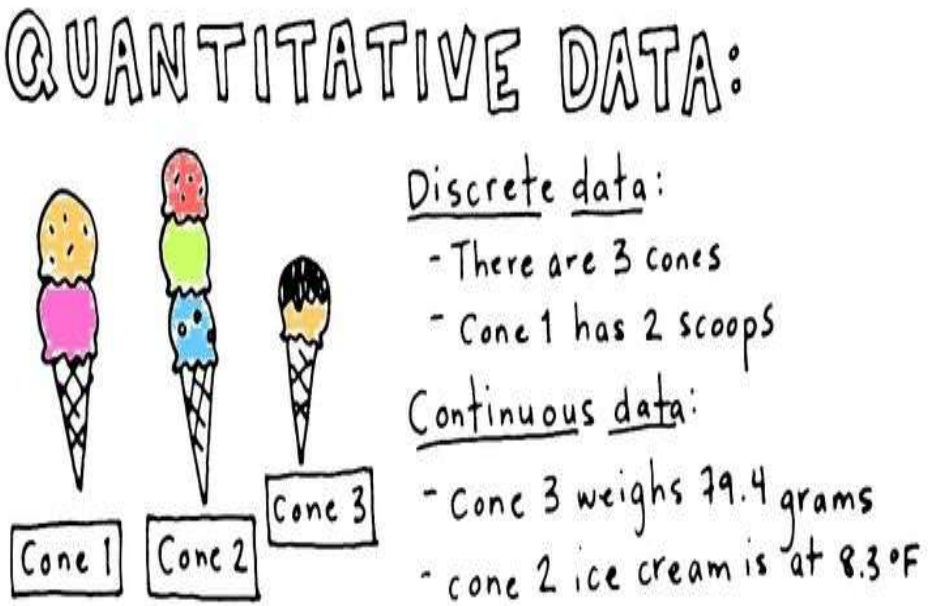
Example: Number of students, Number of accidents.

