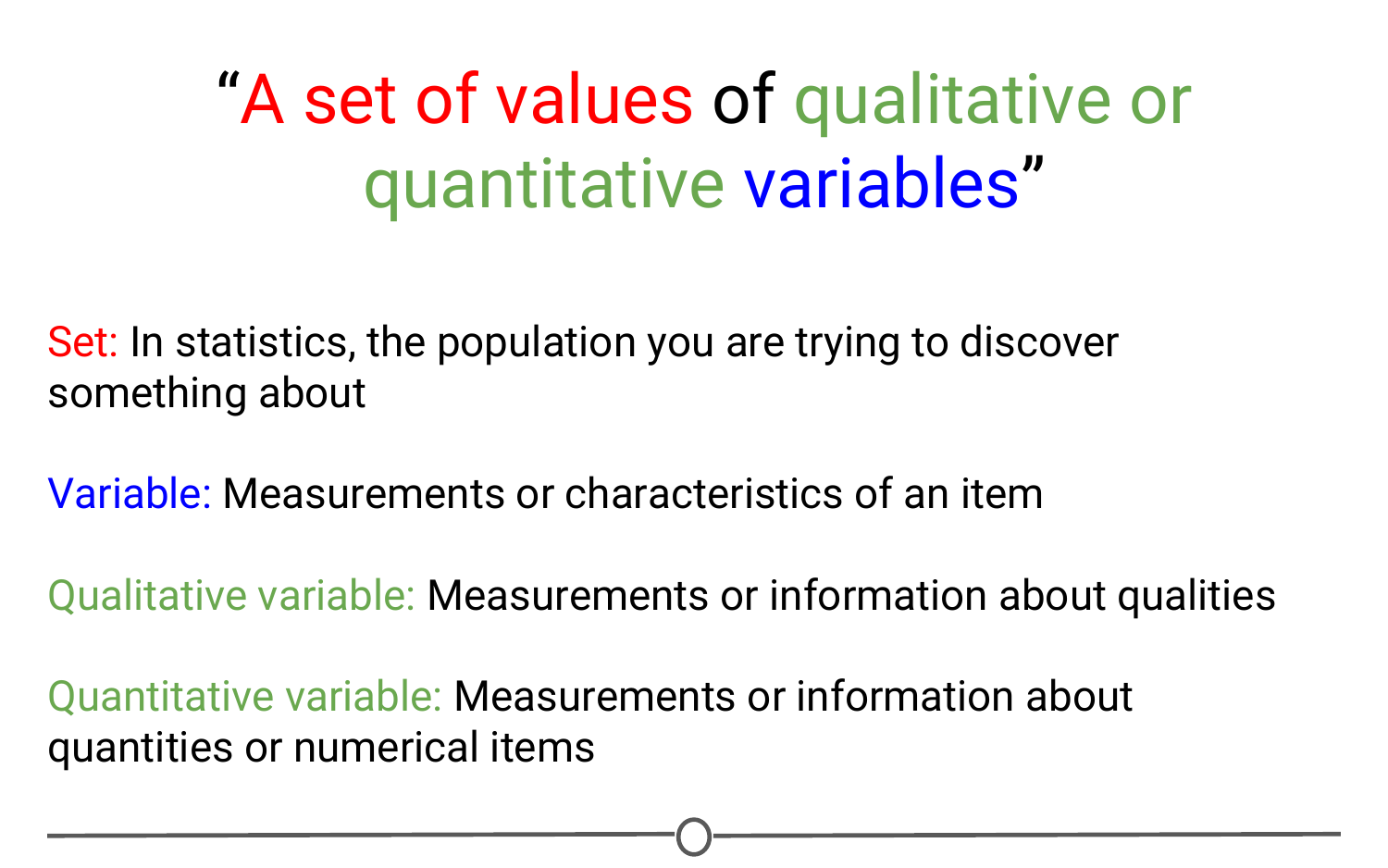
**What is data science: Data science can involve:**

* Statistics, computer science, mathematics
* Data cleaning and formatting
* Data visualization

- Path of learning: R 🡪 R studio 🡪 Version control 🡪 Applying tools to answer the questions

- Substantive skills: proffesional skills, in-depth knowledge relating specific industries, domains, etc. Nghiệp vụ

**What is data ?**



Good data scientists **ask for the questions first** and 🡪 Seek out **the relevant data**

**Stack Overflow and Cross Validated:** Forums to ask questions about data, coding problems, etc.

How to **install R Packages** **in this COURSE**

After installation 🡪 Load it

! Sometimes updating can change codes, BE CAREFUL

Understand Version Control 🡪 Keep track of who, when, and where a change was made 🡪 revert back