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Quantitative Data

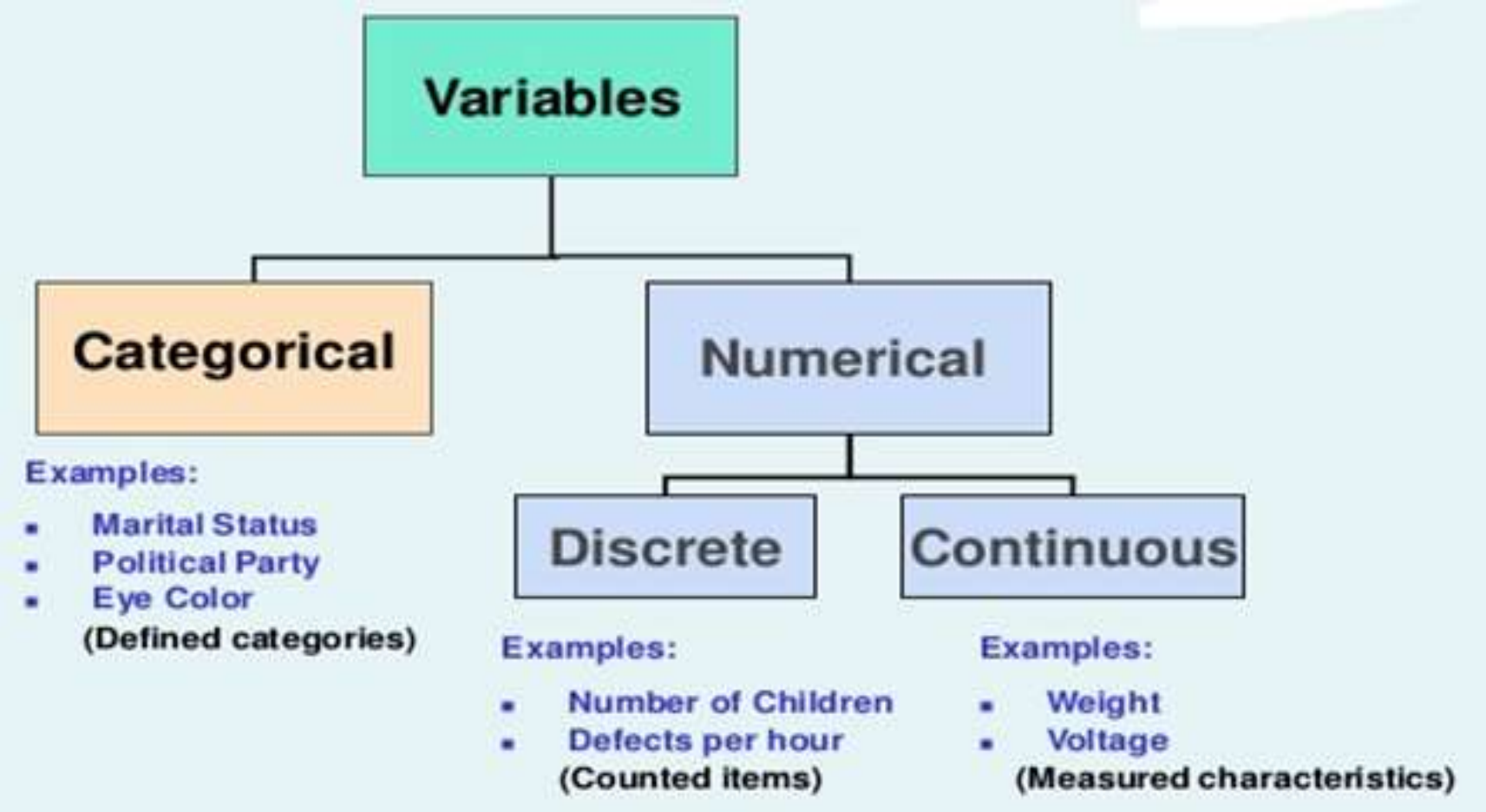




Sources: mathisfun.com,www.pinterest.com

STATISTICS FOR DATA SCIENCE

Types of Data



Sources: sheilapontis.wordpress.com

STATISTICS FOR DATA SCIENCE

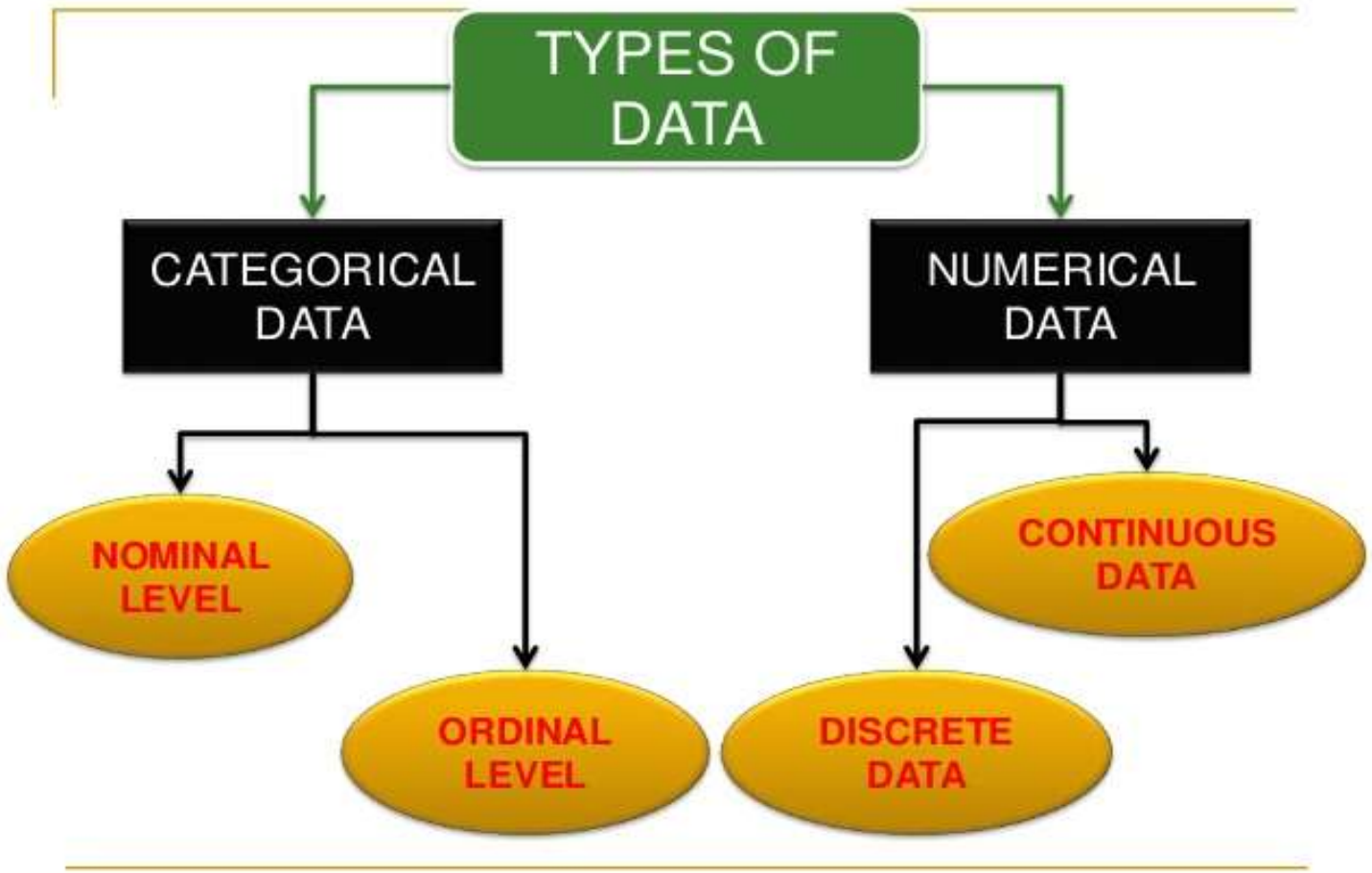
Types of Data

Qualitative Data	Quantitative Data
Overview: • Deals with descriptions. • Data can be observed but not measured. • Colors, textures, smells, tastes, appearance, beauty, etc. • Qualitative → Quality	**Overview:** • 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

Sources: viciousvix.wordpress.com

STATISTICS FOR DATA SCIENCE

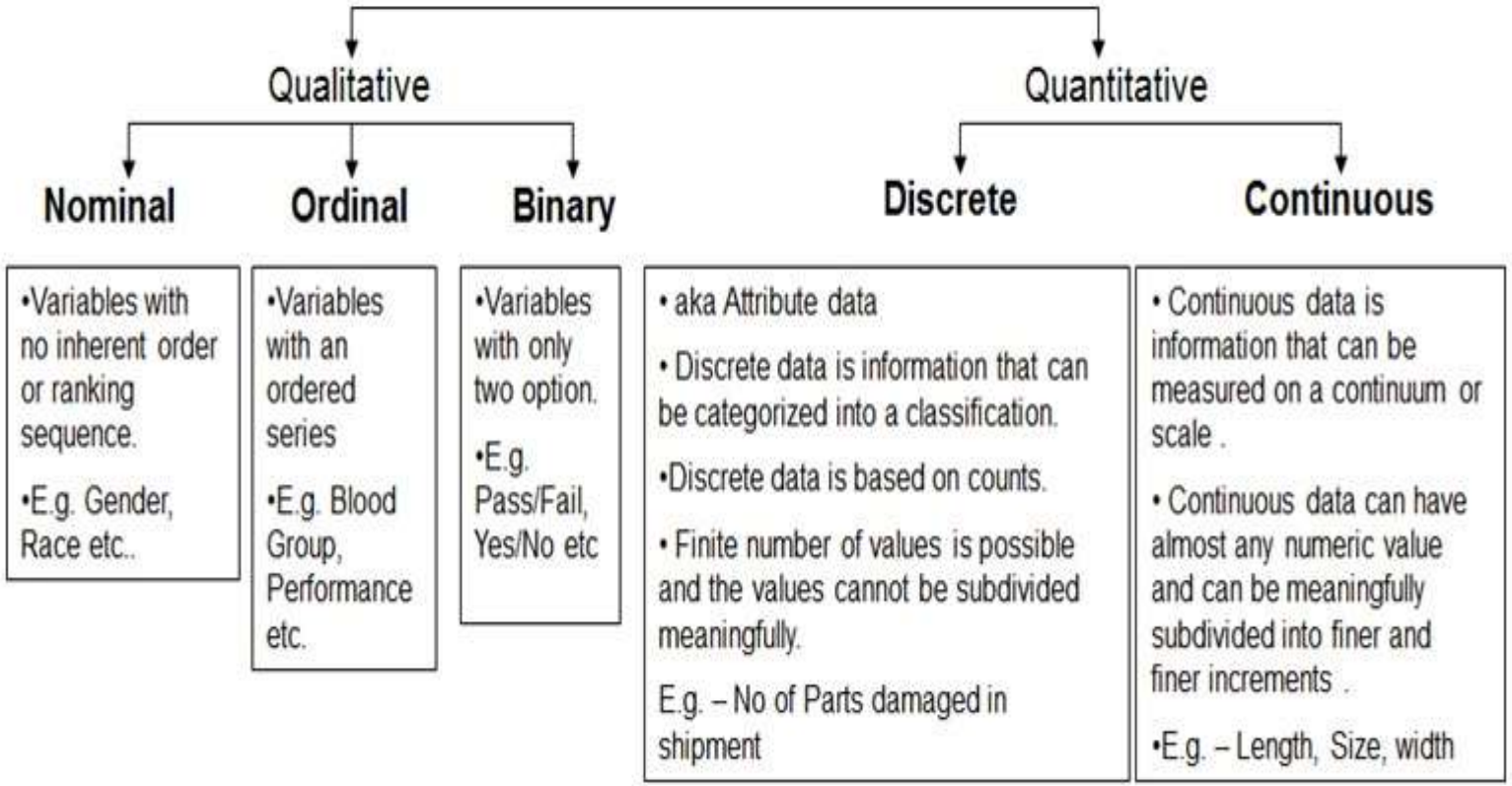
Types of Data



Sources: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Types of Data



Sources: www.sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Continuous Data

Continuous Data can take any value (within a range).

Examples:

•A person's height: could be any value (within the range of human heights), not just certain fixed heights.

•Time in a race: you could even measure it to fractions of a second.

Sources: www.mathisfun.com

Discrete Data

Discrete Data can only take certain values.