**What is Git ?**

Link with RStudio 🡪 Keep all files and track the versions changed

**Types of data science questions:**

1. Descriptive

🡪 Summarize a set of data

* Early analysis of new data, simple summaries about the samples and their measurements
* NOT for generalizing the results of the analysis to a larger population or conclusions

1. Exploratory

🡪 Examine the data, find relationships that were not previously known

How different variables might be related ?, new connections 🡪 formulate hypotheses 🡪 design of future studies and data collection

**\*\* Correlation does not imply causation!**

1. Inferential (phan tich suy luan)

🡪 use relatively small sample of data 🡪 Say about population at large

* Provide the estimation of the variable for the population and provide uncertainty about the estimation
* Accurately inferring (suy ra) info about larger population depends on sampling scheme

1. Predictive

🡪 Use current and historical data to find patterns and make predictions about future data

* Accuracy in predictions is dependent on measuring RIGHT variables
* Many ways to build up prediction models with some being better or worse for specific cases (more data and a simple model generally performs well at predicting future outcomes)

**\* just because one variable predicts another does not mean it causes the other !**

1. Casual (phan tich nhan qua)

🡪 What happends to one variable when we manipulate another variable ?

* Gold standard in data analysis
* Often applied to the **results** **of randomized studies** that were **designed to identify causation**
* Usually analysed in aggregate and observed relationships are usually average effects.

1. Mechanistic( a method for deeply understanding how complex systems and phenomena work by dissecting the underlying processes, interactions, and components)

🡪 **Understand exact changes** in variables leading to exact changes in other variables

* Applied to simple situations or those are nicely modeled by deterministic equations
* Commonly applied to physical or engineering sciences (biological sciences are far too noisy to use mechanistic analysis)
* Often, the only noise in the data is measurement error

**Experimental Design**

- Experimental design is organizing an experiment so that you have the correct data (**and enough** of it) 🡪 clearly and effectively answer your data science question.

- This process involves clearly **formulating your question** in advance of any data collection, **designing the best set-up possible** 🡪 **gathering the data** to answer your question 🡪 **identifying problems** or **sources of error in your design** 🡪 **collecting the appropriate data.**

**Why should we care about Experimental Design ?**

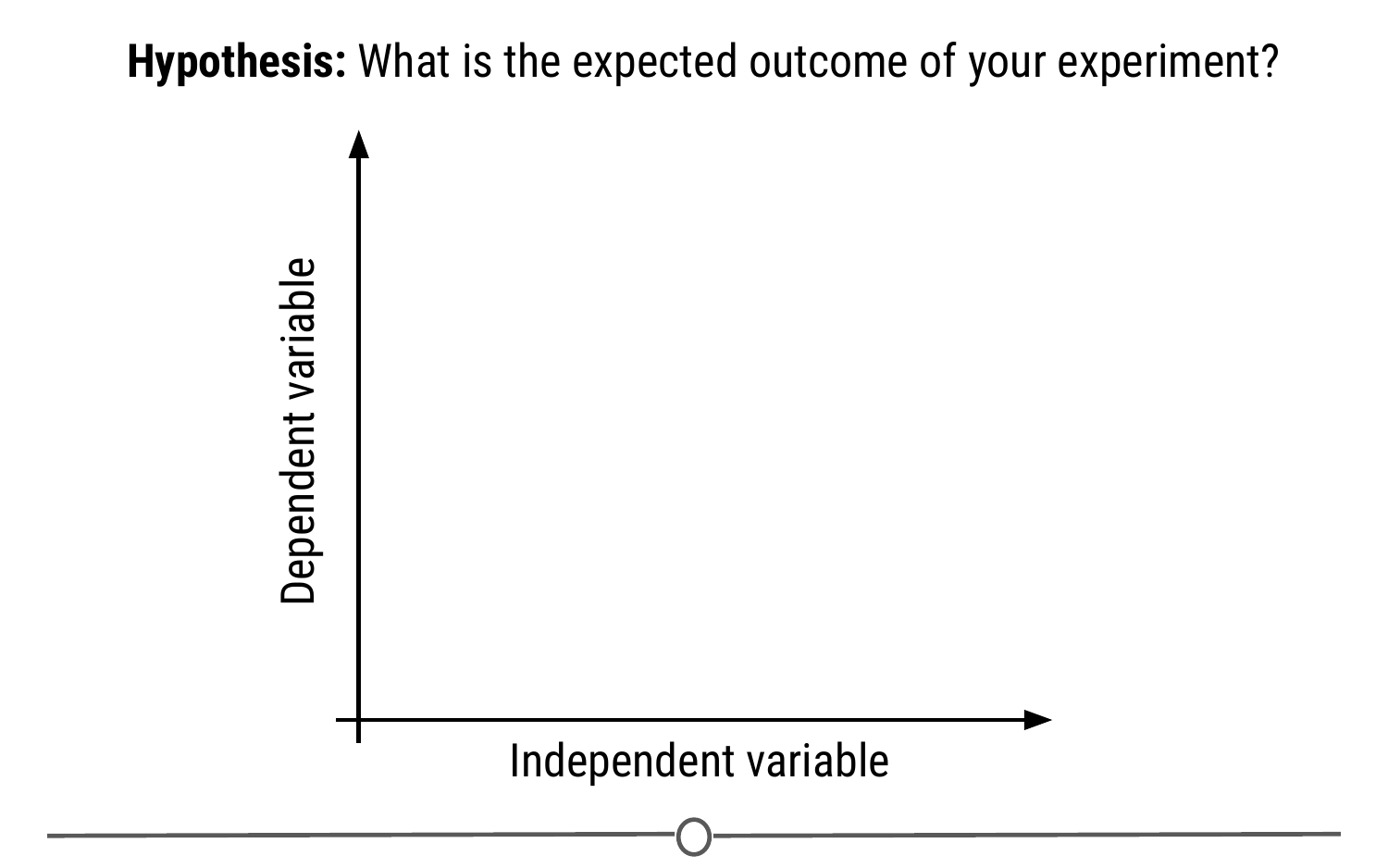
If you do the wrong analysis, you can come to the wrong conclusions!