Example:

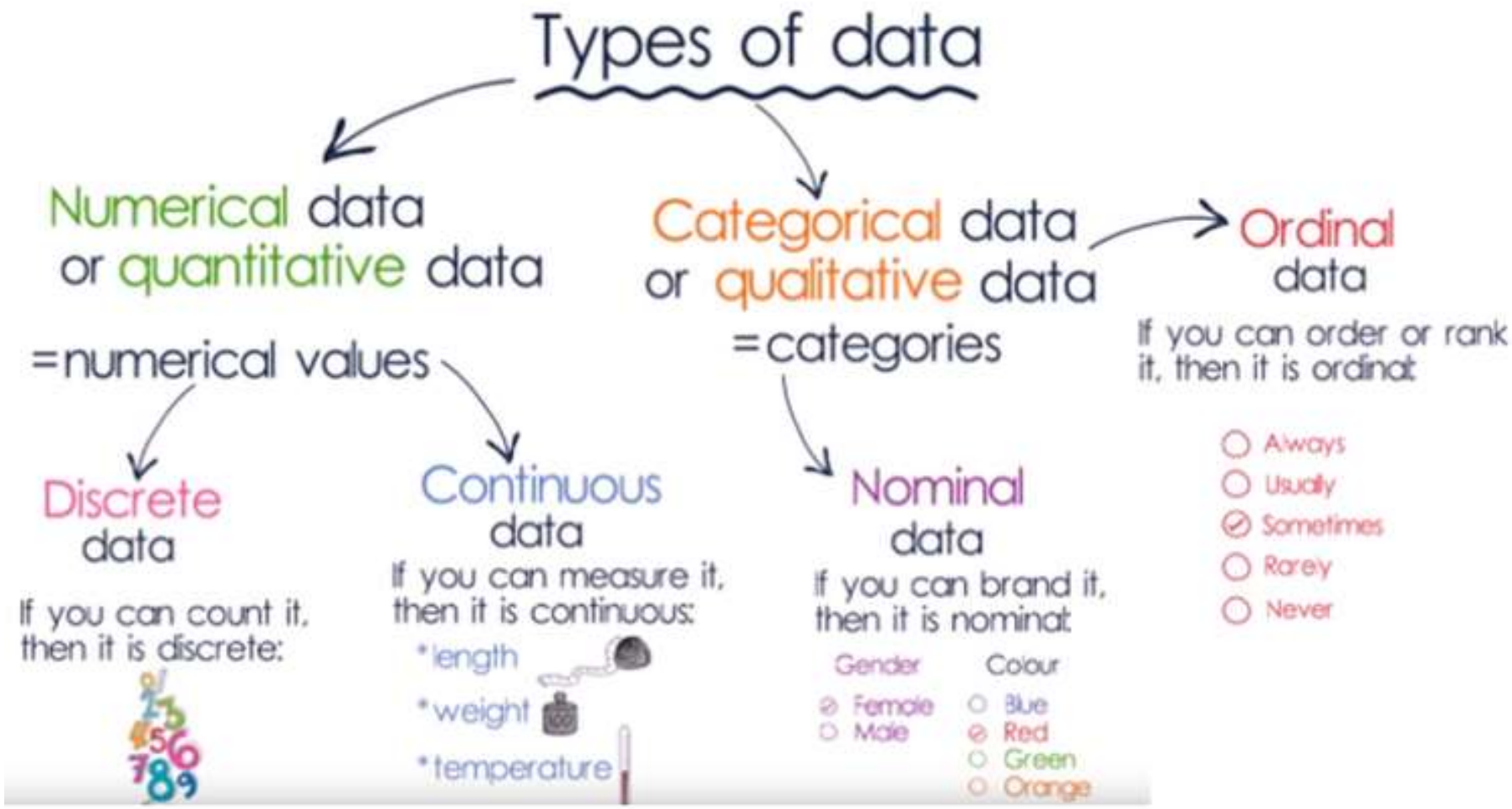
The results of rolling 2 dice only has the values

2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12

Sources: www.mathisfun.com

STATISTICS FOR DATA SCIENCE

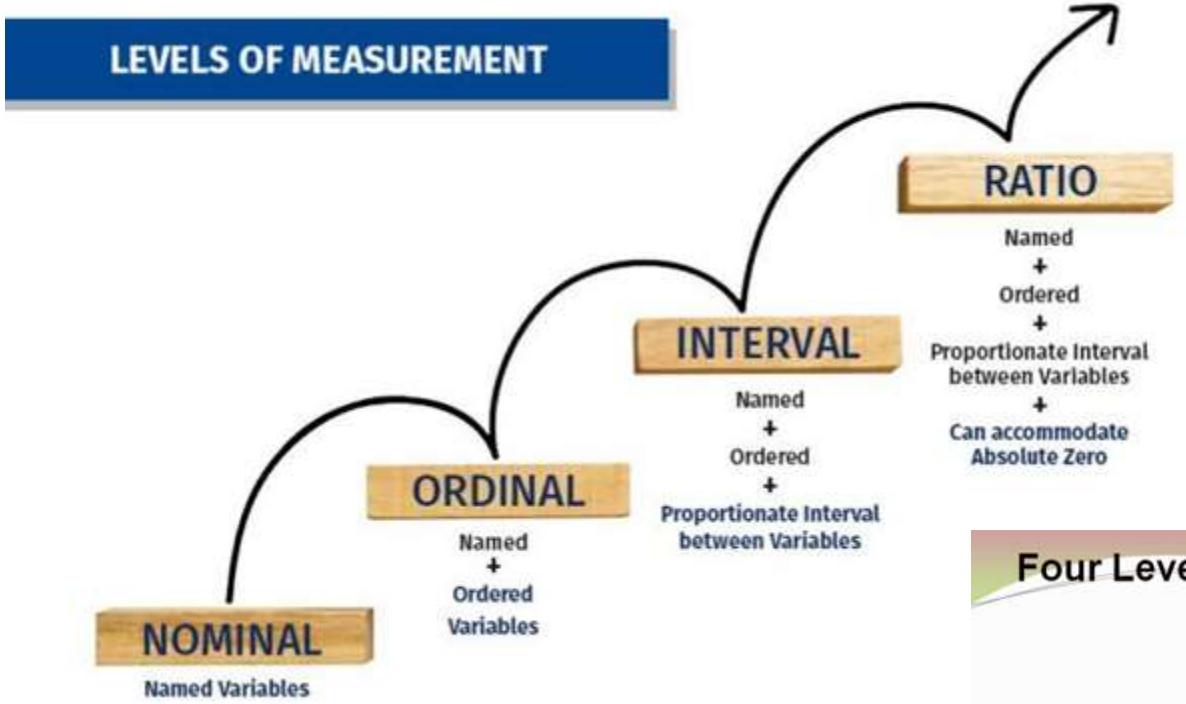
Types of Data : Summary

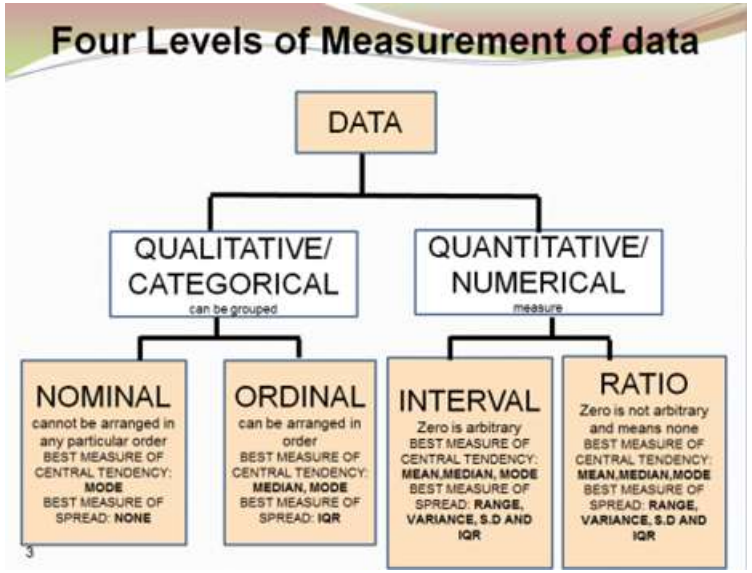


Sources: https://www.youtube.com/watch?v=GlgA8OMgLxE

STATISTICS FOR DATA SCIENCE

Types of Data : Summary





Sources: QuestionPro, Medium, slideshare.net

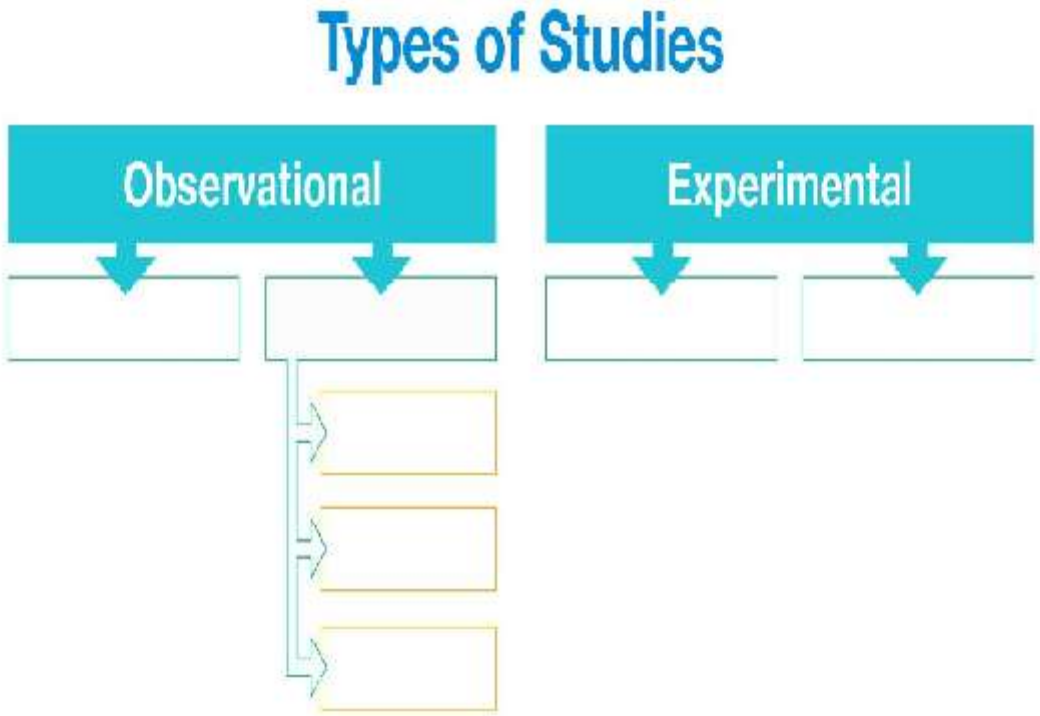
STATISTICS FOR DATA SCIENCE

Do it yourself !!!!!

- 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.

STATISTICS FOR DATA SCIENCE

Types of Studies



Definition: Text-Book, Chapter 1.1,Pg.No:11&12

Sources: www.acne.org

STATISTICS FOR DATA SCIENCE