For example, a paper that was trying to predict the effects of a person’s genome (bộ gen) on their response to different chemotherapies (hoá trị), to guide which patient receives which drugs to best treat their cancer. 🡪 Was retracted. Why ? 🡪Data have problems in their set-up and cleaning

**Principles of experimental design**

* **Independent variable (AKA factor):** The variable that the experimenter manipulates; it does not depend on other variables being measured. Often displayed on the **x-axis.**
* **Dependent variable:** The variable that is expected to change as **a result of changes** in the **independent variable.** Often displayed on the **y-axis,** so that **changes in X**, the independent variable, **effect changes in Y.**

**🡪** When we are designing an experiment 🡪 have to decide what variables will be measured and which will be manipulated to effect changes in other measured variables 🡪 Must develop your hypothesis 🡪 The relationship between **your variables** and the **outcome of your experiment**

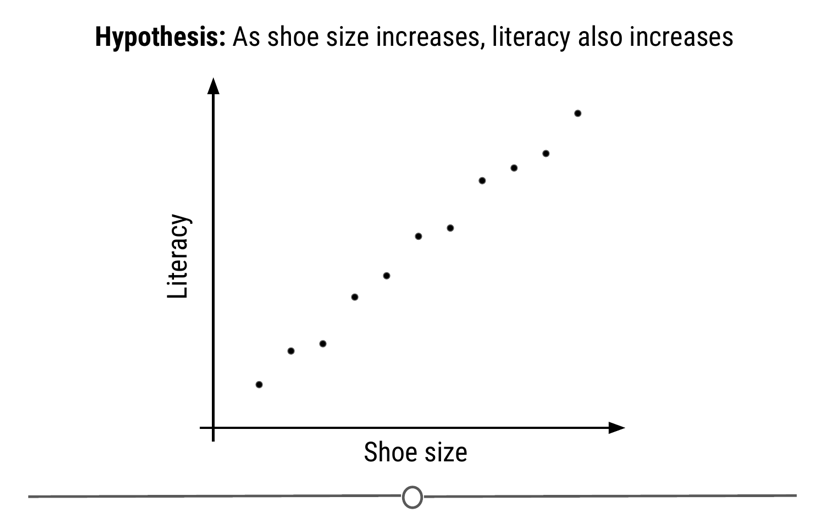


**Example:** Experimental set-up: Hypothesize that literacy level depends on shoe size

Design an experiment of 100 individuals. Before collect data:

* Consider if there are problems with this experiment that causes an erroneous (sai lệch) result 🡪 In this case, the experiment may be fatally flawed by a confounder.
* **Confounder (yếu tố gây nhiễu)**: An extraneous variable that may affect the relationship between the dependent and independent variables.
* Since age affects foot size and literacy is affected by age 🡪 if any relationship between shoes size and literacy 🡪 the relationship may actually be **due to age – age** is confounding our experimental design 🡪 To control:

Measure the age of each individual 🡪 take into account the effects of age on literacy. Another way to control **the effect of age on literacy** is to **fix** the age of **all pariticpants**



:

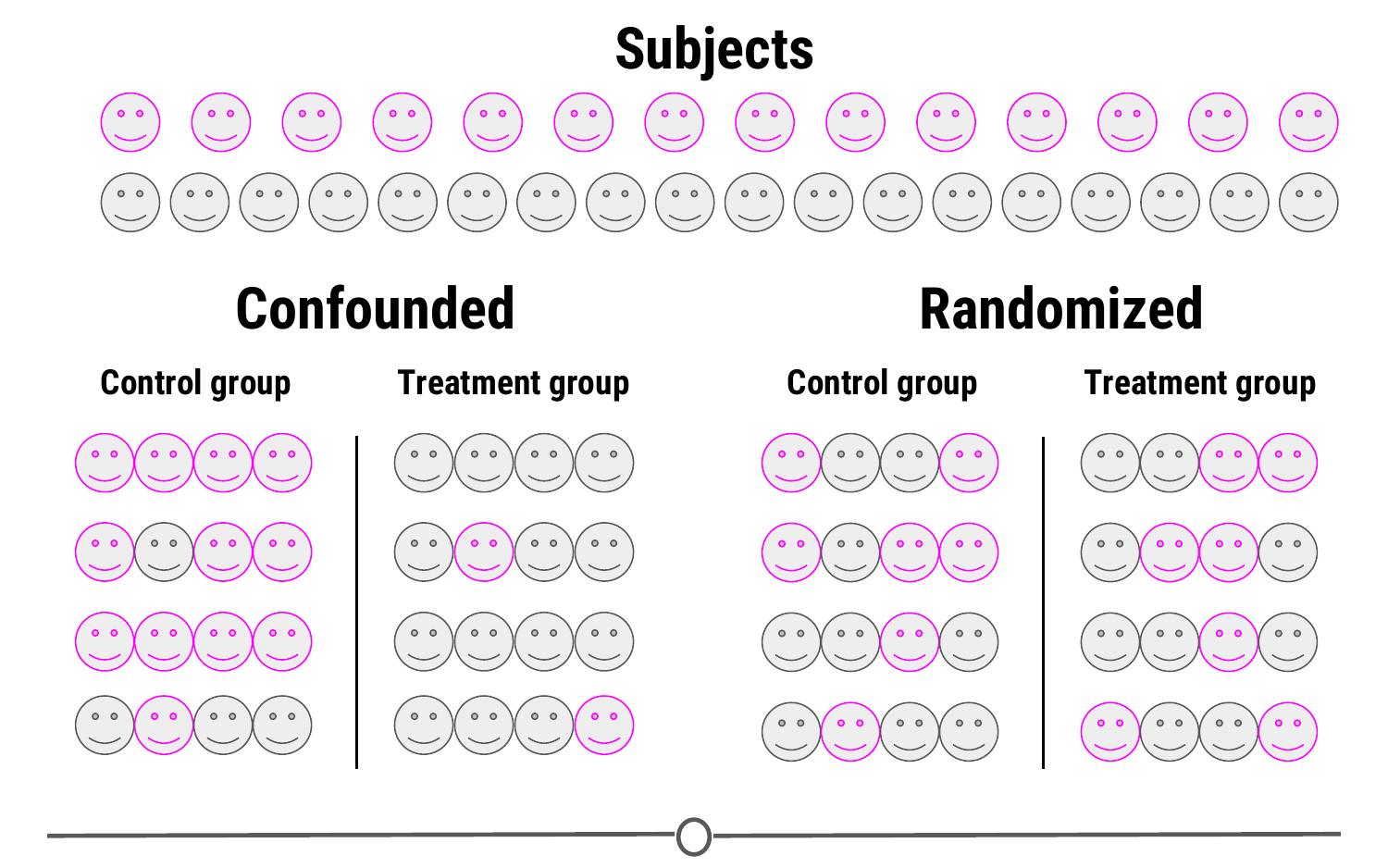
A diagram of a diagram

AI-generated content may be incorrect.

Strategies to control for confounding effects.

**Ex:** Blinding your study means that your subjects don’t know what group they belong to - all participants receive a “treatment” (thí nghiệm cho 2 nhóm sử dụng thuốc thật và thuốc giả làm từ đường để vượt qua **placebo effect** (ở đây là counfounding effect).

Balancing of confounders 🡪 randomization 🡪 to help lessen the risk of accidentally biasing one group to be enriched for a confounder, you can randomly assign individuals to each of your groups.



**Replication Concept**

Replication is repeating an experiment with different experimental subjects.

A single experiment’s results may **have occurred by chance**; **a confounder was unevenly distributed across your groups**, there was a **systematic error** in the **data collection**, there were **some outliers,** etc.

🡪 However, if you can **repeat the experiment** and **collect a whole new set of data** and still come to the **same conclusion** 🡪 your study is much stronger. Also at the heart of replication is that it allows you to measure the variability of your data more accurately, which allows you to better assess whether any differences you see in your data are significant.

**Beware p-hacking!**

**3 QUALITIES OF BIG DATA**