Observational Study

- An observational study is a study in which the researcher simply observes the subjects without interfering.
- That is, the researcher has no control over any treatments the subjects may be given or which groups the subjects may be separated into, etc.
- They just observe the subjects and record data based on their observations.

Sources: www.study.com

STATISTICS FOR DATA SCIENCE

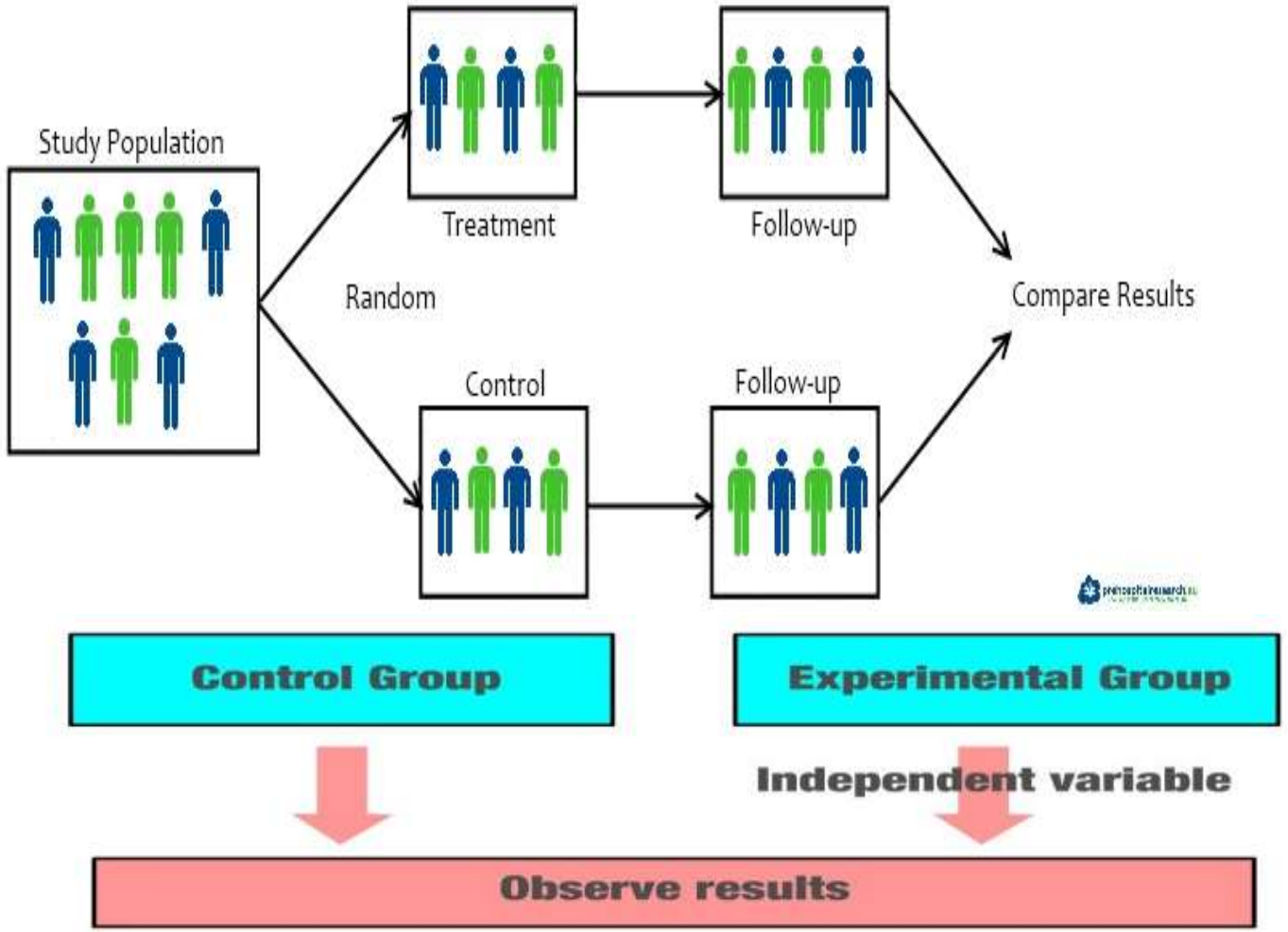
Experimental Study

- Experimental studies are ones where researchers introduce an intervention and study the effects.
- Experimental studies are usually randomized, meaning the subjects are grouped by chance.

Sources:https://www.iwh.on.ca/

STATISTICS FOR DATA SCIENCE

Experimental Study



Sources: prehospitalresearch.eu,www.scienceabc.com

Control vs. Experimental Group

Experimental and control groups:

In an experimental, a group is exposed to usual conditions, it is termed as 'control group', but when a group is exposed to some novel or special condition, it is termed as 'experimental group'.

Example:

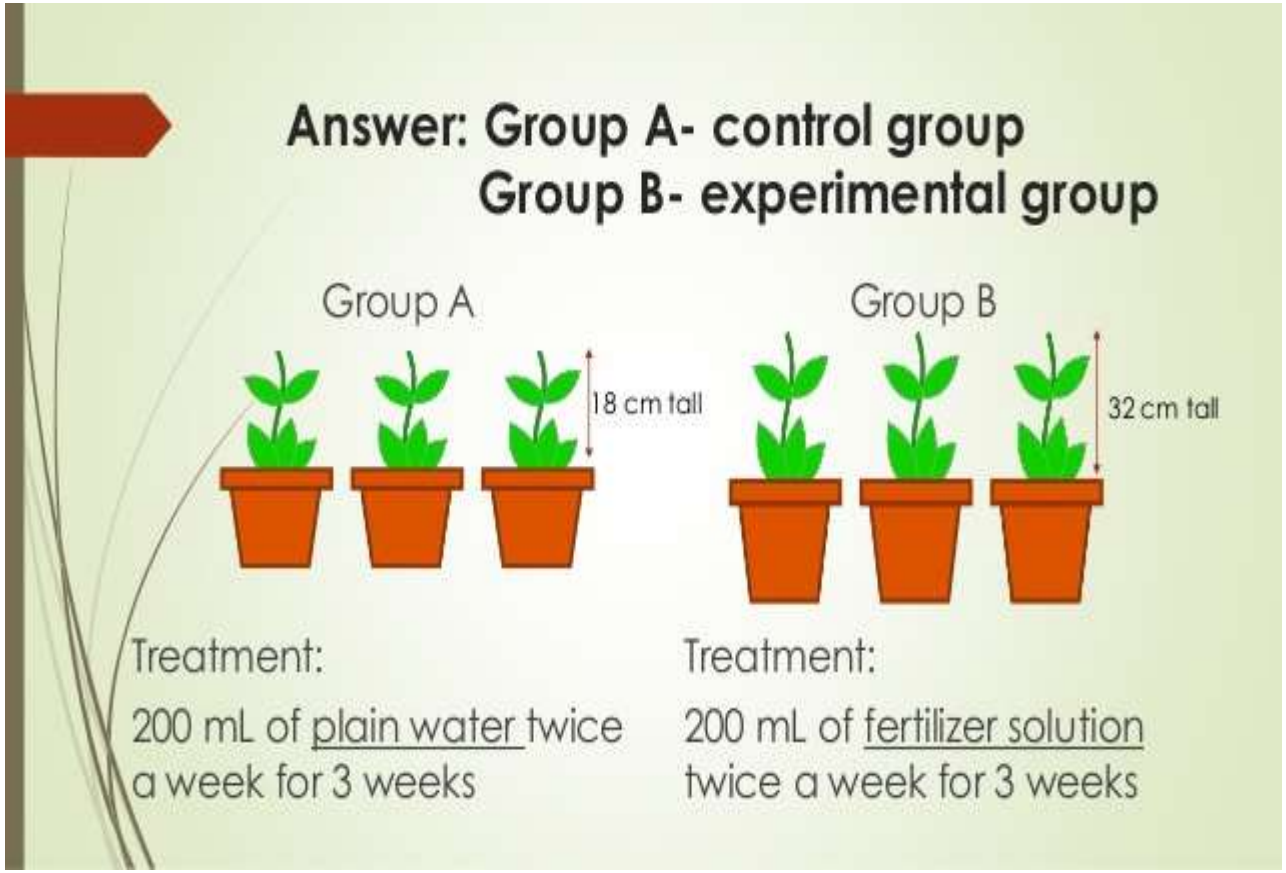
A student is testing to see if plants will grow without sunlight.

Which would be the experimental group and which would be the control group?

Sources: www.acne.org

STATISTICS FOR DATA SCIENCE

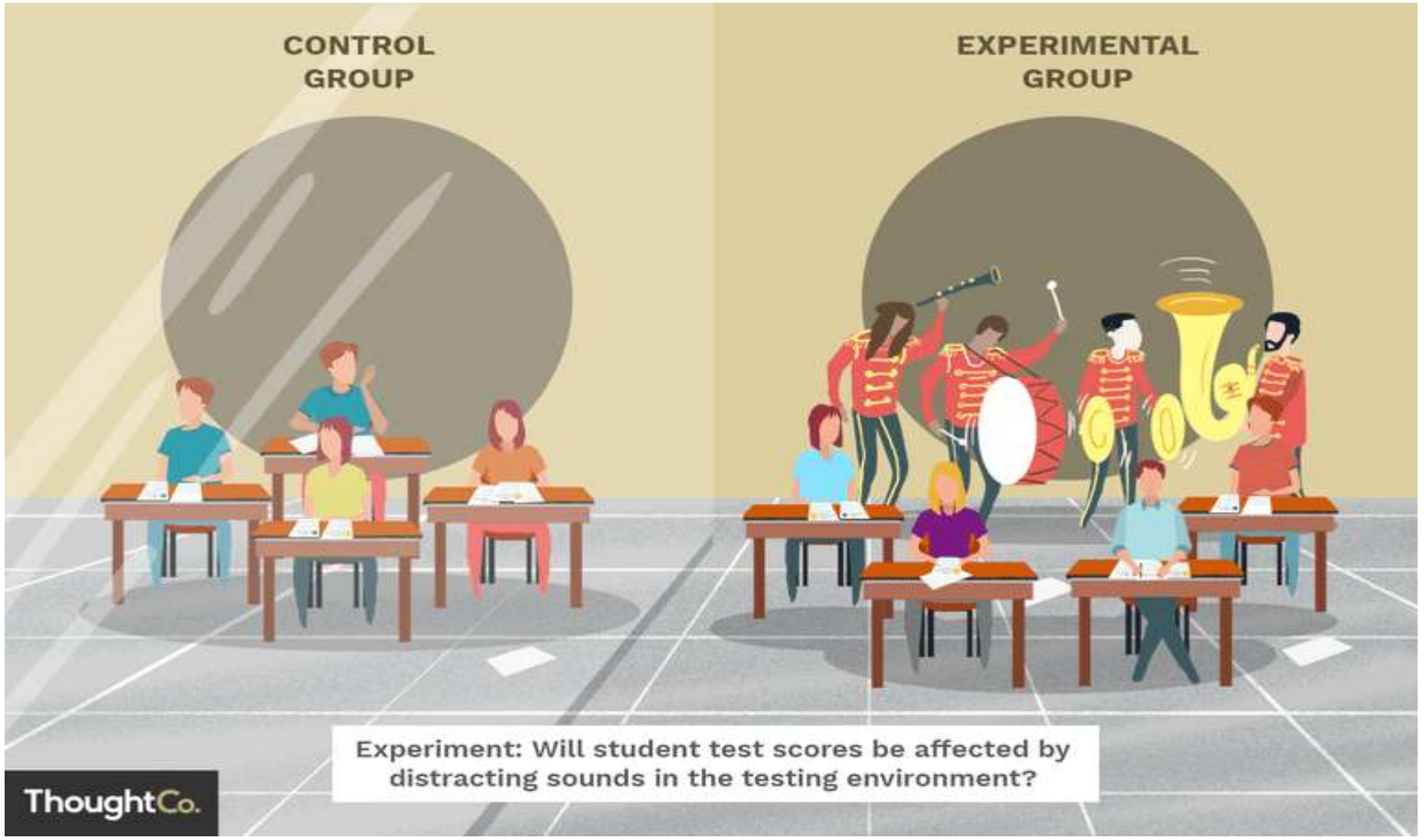
Control vs. Experimental Group



Sources: www.acne.org

STATISTICS FOR DATA SCIENCE

Types of Experiments



Sources: www.thoughtco.com

STATISTICS FOR DATA SCIENCE

Control Group & Experimental Group

- The group that does not receive the treatment is called the control group.
- An experimental group (sometimes called a **treatment group**) is a group that receives a treatment in an experiment.

Sources: https://www.statisticshowto.com/experimental-group

STATISTICS FOR DATA SCIENCE

Control Group & Experimental Group

Examples:

1.You are testing to see if a new plant fertilizer increases sunflower size. You put 20 plants of the same height and strain into a location where all the plants get the same amount of water and sunlight. One half of the plants–the control group–get the regular fertilizer. The other half of the plants–the experimental group–get the fertilizer you are testing.

2. You are testing to see if a new drug works for asthma. You divide 100 volunteers into two groups of 50. One group of 50 gets the drug; they are the experimental group. The other 50 people get a sugar pill (a placebo); they are the control group.

Sources: https://www.statisticshowto.com/experimental-group

Observational study vs. Experimental study

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.

Examples

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

Examples

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

Examples

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

Examples

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

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering