

# Human-Computer Interaction SoSe 25


## *Evaluation - Part 1*



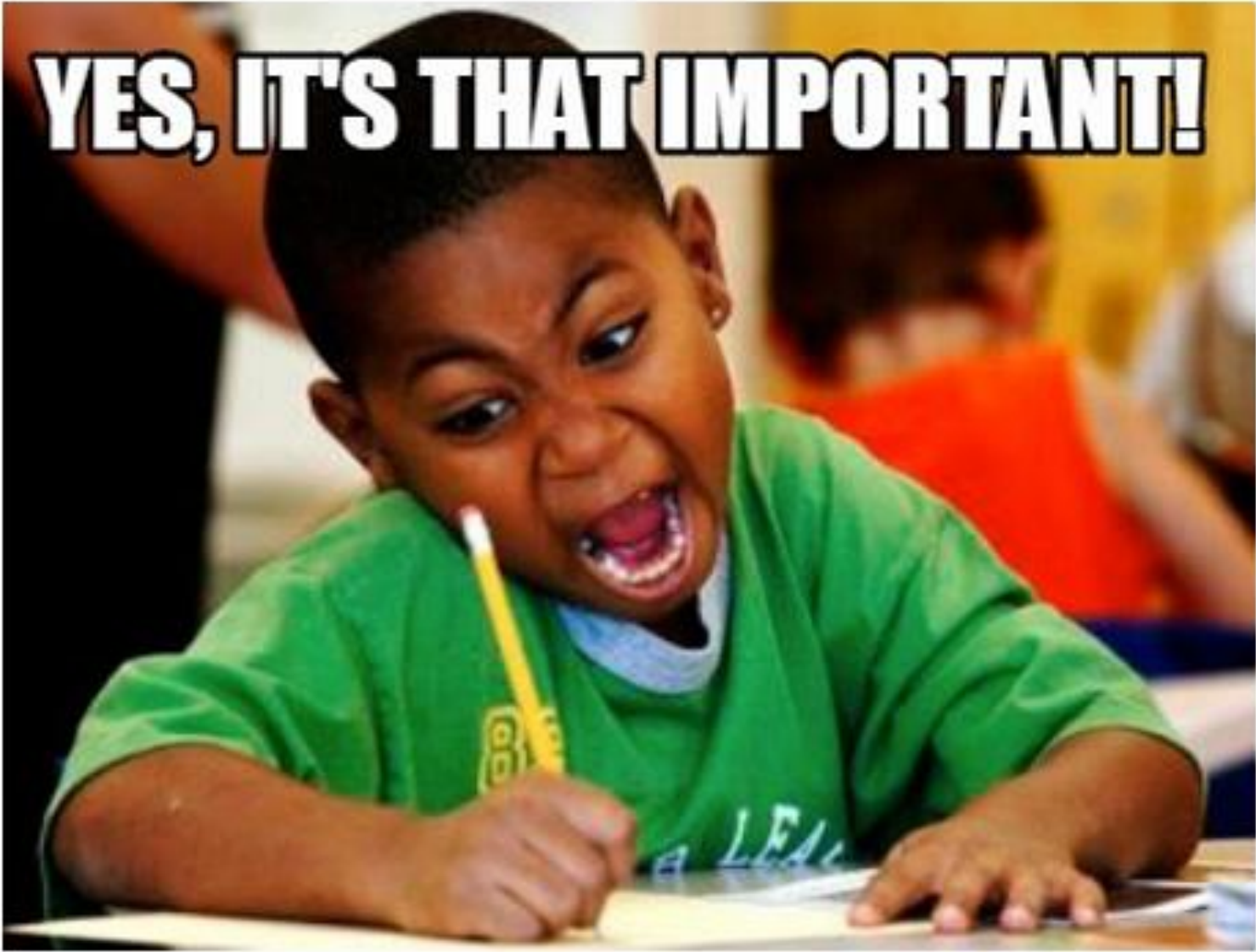
# Lecture

1. What is Human-Computer Interaction
  1. Basics of HCI
  2. History of HCI
2. Cognitive Basics
  1. Cognition and Perception
  2. Metaphors and Mental Models
  3. *In-class Exercise*
3. Usability Engineering
  1. (Human) Design Processes
  2. Prototyping
  3. *In-class Exercise*
4. Evaluation Methods
  1. Study Design
  2. Methodology
  3. *In-class Exercise*
5. Post-WIMP + Input Devices
6. Mixed-Reality
7. Wrap-Up and Q&A

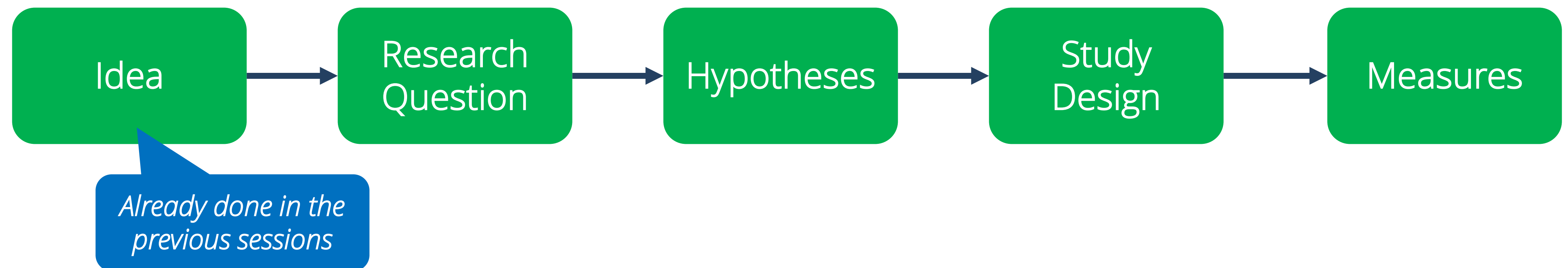




A1	A2	A3	A4	A5	A6	A7	A8	A9	B	$\Sigma$	Note
15	8	9	9	7	5	22	10	5	?	90	









### Research question:

A statement that describes or explains a **relationship between or among variables** and is a proposal to be tested.

Therefore, need to identify **variables** and **research question** for your observation.

### What are Variables?

Characteristics or conditions that change or have different values for different individuals.



## Good and Bad Research Questions:

### 1) Questions Should Have Complex Answers (not just yes/no)

**Bad:** Does owning a pet improve quality of life for older people?

**Good:** In what ways does owning a pet improve quality of life for older people?

### 2) Good Research Questions Need Focus (not too generalized)

**Bad:** Does medication help alleviate attention deficit hyperactivity disorder (ADHD) symptoms? And do kids need more exercise?

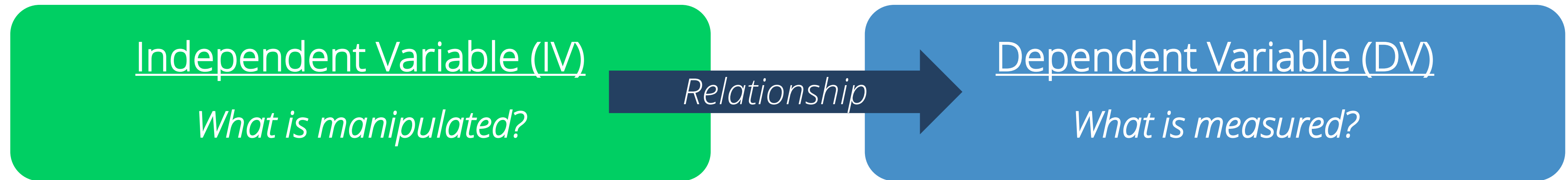
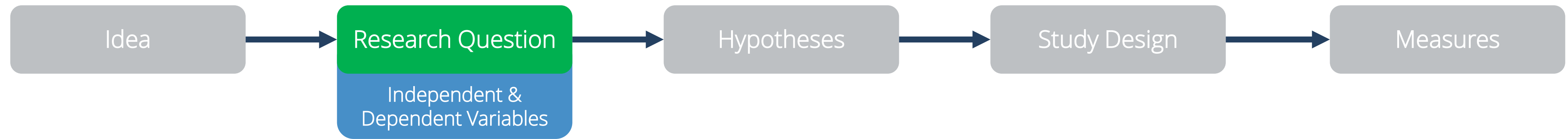
**Good:** How effective are the various types of medication in treating elementary students with ADHD?

### 3) Questions Should Be Specific and Precise

**Bad:** How do artificial sweeteners affect people?

**Good:** How does aspartame affect elderly women older than 70 who suffer from migraines?





IV refer to the factors that the researchers are interested in studying or the possible “*cause*” of the change in the dependent variable,  
*i.e., variables that we change during the experiment*

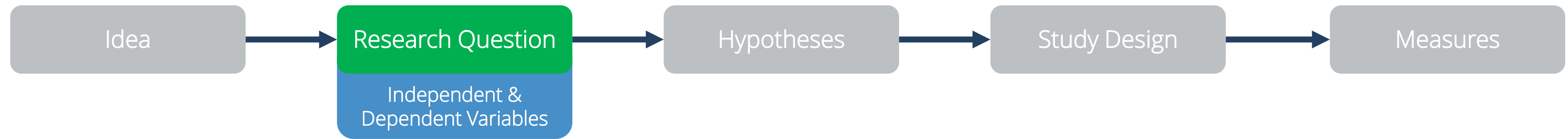
Examples:

- Type of Interaction
- Context
- Input Modalities

DV refer to the *outcome* or *effect* that researchers are interested in,  
*i.e., variables that we measure during the experiment*

Examples:

- Efficiency: task completion time, speed
- Accuracy: error rate
- Subjective satisfaction
- Ease of learning and retention rate
- Physical or cognitive demand



## Independent Variable (IV)

*What is manipulated?*

Often use control condition or a baseline technique as one level

Are they:

- Similar in all non-essential aspects?
- Biased ? Best possible alternative ?
- Representative ?
- Realistic ?

### **Example:**

**IV:**

Input Device

**Levels:**

Mouse, Controller, Touch

**Conditions:**

Mouse, Controller, Touch

**IV:**

Input Device, Expert Level

**Levels:**

Input Device (Mouse, Controller, Touch)

Expert Level (Beginner, Pro)

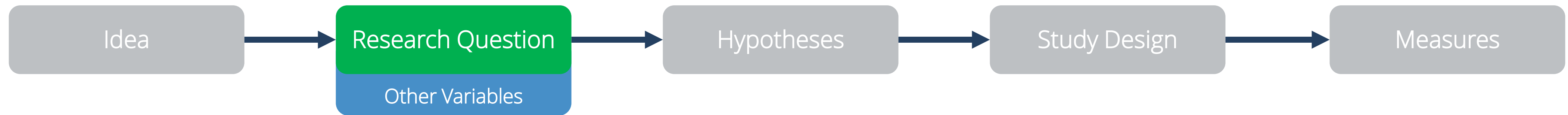
**Conditions:**

Mouse-Beginner, Mouse-Pro,

Controller-Beginner, Controller-Pro,

Touch-Beginner, Touch-Pro,





## Controlled variables

- What is held constant for all trials / participants.
- E.g., lighting conditions, room temperature.

## Random Variables

- What is allowed to vary randomly/uncontrolled.
- E.g., time of the day.
- Needs to have an expected relationship to IV

## Confounding variable

- What correlates with the independent + dependent variable
- Extra variable(s) you didn't account for
- Might suggest correlation when there isn't one

Example:

Activity level  
(Independent Variable)



Weight Gain  
(Dependent Variable)

**Potential Confounding Variables:**  
Eating Habits, Starting Weight, Age, Occupation

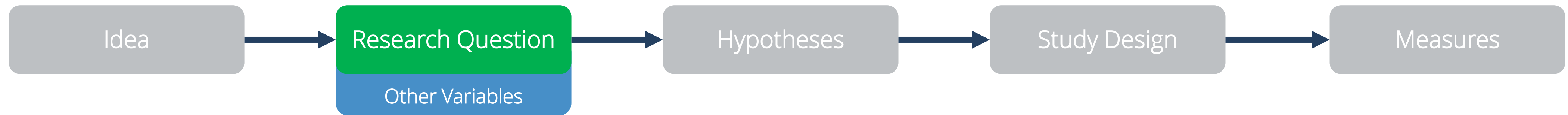
How can these be mitigated?

- E.g., Random sampling with huge sample
- E.g., Control for age (25-40 year olds)

## Hint:

Variables have a name and levels/values, just like in every programming language.

var name = a | b | c | d | ...



### Controlled variables

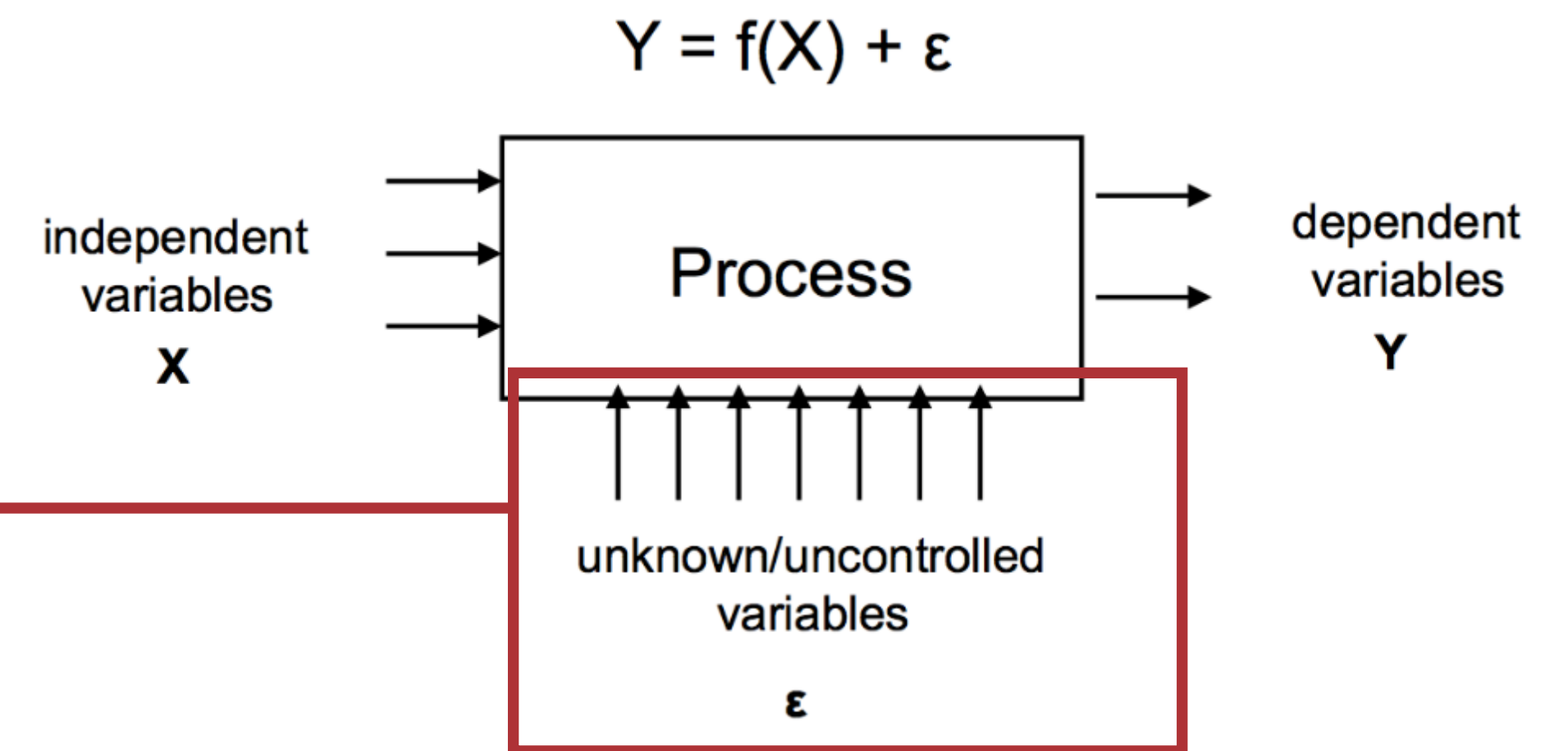
- What is held constant for all trials / participants.
- E.g., lighting conditions, room temperature.

### Random Variables

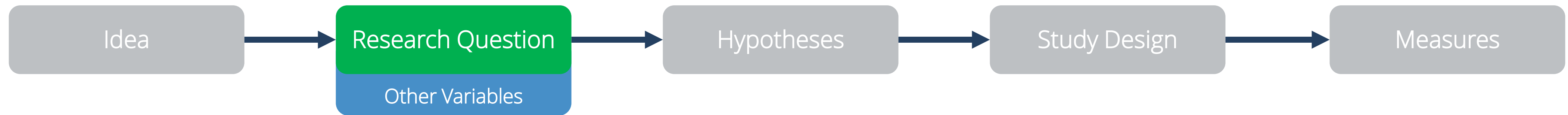
- What is allowed to vary randomly/uncontrolled.
- E.g., time of the day.
- Needs to have an expected relationship to IV

### Confounding variable

- What correlates with the independent + dependent variable
- Extra variable(s) you didn't account for
- Might suggest correlation when there isn't one







### Controlled variables

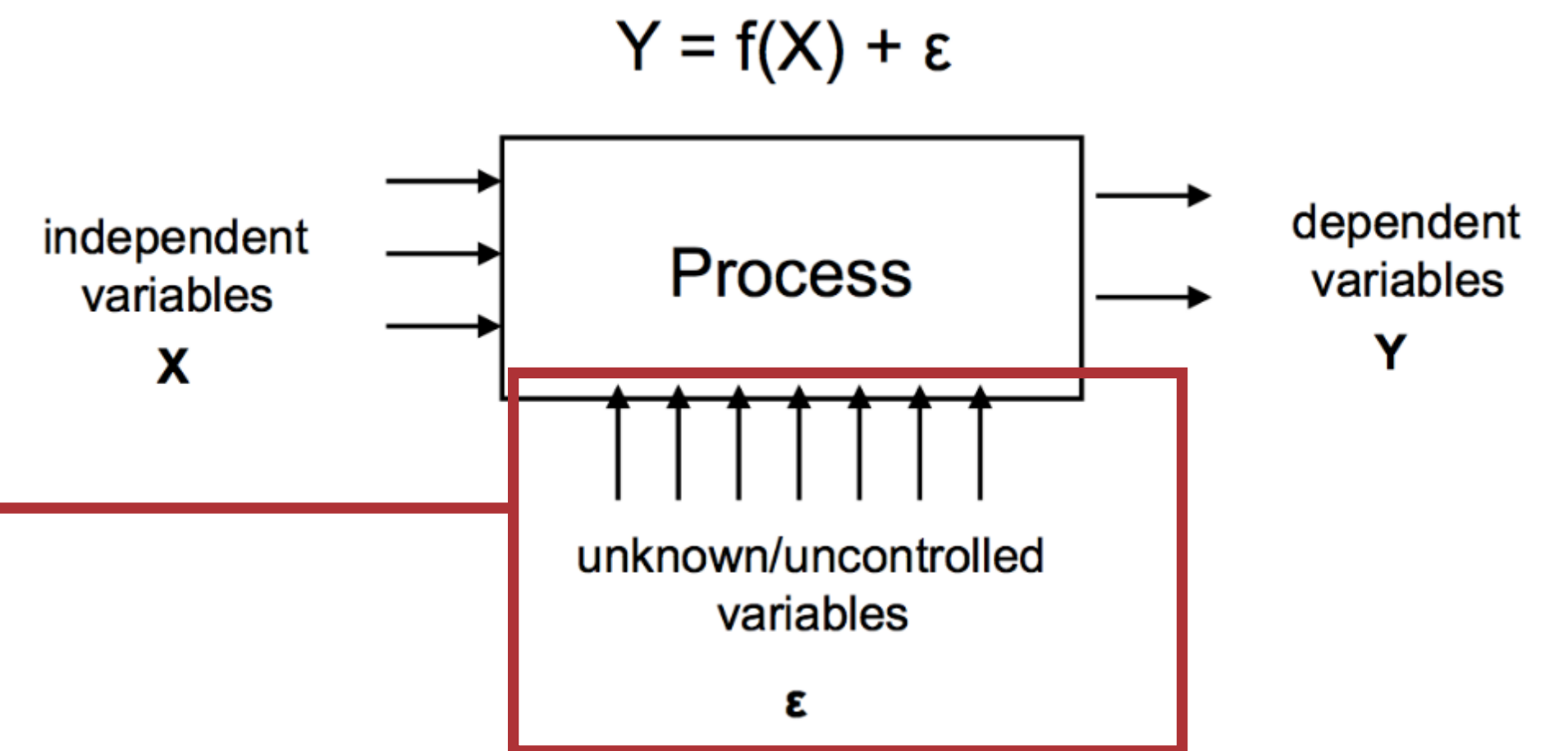
- What is held constant for all trials / participants.
- E.g., lighting conditions, room temperature.

### Random Variables

- What is allowed to vary randomly/uncontrolled.
- E.g., time of the day.
- Needs to have an expected relationship to IV

### Confounding variable

- What correlates with the independent + dependent variable
- Extra variable(s) you didn't account for
- Might suggest correlation when there isn't one



### Example:

**Controlled**  
**Random**  
**Confounding**

Quality/Price of Input Device  
Seating Position  
Sampling Rate of Input Device



## What is a hypothesis?

- A claim that **predicts the outcome of experiment**
- Concrete and testable statements derived from the research question
- An experiment normally starts with a research hypothesis..  
.. which is a precise problem statement that can be directly tested through an empirical investigation

## Experiment goal: Confirm hypothesis

- Approach: Reject *null hypothesis* ( $H_0$ ; inverse, i.e., “no influence”) by finding statistical evidence to refute or nullify the null hypothesis in order to support the alternative hypothesis
- Null hypothesis* is a term from statistical testing: The samples are drawn from the same statistical distribution

A well-defined hypothesis clearly states the dependent and independent variables of the study





Example from Research Question to Hypothesis:

*Research Question:*

Which Input Device allow the user to perform the best in competitive 3<sup>rd</sup> person games

*Hypothesis:*

Participants will receive higher scores using mouse input compared to touch and controller input when playing X minutes of Game Y

Hint:

Check you included your DVs and IVs

Independent Variable

Dependent Variable

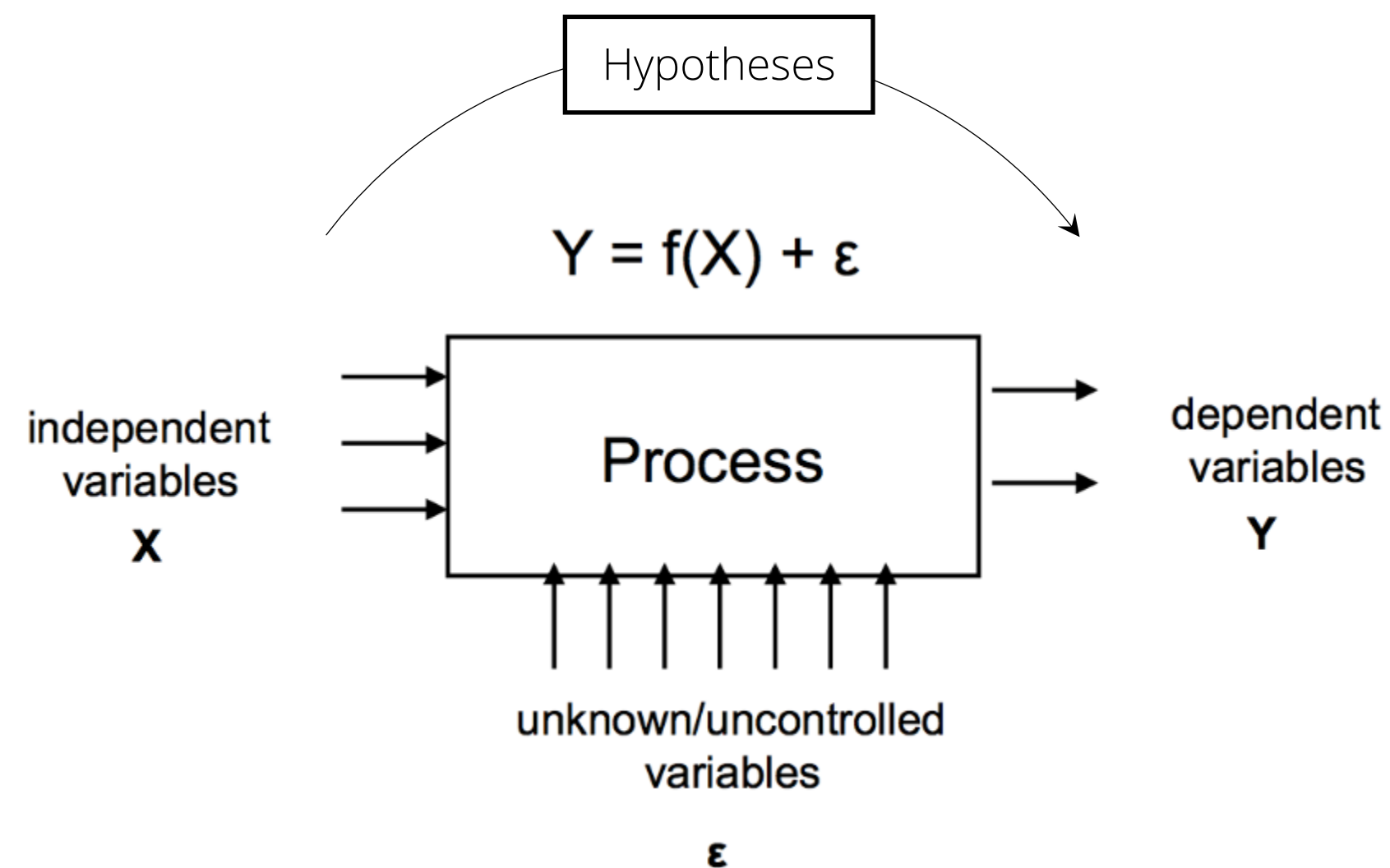


### *Research Question:*

Which Input Device allow the user to perform the best in competitive 3<sup>rd</sup> person games

### *Hypothesis:*

Participants will receive higher scores using mouse input compared to touch and controller input when playing X minutes of Game Y







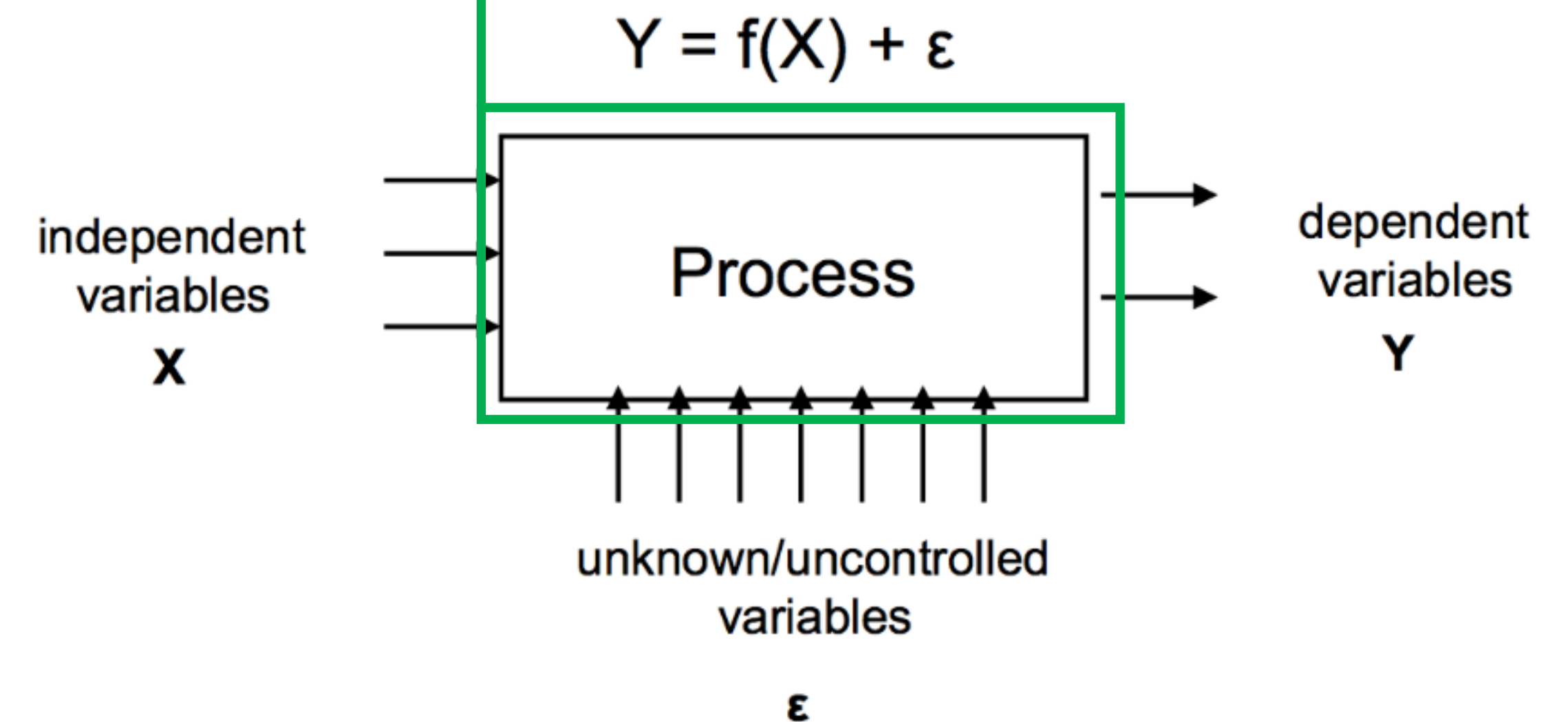
## How to get from a hypothesis to a study?

1. Identify research questions and hypotheses (*done yaye*)
2. Specify the design of the study:
  - Type of experiment
  - Within- or between-subjects design
  - Experimental conditions
  - Task description and procedure
  - Experimental measures
3. Run a pilot study and re-iterate if necessary
4. Recruit participants
5. Conduct the actual study (incl. data collection)
6. Analyze the data



## How to get from a hypothesis to a study?

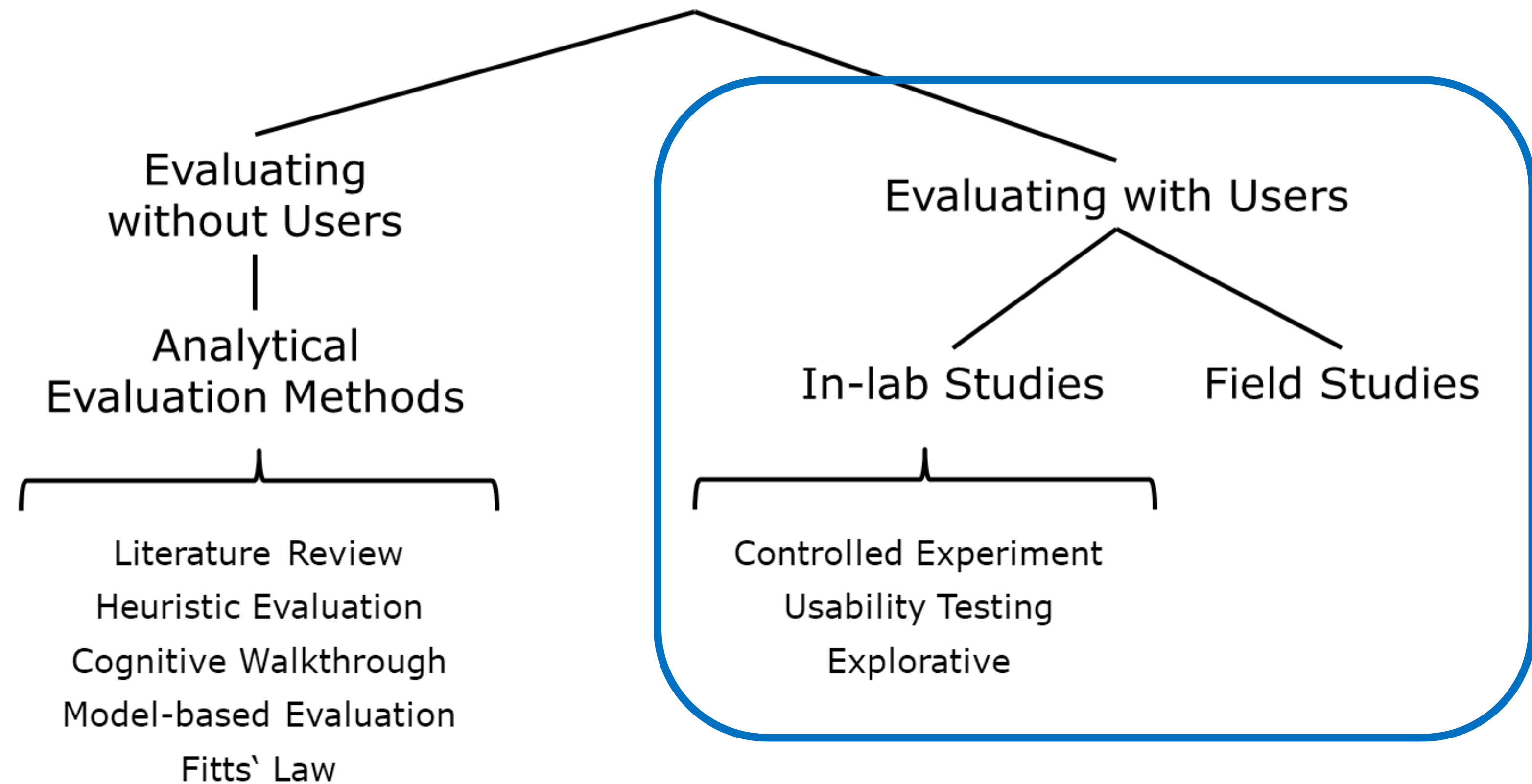
1. Identify research questions and hypotheses (*done yaye*)
2. Specify the design of the study:
  - Type of experiment
  - Within- or between-subjects design
  - Experimental conditions
  - Task description and procedure
  - Experimental measures
3. Run a pilot study and re-iterate if necessary
4. Recruit participants
5. Conduct the actual study (incl. data collection)
6. Analyze the data

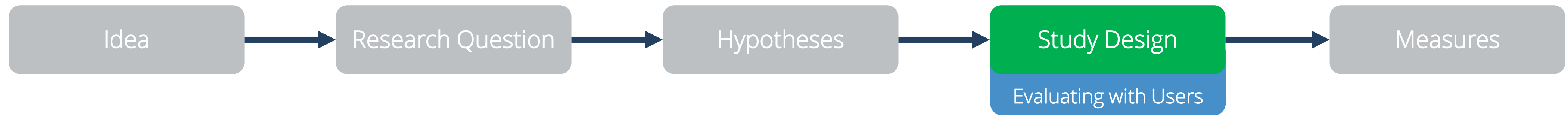






## Types of Experiments





### Evaluating with Users:

- Unlike the analytical methods, evaluating with users involves actual (or potential) users
- Participants should be representative (e.g., age, disabilities, ...) for the identified target audience
- Requires more time but leads to deeper insights on how users actually interact with a something
- Allows studying effects for which no models/heuristics/etc. exist

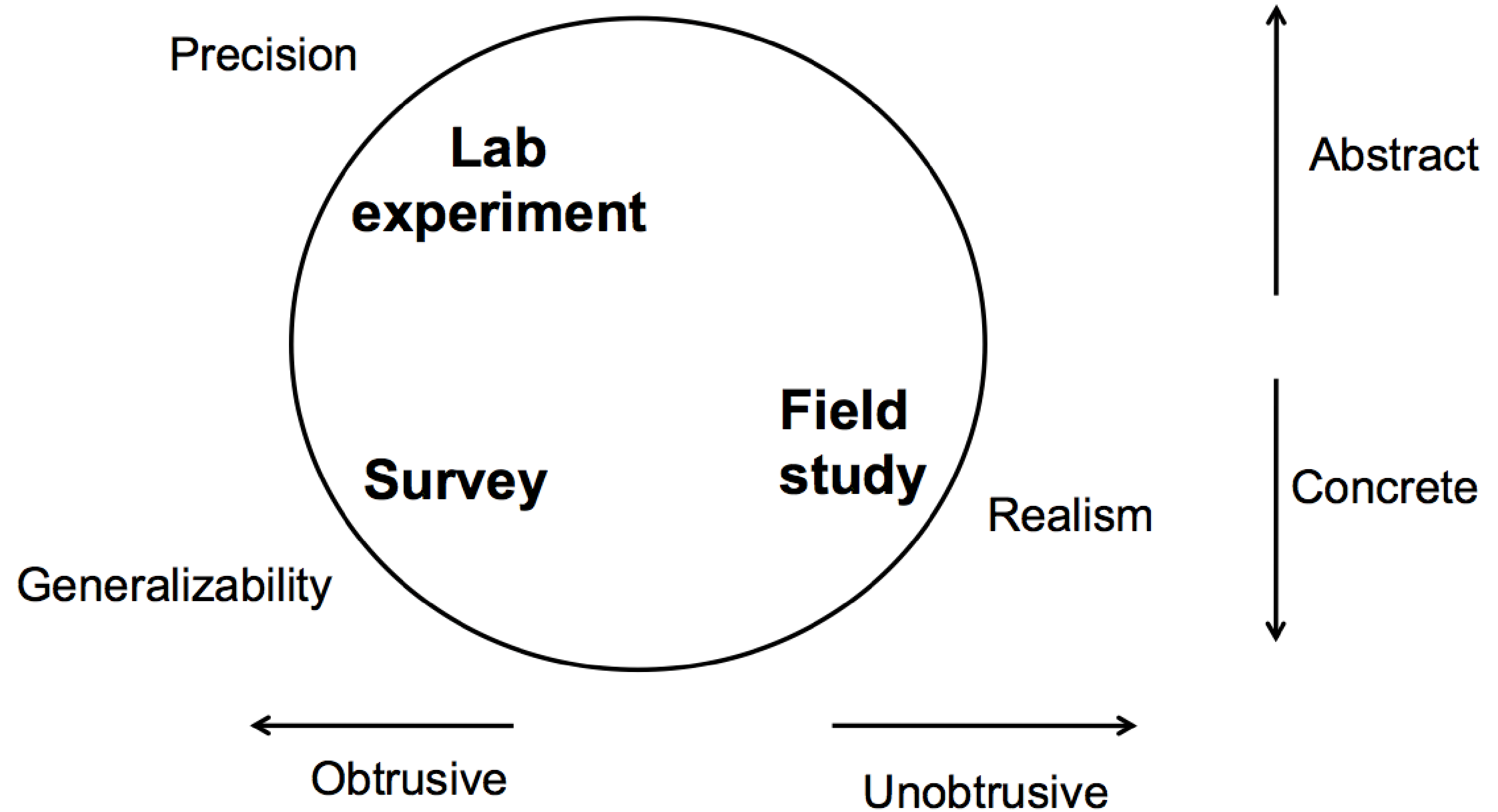
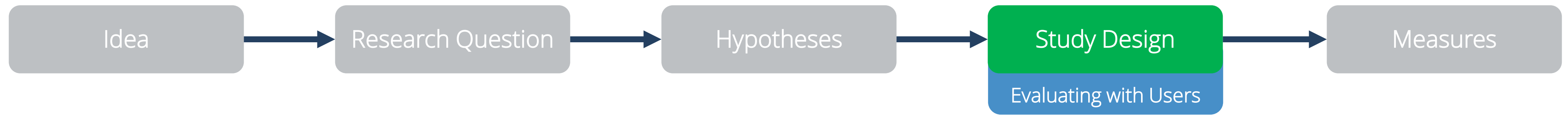
### Controlled Experiment / Lab Studies:

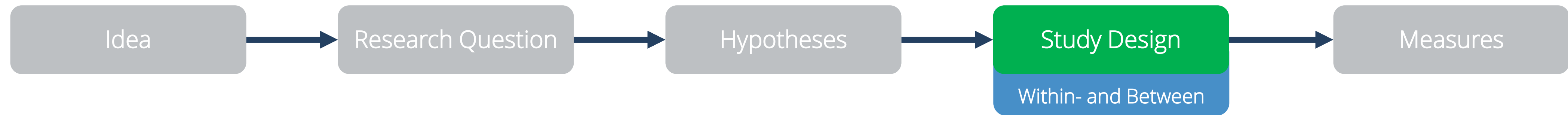
*A lab experiment is a **controlled environment**, that typically is used to isolate variables and establish cause-and-effect relationships.*

### Field study:

*A field study is a **real (or natural) situation**, happening in the actual environment and using real tasks (rather than tasks concocted by the experimenter) to observe phenomena in their real-world context.*



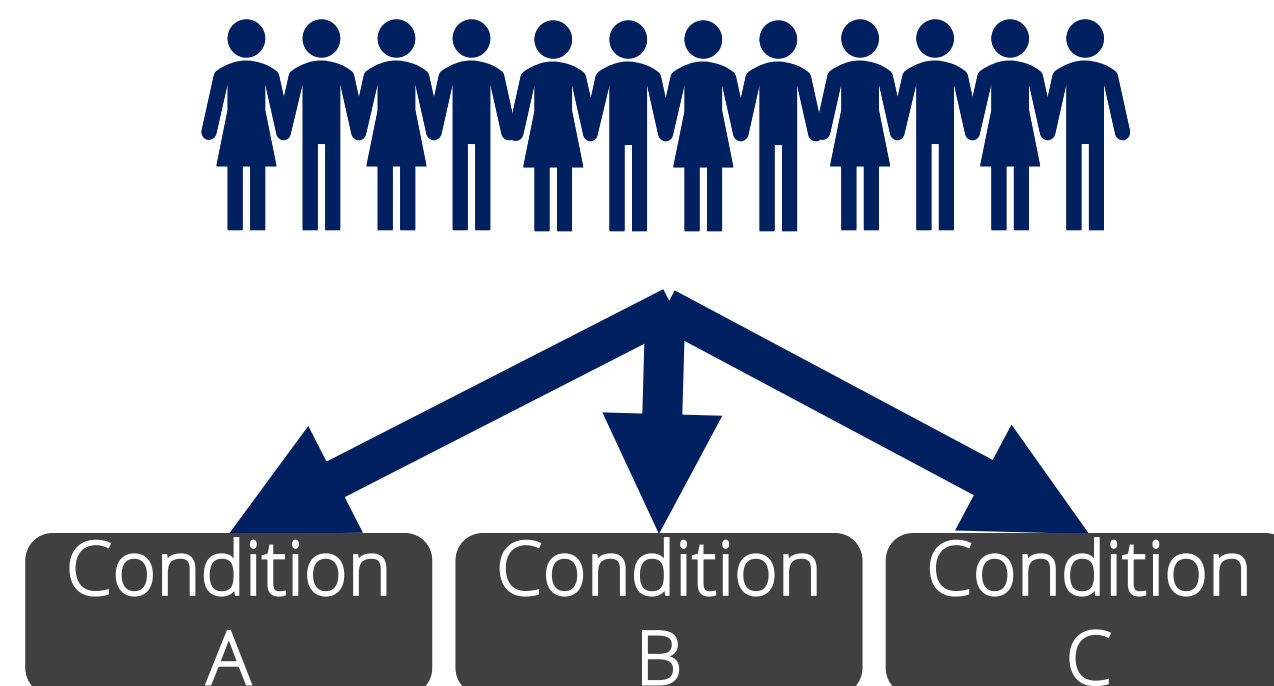




### Within-Subject Design:

- Each participant is exposed to all experimental conditions.
- Only **one group of participants** is needed for the entire experiment.

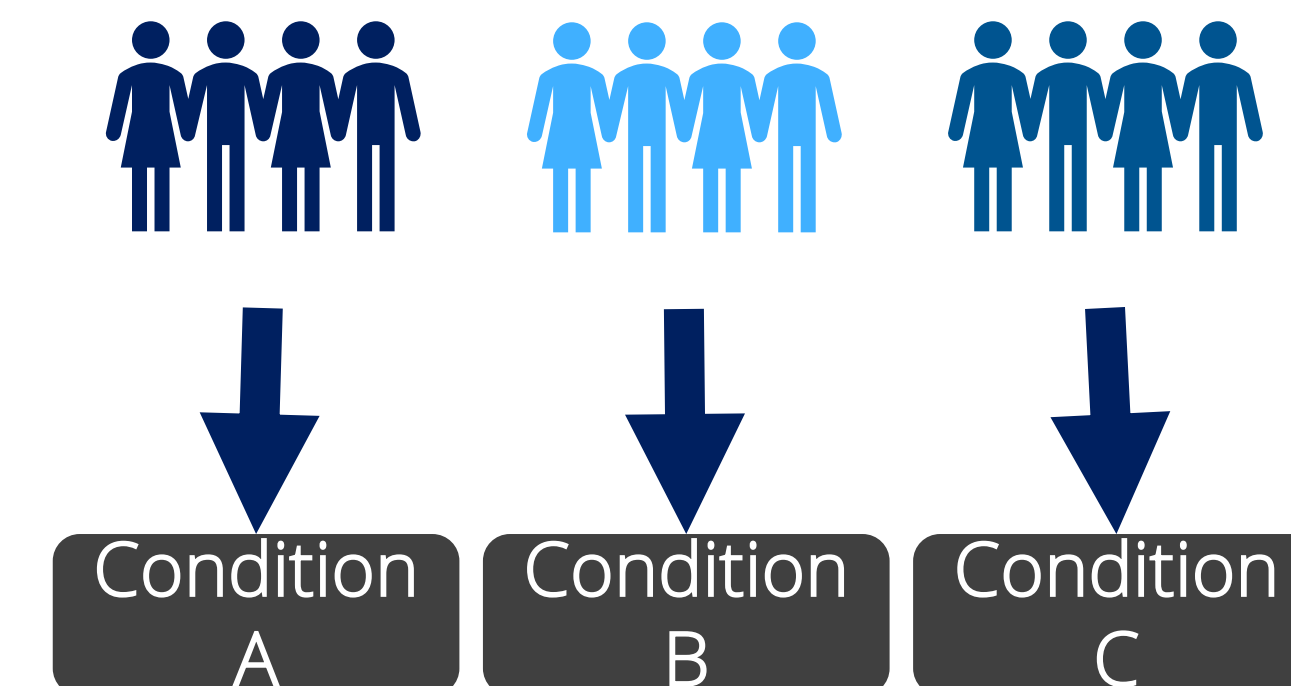
- + Smaller number of participants needed for the same amount of data.
- + Smaller influence of interpersonal differences.
- Learning effects between conditions\*



### Between-Subject Design:

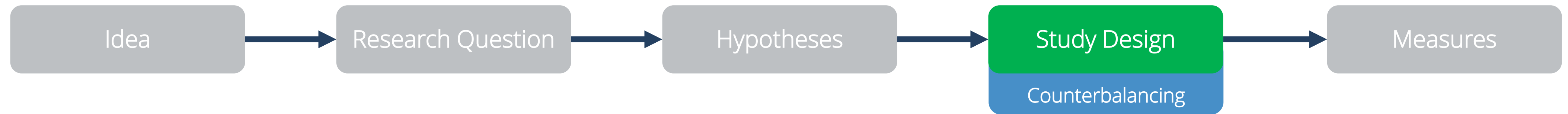
- Each participant is only exposed to one experimental condition.
- The number of participant **groups** directly correspond to the number of experimental conditions.

- + Simple Design
- + Low influence of fatigue (or other effects that appear over time)
- Large number of participants needed.
- Impact of interpersonal differences.



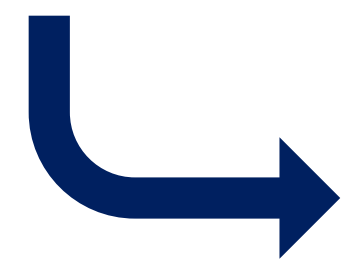
\* For example: participant may perform better when subsequently completing the same tasks using another interface.



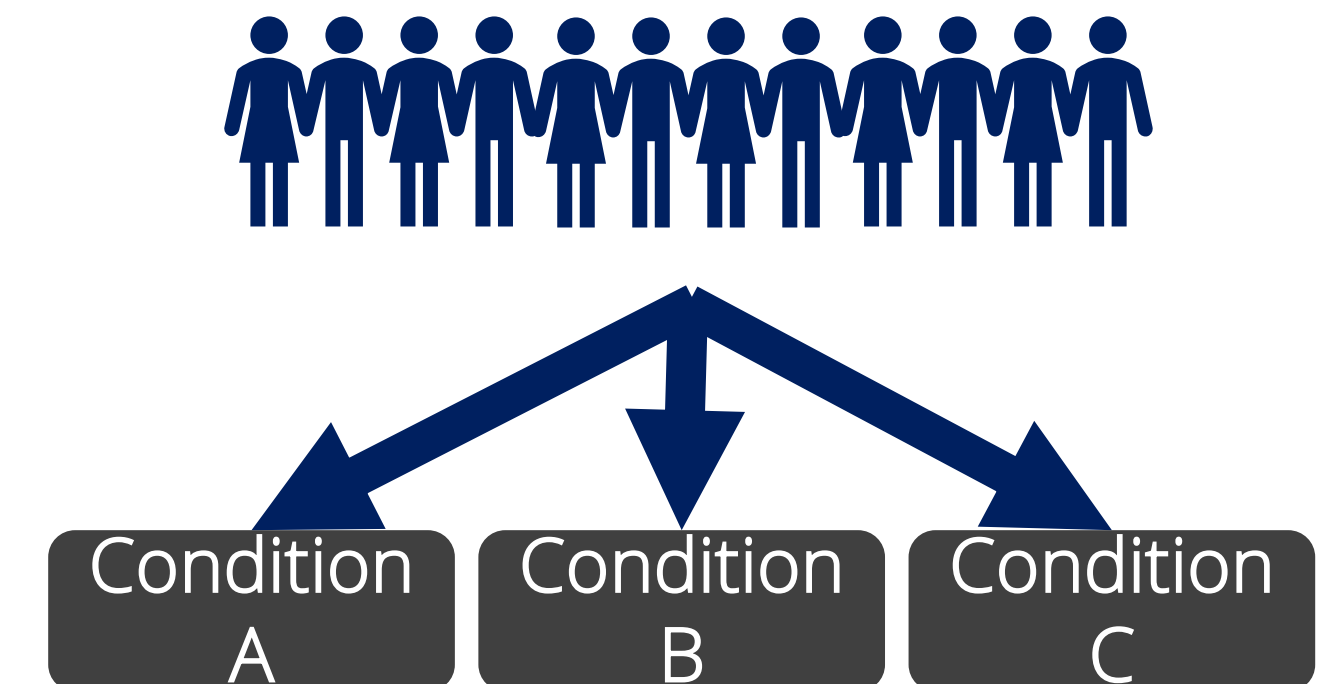


## Counterbalancing:

- Learning effects between conditions\*



Solution: Counter-balance the order of conditions across the participants.



## Balanced Latin Square

The balanced latin square ensures that each condition will precede each other condition exactly once.

- **Not all participants perform the conditions in the same order**
- *Not just:* P1-P4: ABCD
- **But rather:** P1: ABDC; P2: BCAD; P3: CDBA; P4: DACB
- **Learning effects** are spread across conditions and cancel-out
- Also protects against **carry-over effects**:  
*Something learned in a condition can be transferred to another condition.*

n = 3

A	B	C
B	C	A
C	A	B
C	B	A
A	C	B
B	A	C

n = 4

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

Important: Balanced Latin Squares for odd numbers of conditions are **2n** arrays!  
(e.g.,  $n=3$  results in 6 orders,  
 $n=4$  in 4;  $n=5$  in 10;  $n=6$  in 6; ...)

\* For example: participant may perform better when subsequently completing the same tasks using another interface.

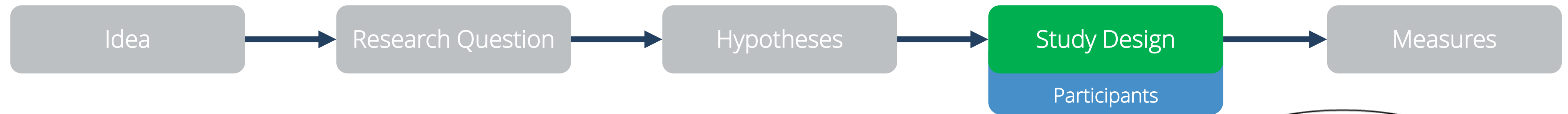


## How many participants do I need?

- Use tools to calculate/estimate your participants
- Tools need input that needs to be determined or estimated
- This can be hard and sometimes impossible, leading to the use of established, predefined values
- Not using such tools can lead to under- or overpowered study
  - Underpowered: too few samples to answer the research question of interest
  - Overpowered: too many samples show significances that are not relevant (effect size)

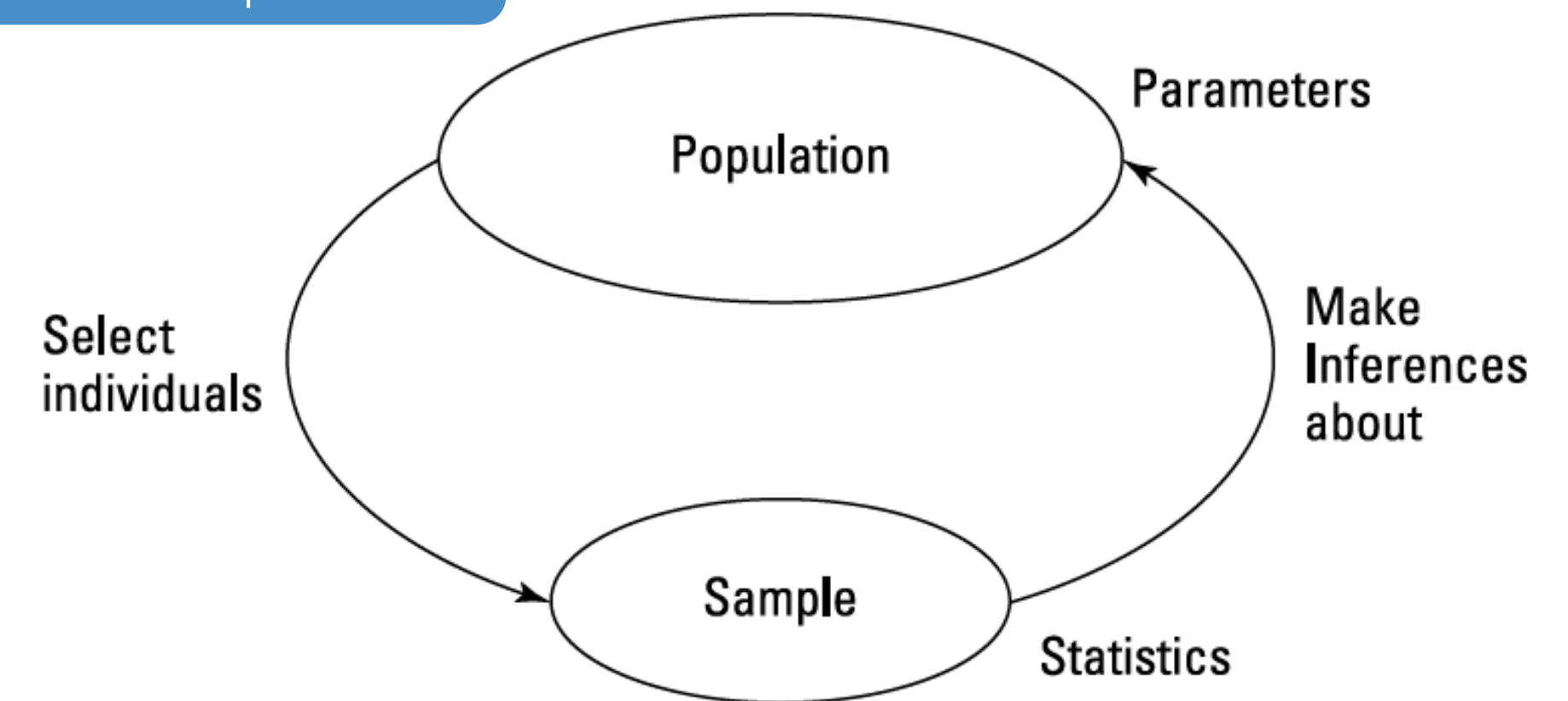






## Sampling:

Typically, you don't want to make statements about individuals but **larger populations**.

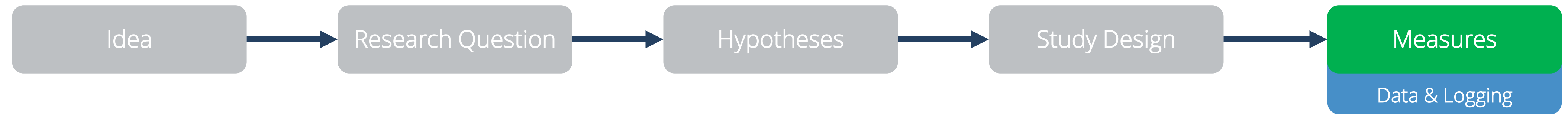


- You are **not** interested in individuals but in entire populations.
- However, we **usually** cannot test the entire population

Example: "Children are faster using system A" -> We would need to test the system with all children on the planet.

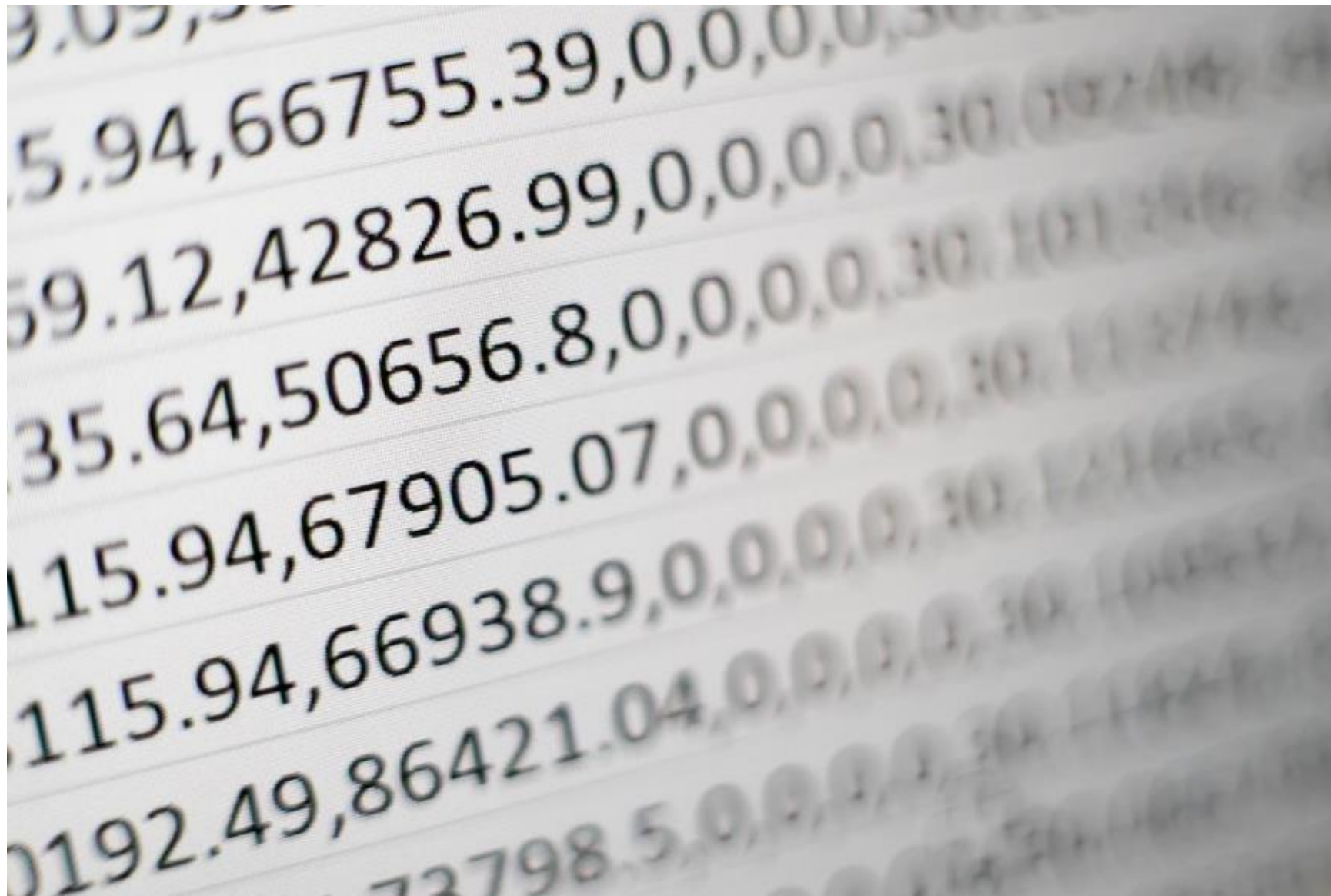
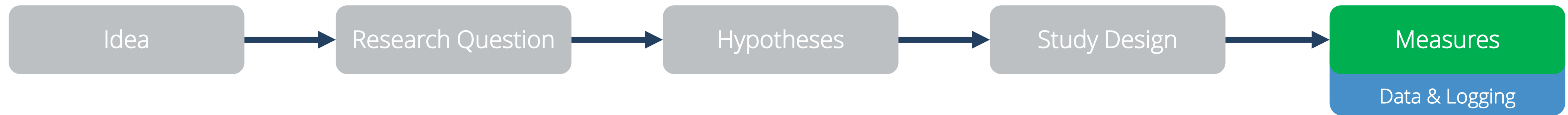
That's why we need to **sample** parts of the population.

- We use this sampled group to perform tests.
- Based on the results, we **make inferences** about the underlying population.
- **Significance tests** allow us to determine how confident we are that the results observed from the sampling population can be generalized to the entire population.



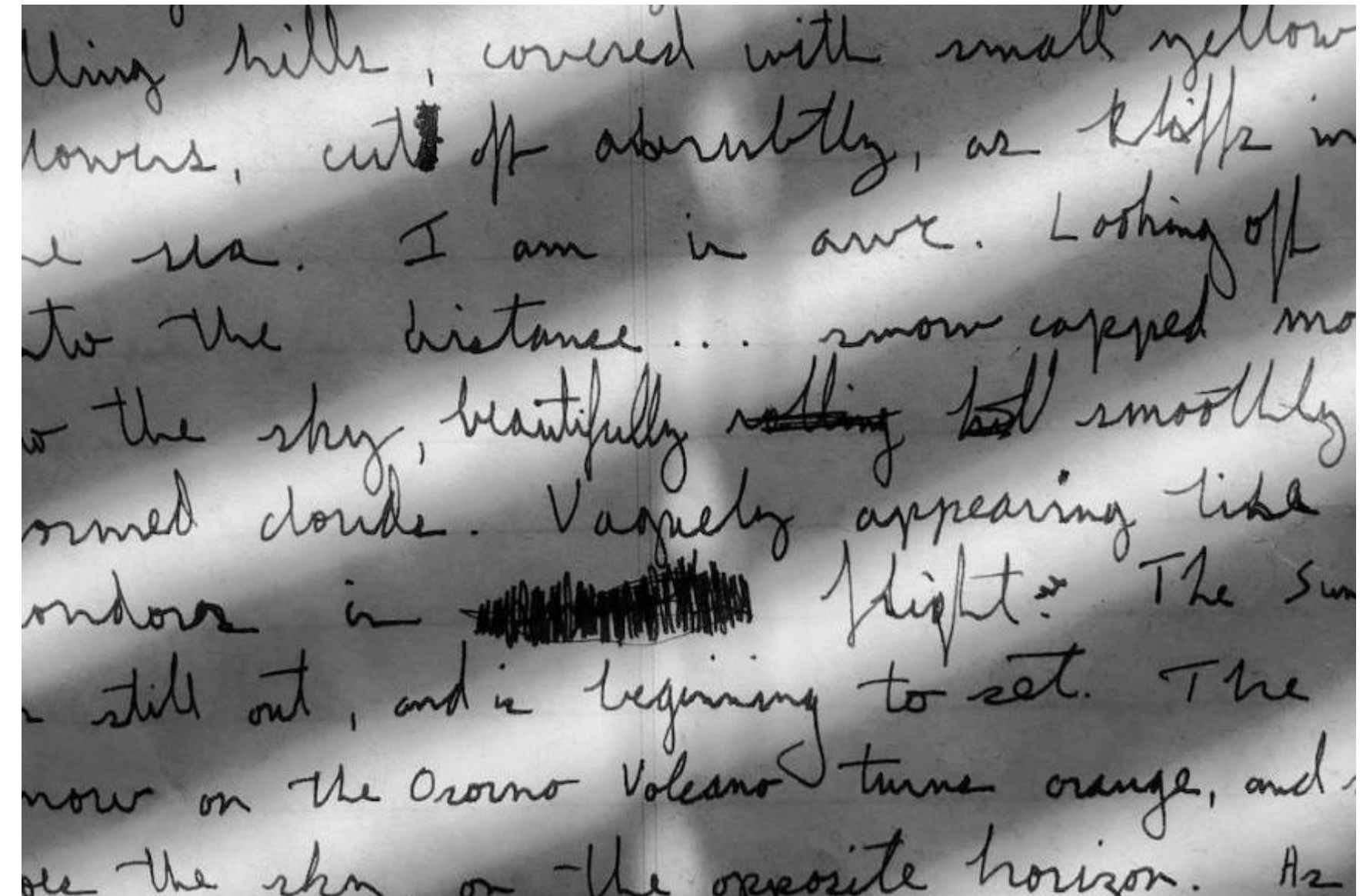
# What is data?





## Quantitative

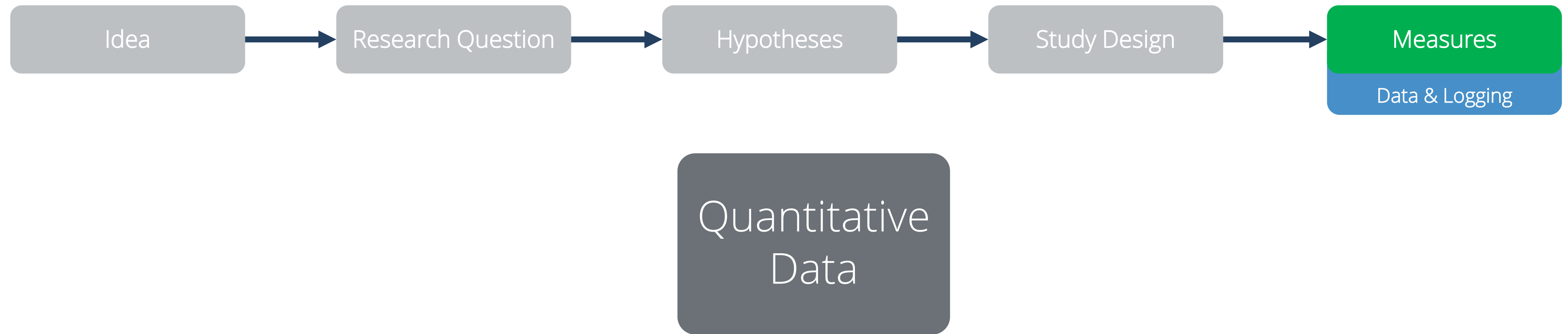
- Numeric information
- Measures, Counts, ...



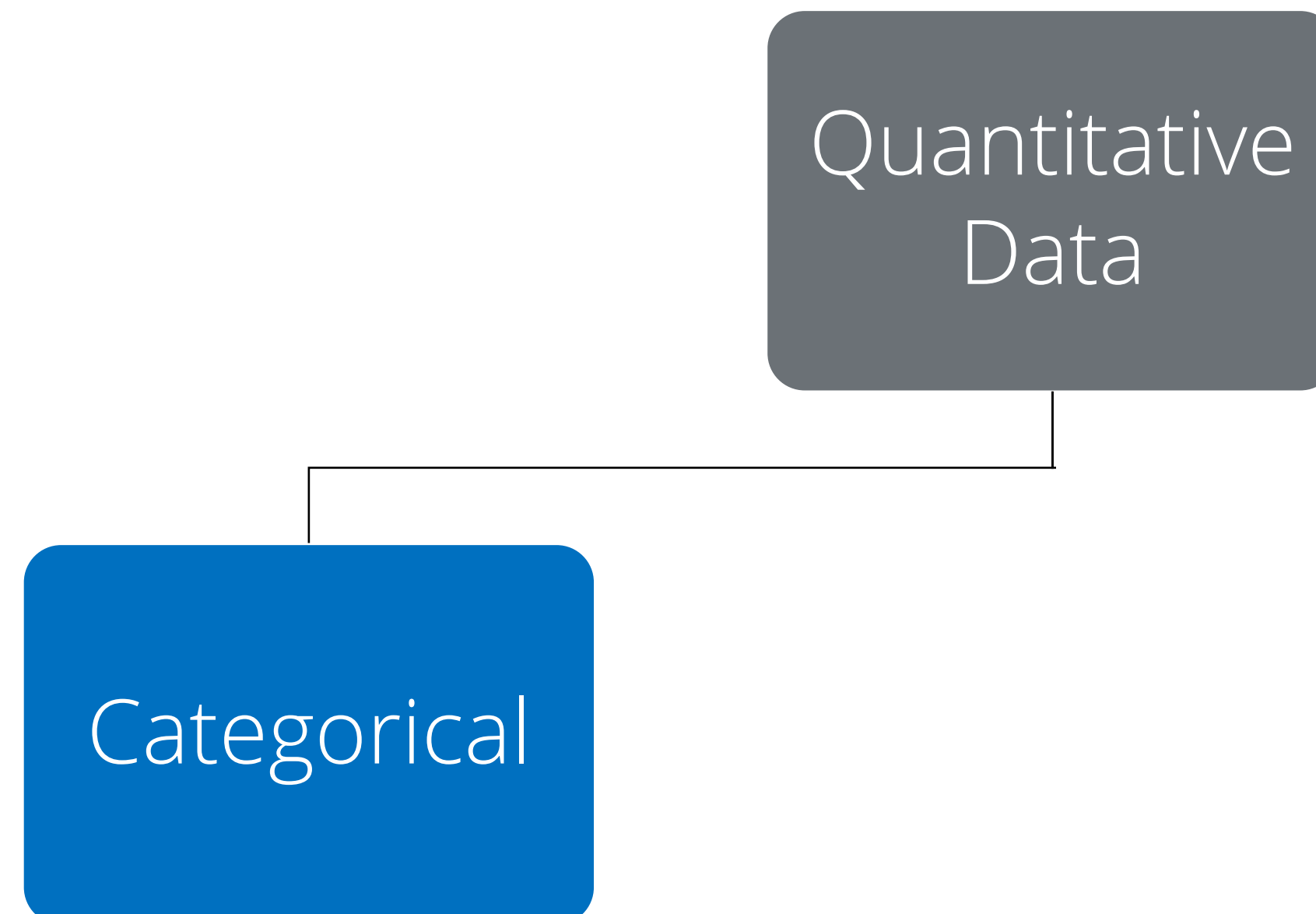
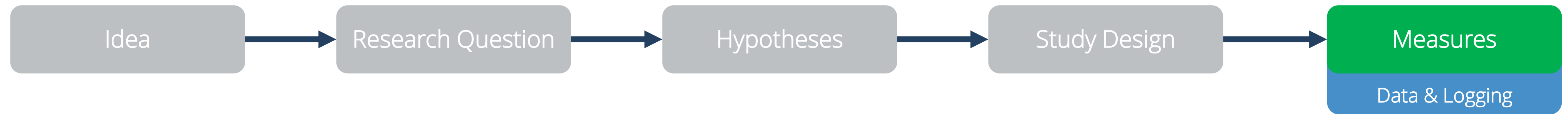
## Qualitative

- Non-numeric information
- Interviews, Audio-/ Visual-Recording, ...









Example:

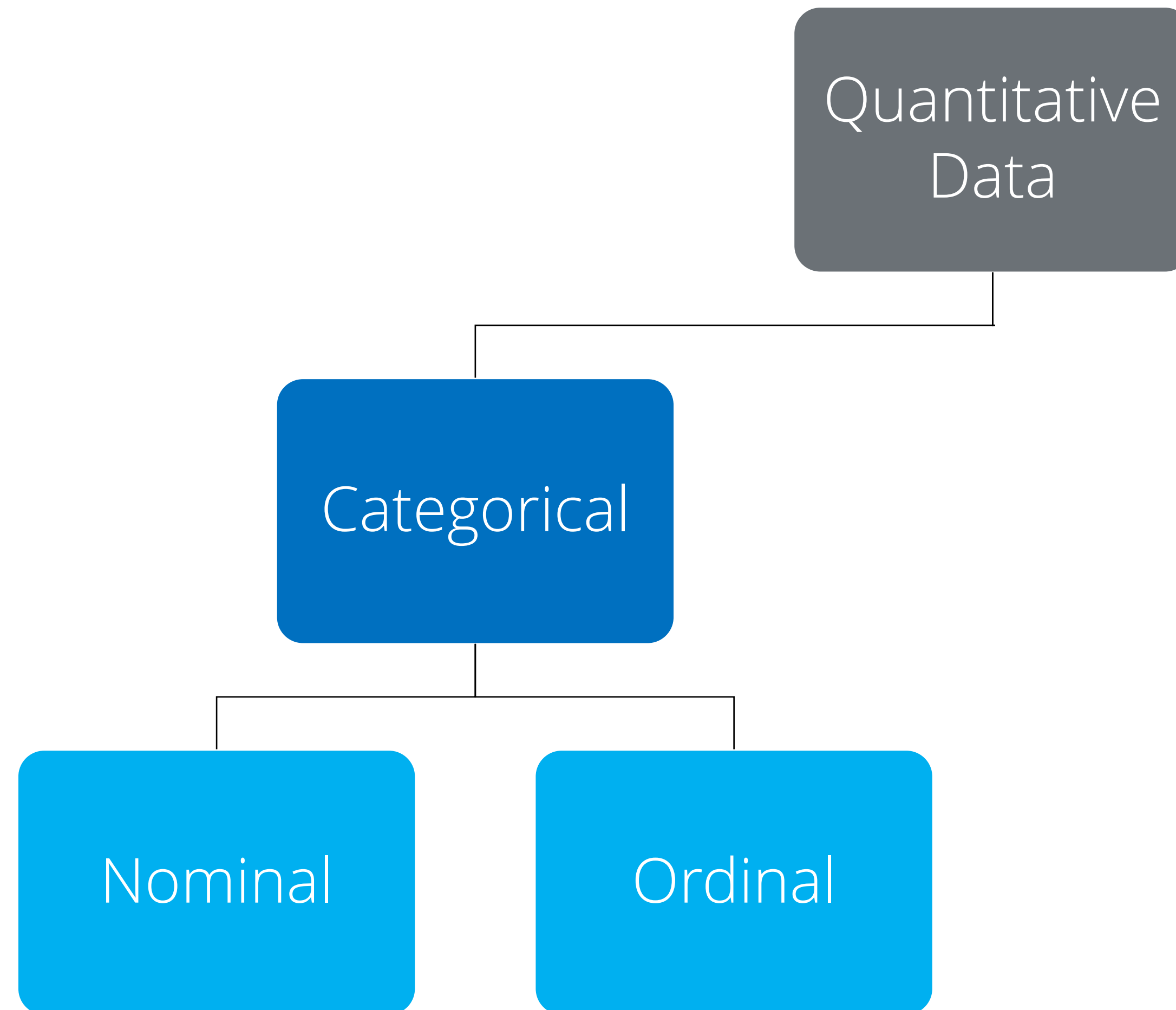
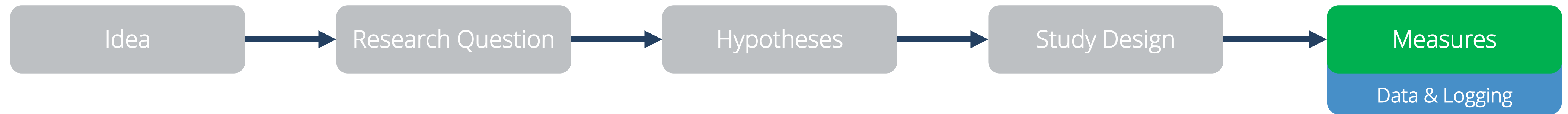
Grade	Points
1	91 - 100
2	76 - 90
3	61 - 75
4	51 - 60
5	< 50

Categorical Data represents **characteristics**

*E.g., Gender, Language, Grades, Satisfaction, ...*

Even though they can be expressed by numbers, **they do not have a mathematical meaning**

- Grades: 1 → very good, 2 → good, 3 → okay, ...
- Is a 1 twice as good as a 2, and four times as good as a 4?
- This is because the **distance between the elements is not necessarily equally distributed**



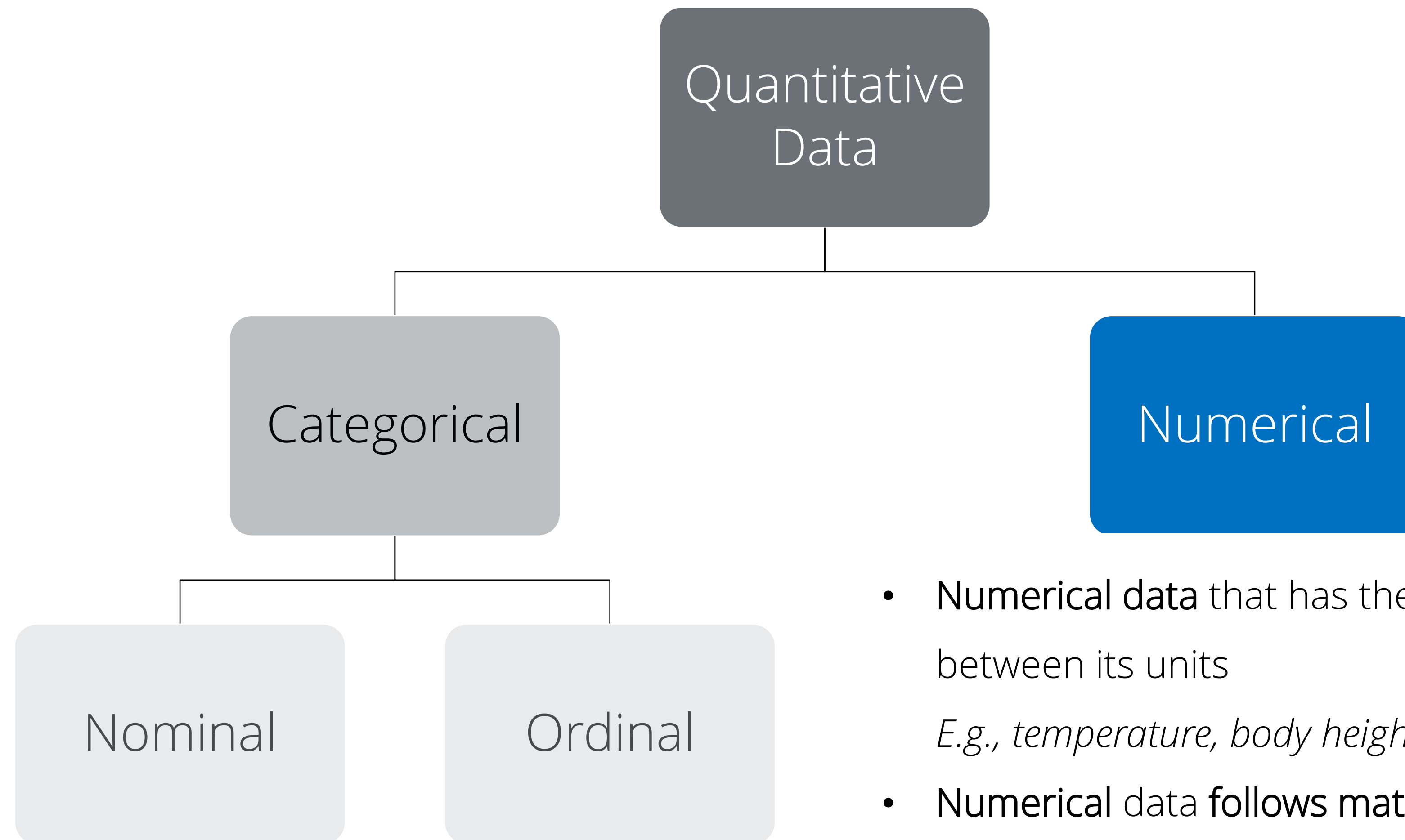
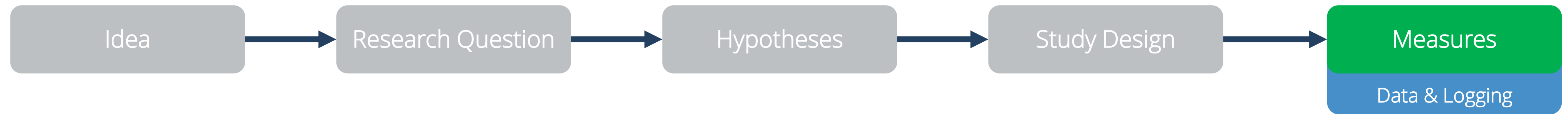
### Nominal

- Data **not** subject to a **natural order**  
*E.g. List of Countries, Languages, ...*

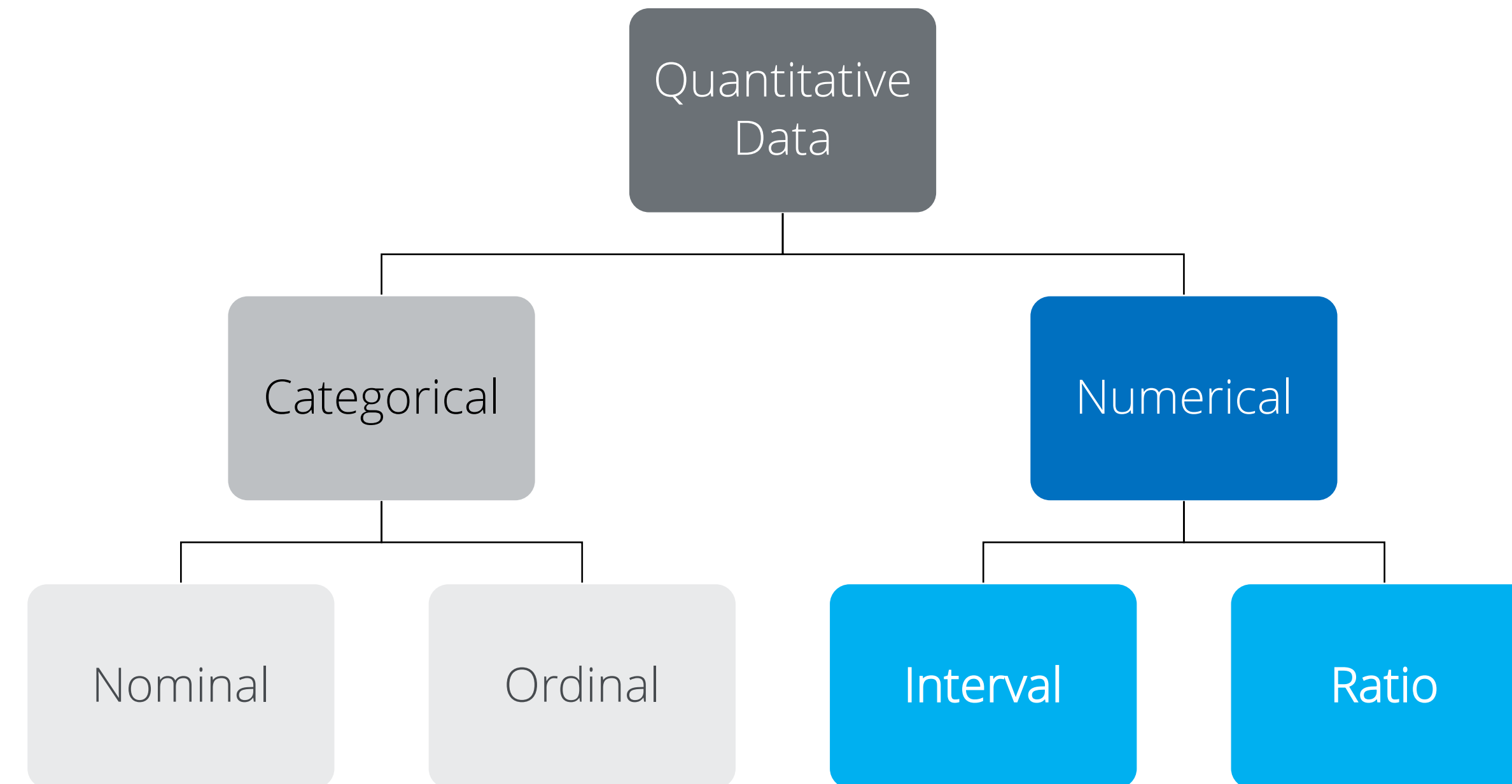
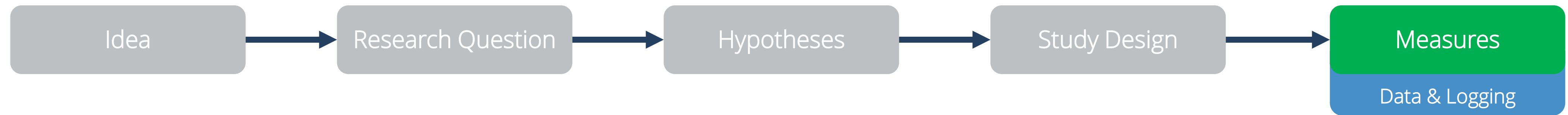
### Ordinal

- Data **being** subject to a **natural order**  
*E.g. Grades, Coffee-Sizes, ...*





- Numerical data that has the same differences between its units  
*E.g., temperature, body height, timespans, ...*
- Numerical data follows mathematical laws
  - $15^{\circ}\text{C} + 5^{\circ}\text{C} = 20^{\circ}\text{C}$
  - $100\text{cm} * 2 = 200\text{cm}$



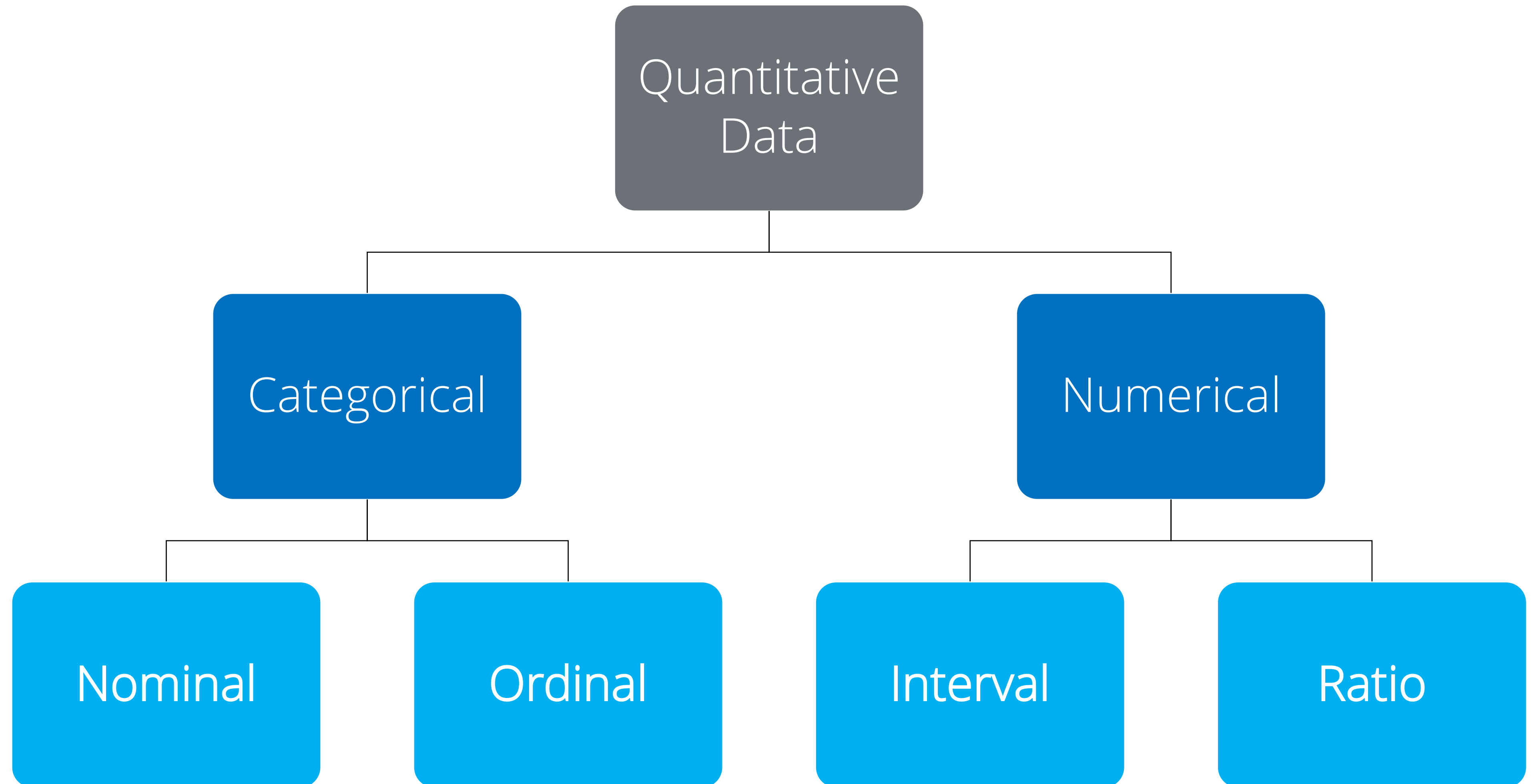
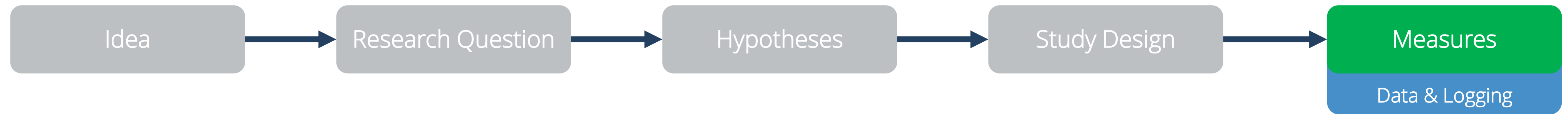
We can further divide based on existence of a **true 0**

- **Interval:** True 0 **does not** exist, the ratio has **no meaningful interpretation**  
*e.g., pH of 3 is not twice as acidic as a pH of 6*
- **Ratio:** True 0 **does** exist, the ratio of two measurements has a **meaningful interpretation**  
*e.g., 10Kg is twice as heavy as 5Kg*

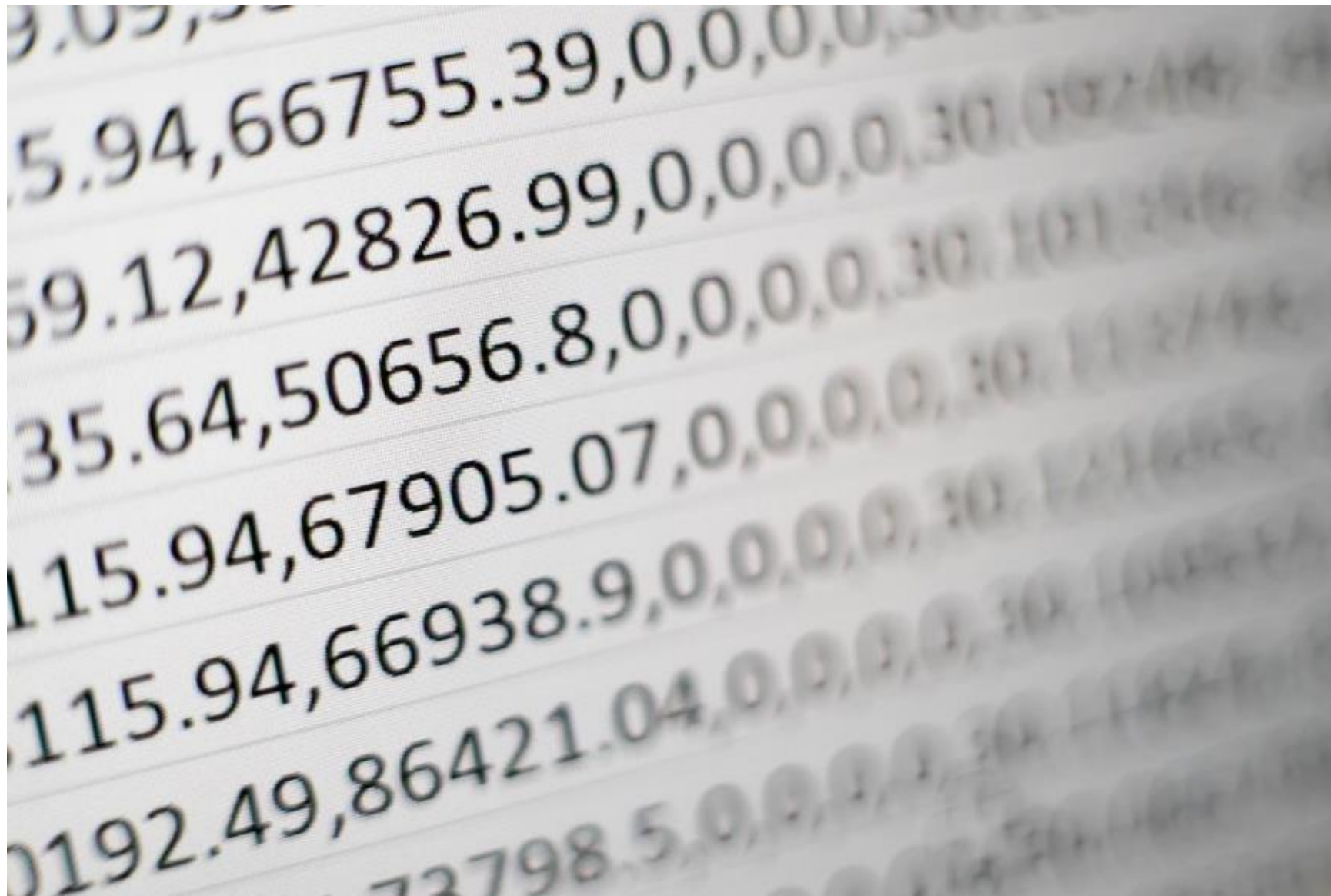
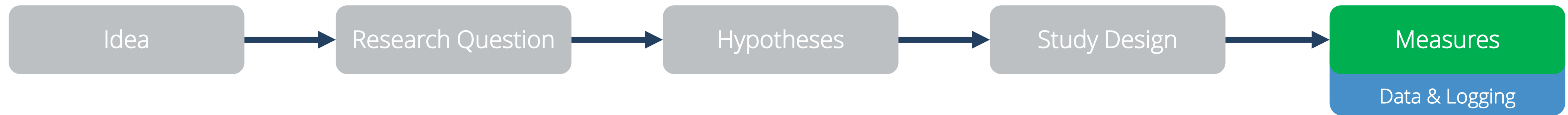
Data can be Ratio or Interval **based on the scale**

- 40°C is not twice as hot as 20°C, 300°K is twice as much thermal energy as 150°K



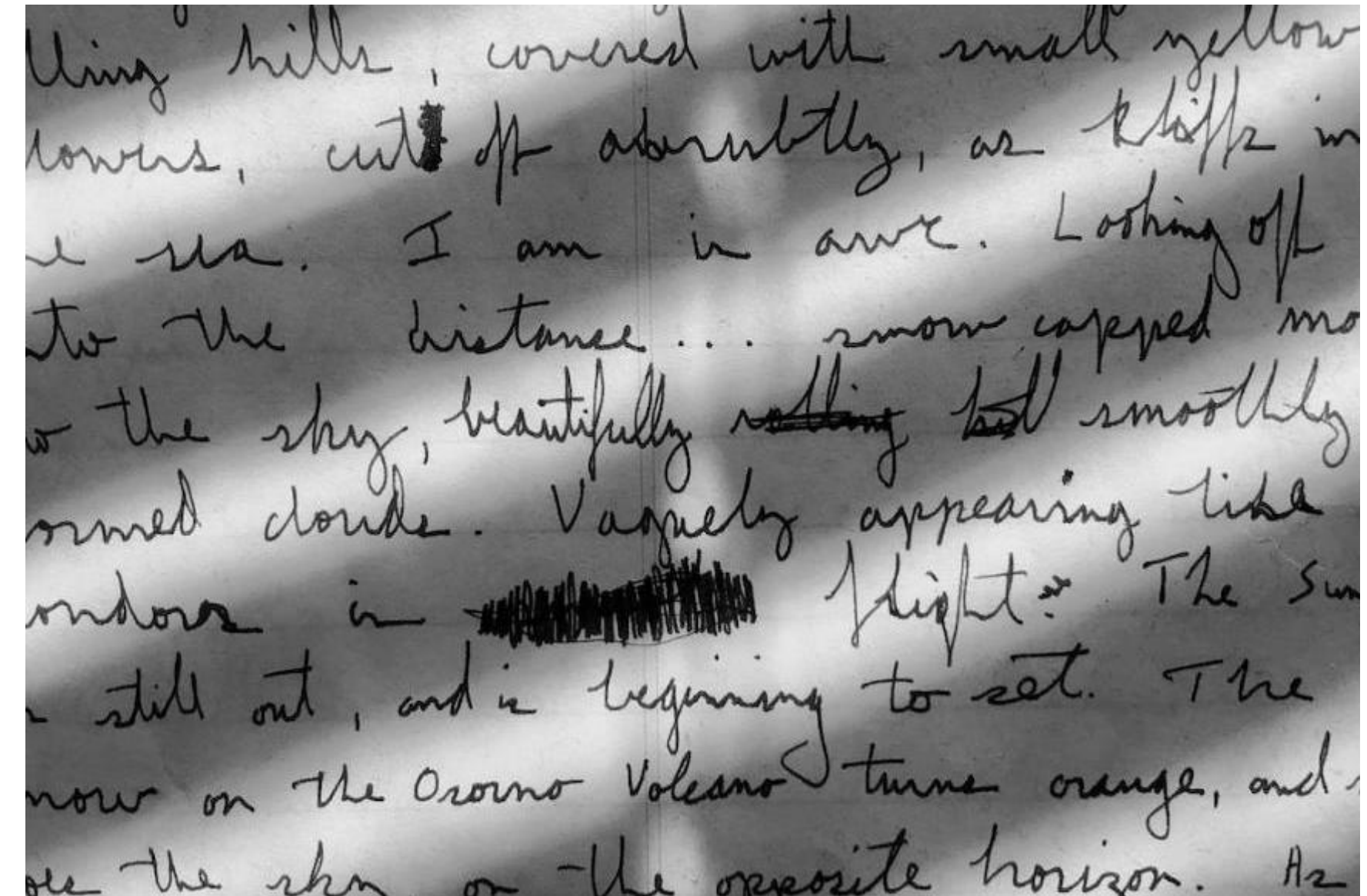






## Quantitative

- Numeric information
- Measures, Counts, ...



## Qualitative

- Non-numeric information
- Interviews, Audio-/ Visual-Recording, ...



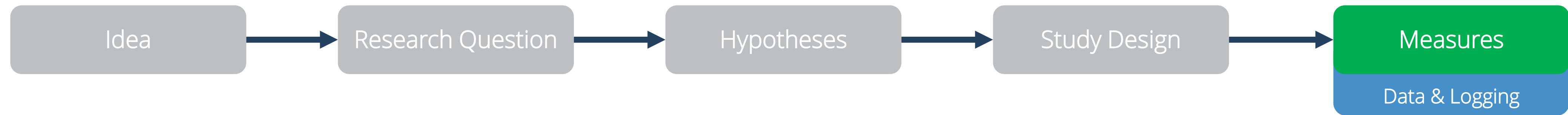


## Qualitative Data

- Non-numeric information
- Interviews, Audio-/ Visual-Recording, ...
- Observations
- Often no automated processing
- Lots of manual, human input needed







## Quantitative Data Collection

Data collection in experiments is mainly achieved through

- (Automated) logging of data and measures
- Collection of data using questionnaires

## Objective data vs. Subjective factors

- For **objective data**, *like accuracy, efficiency, ...*, use techniques to **log** or **measure** data (mostly automated)
- For **subjective factors**, *like ease-of-use, fatigue, ...*, use **questionnaires** to quantify the perception
- Subjective data can still be quantitative (e.g., measuring subjective perception of pain on a Likert scale)

In real experiments **both approaches** are often used **simultaneous**

- **During** the condition data is logged, *e.g., time needed for completion, accuracy, ...*
- **After** the condition participants fill in a questionnaire, *e.g., NASA TLX, SUS, Presence-Questionnaire*





## What to log?

It depends on

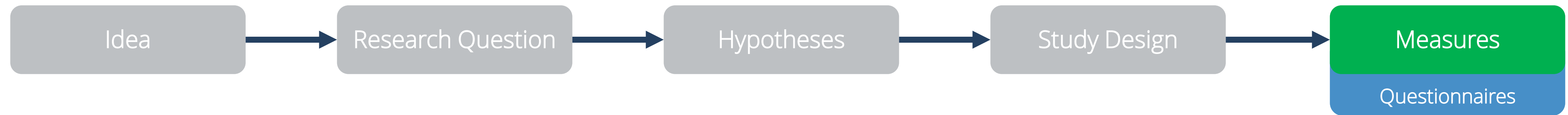
- Research question
- Hypothesis
- Reproducibility
- Fail-Safety

Examples for (*Objective*) Measurements that occur very frequently are

- *Task-Completion Time*: Could even be divided into several sub-steps
- *Reaction Time*: Time till the user starts to react to the system
- *Error Measurement*: Measure of how well users performed their tasks.
- *Movement Data*: The amount of movement (total distance) or even the trajectory

Important: Decide for a suitable **granularity**: Everything or high-level data?  
In any case, also save raw data.





## Questionnaires:

Help to quantify subjective factors, *such as ease-of-use, fun, fatigue, ...*

Questionnaires responses are measured on a **rating scale**

- Yes/ no
- Very Low, Low, High, ...
- Likert Scale
- Custom Scales

Three examples of Likert scale questionnaires are shown, each with a 5-point rating scale from 1 to 5, ranging from 'Strongly Disagree' to 'Strongly Agree'.

Example 1: I am confident I hit the correct buttons. \*

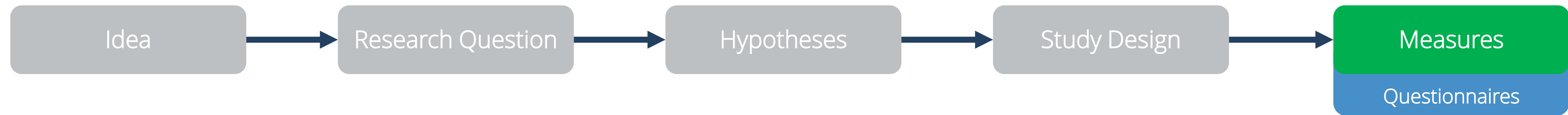
Example 2: The arrangement of the buttons was convenient. \*

Example 3: It was hard for me to press the buttons. \*

Either use **standardized** or **custom made** questionnaires

- Whenever there is a standardized questionnaire, **use it**
- When someone used a questionnaire to quantify same factor, use it or at least get inspiration





## Likert Scale

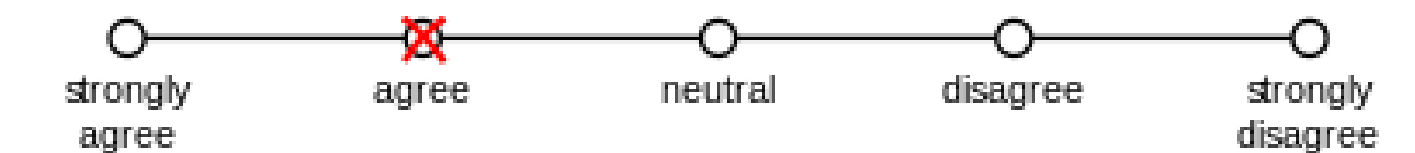
*Named after Rensis Likert*

Rather than asking questions, we **measure the participants' agreement** with certain statements:

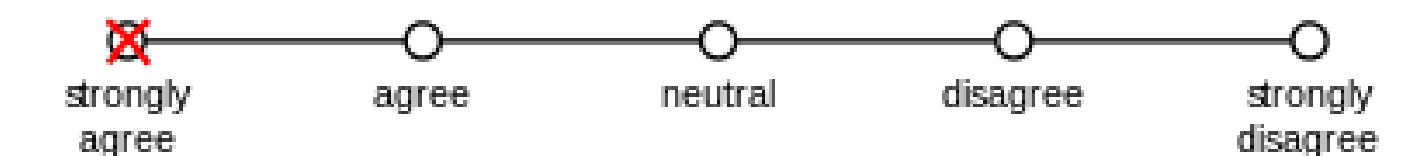
- Ranging from *strongly disagree* to *strongly agree*
- Typically in 5 or 7 steps, with a neutral middle
- Variants with even number of options forces participants to choose either positive or negative tendency

## Website User Survey

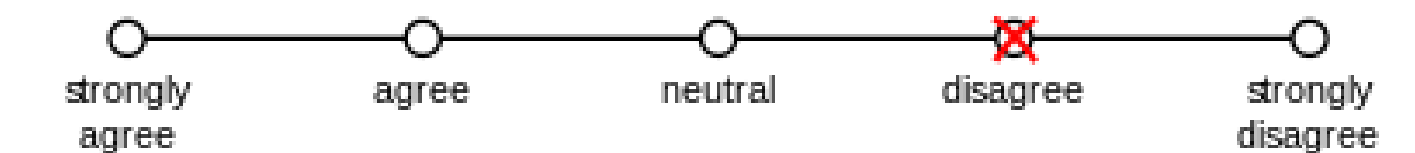
1. The website has a user friendly interface.



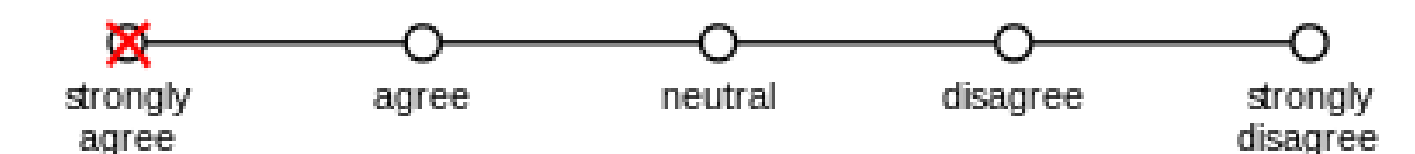
2. The website is easy to navigate.



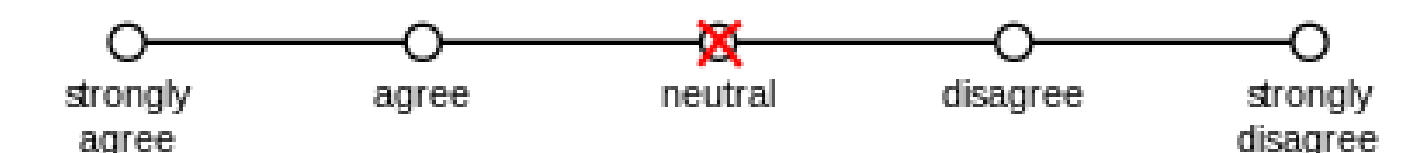
3. The website's pages generally have good images.



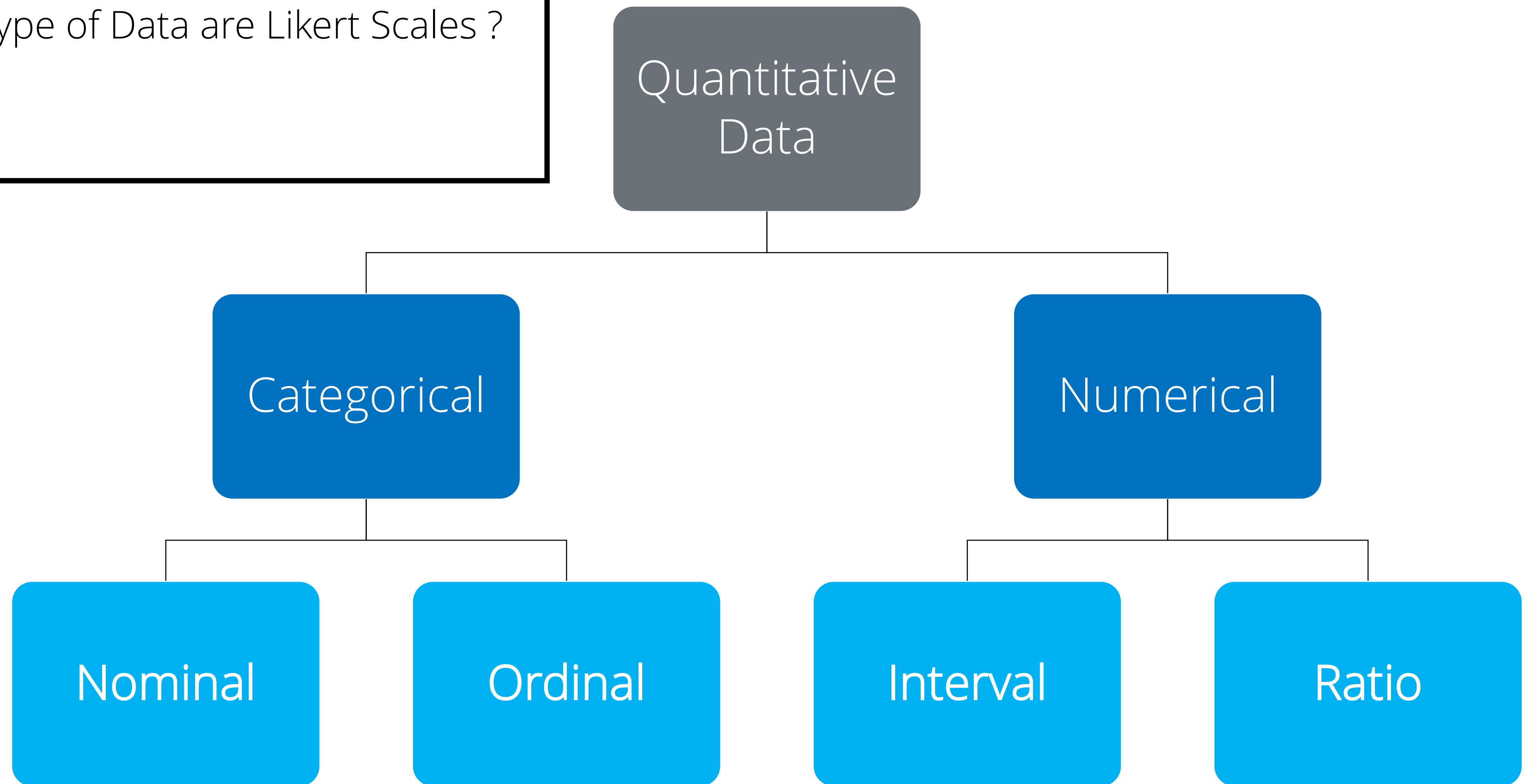
4. The website allows users to upload pictures easily.



5. The website has a pleasing color scheme.



What type of Data are Likert Scales ?







## Standard Questionnaires

Well evaluated and known questionnaires

How to assess whether a questionnaire is accepted?

- Look at who published it?
- Where was it published?
- How often was it cited?
- Was it used in previous related Work?

### Development of **NASA-TLX** (Task Load Index): Results of empirical and theoretical research

SG Hart, LE Staveland - *Advances in psychology*, 1988 - Elsevier

The results of a multi-year research program to identify the factors associated with variations in subjective workload within and between different types of tasks are reviewed. Subjective evaluations of 10 workload-related factors were obtained from 16 different experiments. The ...

☆ 77 Zitiert von: 11120 Ähnliche Artikel Alle 7 Versionen In BibTeX importieren »

### **SUS**: a “quick and dirty” usability

J Brooke - *Usability evaluation in industry*, 1996 - books.google.com

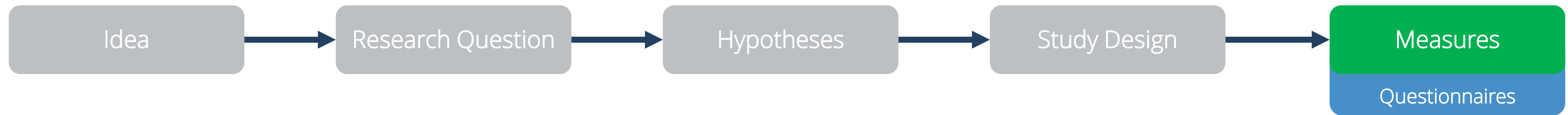
CHAPTER TWENTY ONE **SUS**: a “quick and dirty” usability Scale JOHN BROOKE Redhatch Consulting Ltd, Earley, Reading, UK US ABILITY. ANDCONT EXT Usability is not a quality that exists in any real or absolute sense. Perhaps it can be best summed up as being a general quality of ...

☆ 77 Zitiert von: 10201 Ähnliche Artikel Alle 22 Versionen In BibTeX importieren

### **NASA Task Load Index**

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date
<p><b>Mental Demand</b> How mentally demanding was the task?</p> <p>Very Low   Very High</p>		
<p><b>Physical Demand</b> How physically demanding was the task?</p> <p>Very Low   Very High</p>		
<p><b>Temporal Demand</b> How hurried or rushed was the pace of the task?</p> <p>Very Low   Very High</p>		
<p><b>Performance</b> How successful were you in accomplishing what you were asked to do?</p> <p>Perfect   Failure</p>		
<p><b>Effort</b> How hard did you have to work to accomplish your level of performance?</p> <p>Very Low   Very High</p>		
<p><b>Frustration</b> How insecure, discouraged, irritated, stressed, and annoyed were you?</p> <p>Very Low   Very High</p>		



## Custom Questionnaires

Capture **additional aspects** not covered by standard questionnaires and can be **tailored to your experiment** to be more **specific**.

Be mindful with your formulations!

### Bad Examples:

- No leading question: *How great was your experience with our awesome system?*
- No assumptive question: *What made the interaction better?*
- No double-negatives: *Was the interaction no unfun?*
- Poor answer scales: *How many hours do you use VR in your daily routine? A) 0-1 hours B) 2-6 hours C) 7-14 hours*

*Remember also good and bad examples known from Research Questions.*

I am confident I hit the correct buttons. \*

1 2 3 4 5

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

The arrangement of the buttons was convenient. \*

1 2 3 4 5

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

It was hard for me to press the buttons. \*

1 2 3 4 5

Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree





## How to get from a hypothesis to a study?

1. Identify research questions and hypotheses (*done yaye*)
2. Specify the design of the study:
  - Type of experiment
  - Within- or between-subjects design
  - Experimental conditions
  - Task description and procedure
  - Experimental measures
3. Run a pilot study and re-iterate if necessary
4. Recruit participants
5. Conduct the actual study (incl. data collection)
6. Analyze the data

### Things to consider:

- Run pilot tests
- Don't waste participants' time
- Make sure participants feel comfortable and inform them about the study procedure and task
- Guarantee privacy
- Participants are volunteers (*consent forms required*)
- Have a demographics questionnaire

### Hawthorne Effect:

- Participants of a study might **change their natural behavior** because they know that their behavior is being observed
- Most obvious in controlled (unnatural) settings
- But **also applies to field studies** in a natural environment
- Threatens external validity of study results





## How to get from a hypothesis to a study?

1. Identify research questions and hypotheses (*done yaye*)
2. Specify the design of the study:
  - Type of experiment
  - Within- or between-subjects design
  - Experimental conditions
  - Task description and procedure
  - Experimental measures
3. Run a pilot study and re-iterate if necessary
4. Recruit participants
5. Conduct the actual study (incl. data collection)
6. **Analyze the data**





## Data Analysis:

After conducting a study and collecting data, the **collected data has to be analysed** in order to answer the initial research questions.

## Different Approaches

- Descriptives (Mean, Median, ..)
- Quantitative Data Analysis (t-tests, ANOVA, non-parametric tests, ..)
- Qualitative Data Analysis (Codebooks, ..)
- Plotting of Data (Boxplots, Histograms, ..)
- ... etc

Required to identify statistical significance



## Descriptives:

Mathematical Summaries or Measures that provide an Overview of the Data.

- **Mean**: The average value of a distribution.

Adding up all numbers, then divide by how many numbers there are.

Not suitable for ordinal data that might not be equi-distant.

- **Median**: Middle value of the data.

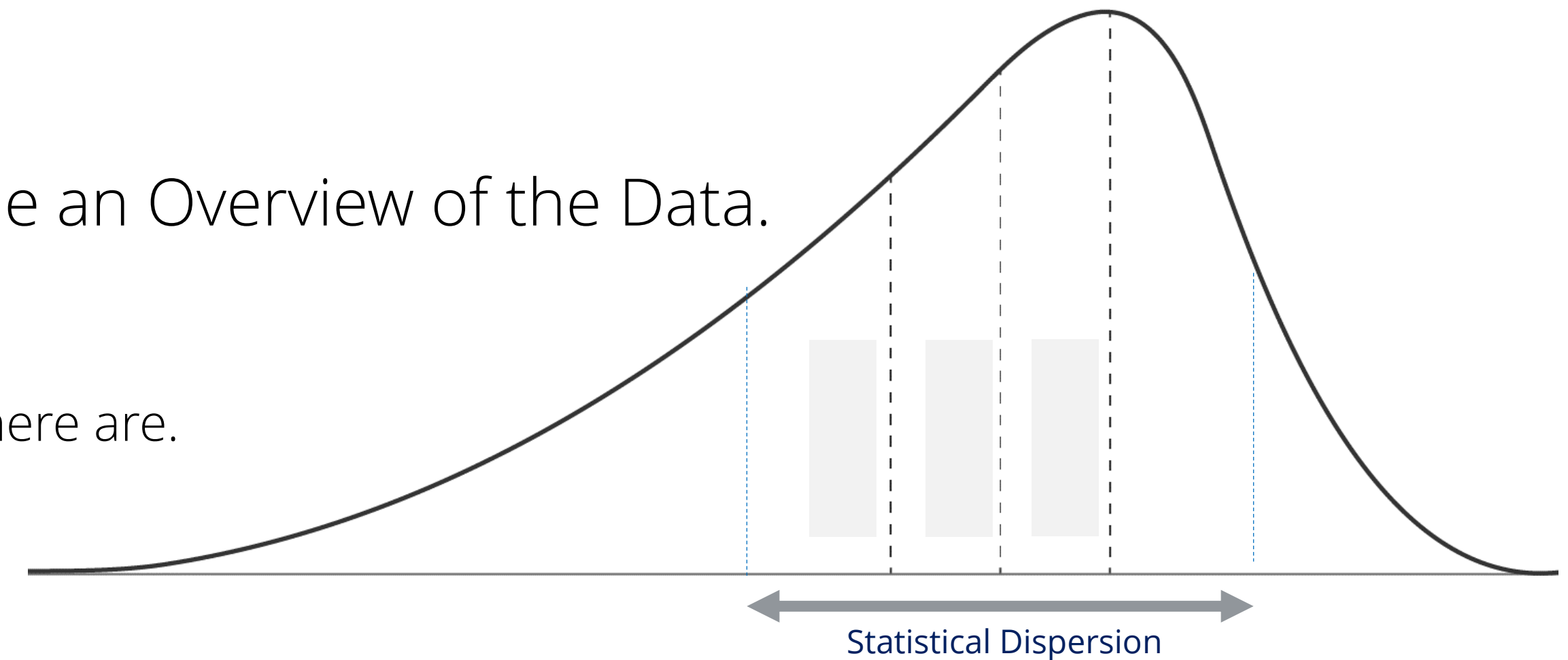
Sort data in numerical order, then pick the element in the middle.

Also works for ordinal data.

- **Standard Deviation** and Variance: Measure of how spread-out the values are.

Variance = Average of the squared differences from the mean; SD = Square root of Variance

- **Mode**: Answers with highest number of responses.
- **Min/Max**: Minimum and maximum values of responses.
- **Interquartile Range**: Difference between 75<sup>th</sup> and 25<sup>th</sup> percentiles.
- **Median Absolute Deviation**: Median of all absolute deviations from the data's median.

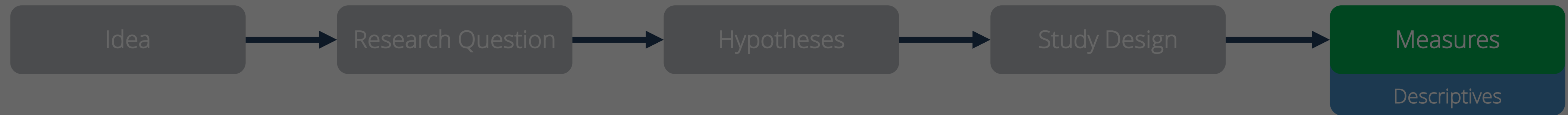


Example Data: [1, 1, 1, 2, 3, 3, 4, 4, 5]

Mean: 2.67, Median: 3, Mode: 1

SD: 1.5, Min/Max: [1, 5], IQR: 4-1=3, MAD: 1





## Descriptives:

Mathematical Summaries or Measures that provide an Overview of the Data.

- **Mean:** The average value of a distribution.

Descriptives help for a **first understanding** of the data.

However, they can **only indicate if somewhere is a difference** which might be only a coincidence.

For any further statement, **we have to use quantitative analysis.**

Variance = Average of the squared differences from the mean; SD = Square root of Variance

- **Mode:** Answers with highest number of responses.
- **Min/Max:** Minimum and maximum values of responses.
- **Interquartile Range:** Difference between 75<sup>th</sup> and 25<sup>th</sup> percentiles.
- **Median Absolute Deviation:** Median of all absolute deviations from the data's median.

Example Data: [1, 1, 1, 2, 3, 3, 4, 4, 5]

Mean: 2.67, Median: 3, Mode: 1

SD: 1.5, Min/Max: [1, 5], IQR: 4-1=3, MAD: 1



## Statistical Significance Testing and the “p-value”:

Significance testing allows us to quantify the chance that an event occurred just by random chance.

We speak of **statistical significance** when it is very unlikely to have occurred given  $H_0$  (*null hypothesis*).

This probability is referred to as “p-value”:

- A small p-value (usually  $<0.05$ ) indicates that the probability of a difference occurring by chance is very unlikely.
- Assuming that there are no differences in reality, p gives the probability of seeing this effect in the data by chance.
- This means: the lower the p-value, the greater the statistical significance of the observed difference.

In HCI, this is typically:

$H_0$ : All samples originate from the same population.

- That means: There is no difference in the data.
- If we can reject  $H_0$  based on our analysis, we say:

*The probability that all of this data that we have seen belongs to the same underlying population is very low. Therefore, we assume that there is a difference between the groups.*





## Different statistical tests required depending on Study Design:

Design	# of IVs	# of Levels	Parametric Test	Non-Parametric Test
Between-groups	1	2	Independent-samples t-test	Mann-Whitney-U-Test
		3+	One-way ANOVA	Kruskal-Wallis test
	2+	2+	Factorial (n-way) ANOVA	Scheirer-Ray-Hare test
Within-groups	1	2	Paired-samples t-test	Wilcoxon test
		3+	Repeated-Measures ANOVA	Friedman test
	2+	2+		Aligned Rank Transform ANOVA
mixed	2+	2+	Split-Pot ANOVA	

HCI researcher are typical no statisticians nor mathematicians.

However, it is **important to understand statistical concepts** and be able to use established statistical analysis.

*Useful tools are:* R, Jamovi, SPSS, Python, ...

**Note:** You don't need to remember all of them, but to understand that there are different methods and that it depends on the design, IVs, and their number of levels. We will go into detail of two types of testing methods; namely t-tests and ANOVA.



## T-Tests:

T-Tests are statistical tests used to determine if there is a **significant difference** between the means of two **groups** and tests whether two sample sets follow the same distribution.

Consider the following **example experiment** (*within-subjects design*) with two conditions (C1 & C2).

Some performance metric says, higher values are better.

**WRONG**

First Intuition:

Condition 2 has higher mean than Condition 1 -> C2 was better than C1.

→ Therefore, it is important to have t-tests (*or in general statistical methods*) to tell whether the two samples are from the same or a different distribution, *i.e., there is an actual difference.*

Participant	Condition 1	Condition 2
1	-0,23	-0,43
2	0,36	-0,48
3	1,3	2,8
4	-0,1	0,43
5	1,2	0,068
6	-0,097	1,1
7	-0,64	-0,69
8	1,8	1,3
9	0,17	0,66
10	-2,3	0,87
<b>Mean</b>	0,1463	0,5628
<b>Variance</b>	1,3429	1,0939





T-Tests:

T-Tests are statistical tests used to determine if there is a significant difference between the means of two groups and tests whether two sample sets follow the same distribution.

→ Different distributions assumed if the significance is below 5% ( $p < 0.05$ ) again, which means, there is a significant difference!  
*(But also, that there is a chance smaller 5% that they are still from the same distribution)*

Example using SPSS:

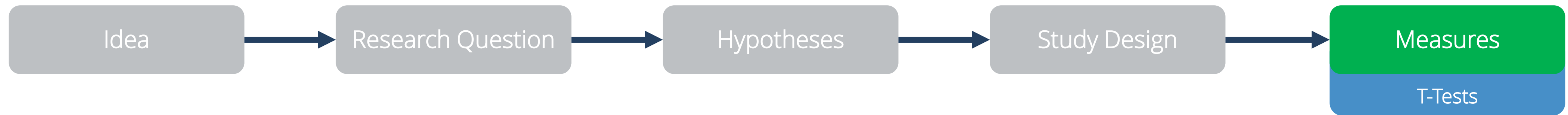
**Statistik für Stichproben mit paarigen Werten**

		Mittelwert	H	Standardabweichung	Standardfehler Mittelwert
Paar 1	Cond1	,1463	10	1,15885	,36646
	Cond2	,5628	10	1,04594	,33076

**Korrelationen für Stichproben mit paarigen Werten**

		H	Korrelation	Sig. p
Paar 1	Cond1 & Cond2	10	,327	,356

Participant	Condition 1	Condition 2
1	-0,23	-0,43
2	0,36	-0,48
3	1,3	2,8
4	-0,1	0,43
5	1,2	0,068
6	-0,097	1,1
7	-0,64	-0,69
8	1,8	1,3
9	0,17	0,66
10	-2,3	0,87
Mean	0,1463	0,5628
Variance	1,3429	1,0939



## T-Tests:

T-Tests are statistical tests used to determine if there is a **significant difference** between the means of two groups and tests whether two sample sets follow the same distribution.

→ Different distributions assumed **if the significance is below 5%** ( $p < 0.05$ )

again, which means, there is a significant difference!

*(But also, that there is a chance smaller 5% that they are still from the same distribution)*

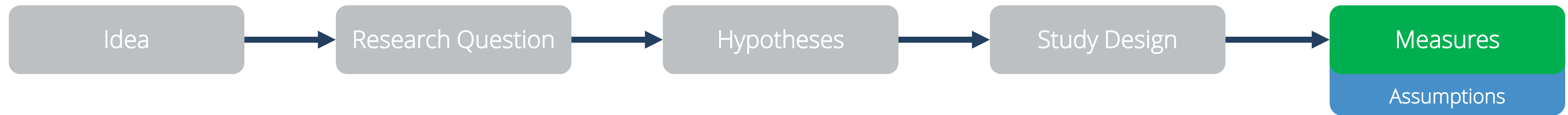
**Beware:** T-tests – similar to other (parametric) methods – have a set of assumptions that must be fulfilled.

For t-tests:

- Continuous Scale (intervals or ratio)
- Homogeneity of variance
- Normal Distribution of data

Participant	Condition 1	Condition 2
1	-0,23	-0,43
2	0,36	-0,48
3	1,3	2,8
4	-0,1	0,43
5	1,2	0,068
6	-0,097	1,1
7	-0,64	-0,69
8	1,8	1,3
9	0,17	0,66
10	-2,3	0,87
<b>Mean</b>	0,1463	0,5628
<b>Variance</b>	1,3429	1,0939



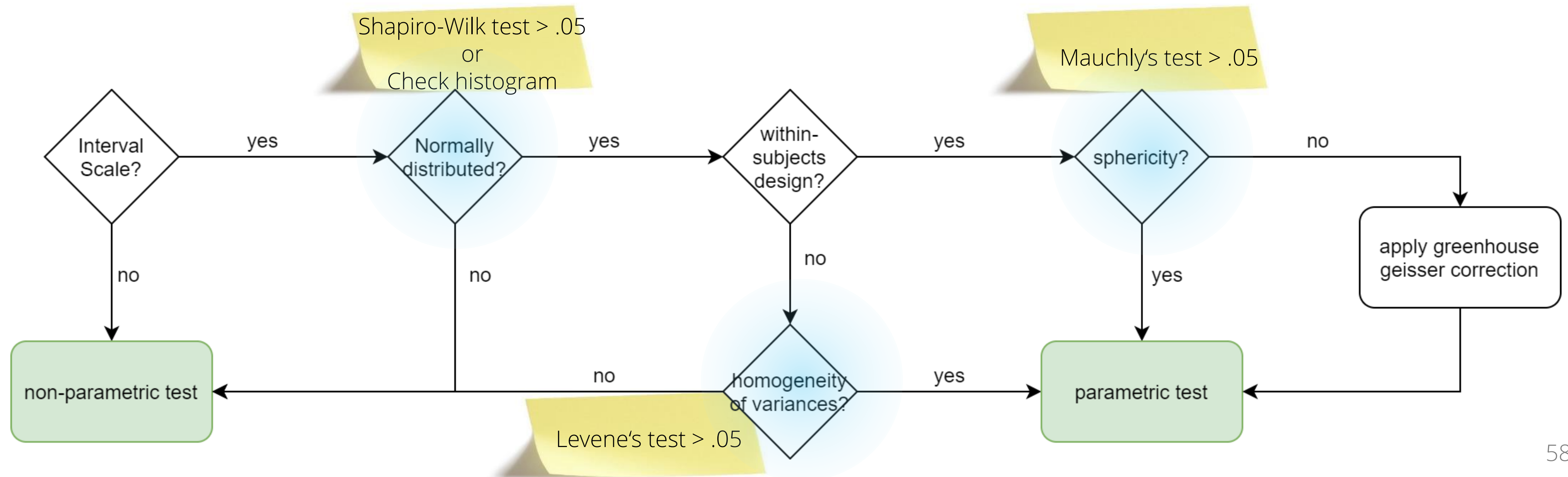


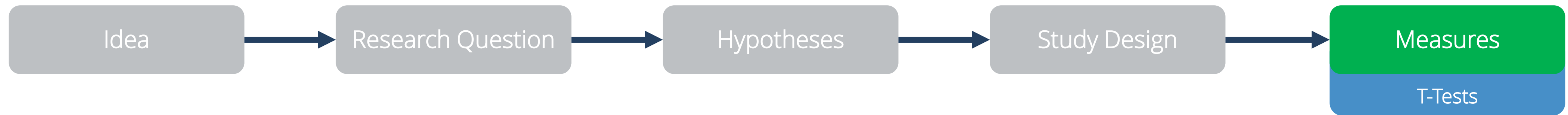
## Assumptions for parametric tests (not limited to t-tests!):

There are two basic groups of tests. **Parametric tests** and **non-parametric tests**.

**Parametric tests have higher power** (i.e., it is easier to obtain significant results), but they have more rigorous requirements for the data (“assumptions”).

*Whenever possible: Use parametric tests. If one of the assumptions is not fulfilled, then switch to the non-parametric equivalent.*





## T-Tests:

T-Tests are statistical tests used to determine if there is a **significant difference** between the means of two groups and tests whether two sample sets follow the same distribution.

What about a third condition?

**Naive idea:** Use multiple t-tests and compare pair-wise.

→ In the example, three tests (*1-2, 1-3, and 2-3*)

Problem: Type I errors ( *$\alpha$ -inflation*)

Single t-tests have a **5% error chance** that would accumulate for multiple tests:

*In the example:  $0.95 * 0.95 * 0.95 = 0.86 = 14\%$  chance of error!*

Participant	Condition 1	Condition 2	Condition 3
1	-0,23	-0,43	1,11
2	0,36	-0,48	1,19
3	1,3	2,8	0,97
4	-0,1	0,43	1,3
5	1,2	0,068	1,42
6	-0,097	1,1	1,18
7	-0,64	-0,69	1,4
8	1,8	1,3	1,32
9	0,17	0,66	1,03
10	-2,3	0,87	0,93
<b>Mean</b>	0,1463	0,5628	1,1850
<b>Variance</b>	1,3429	1,0939	0,0304





## Statistical Methods might result in Errors:

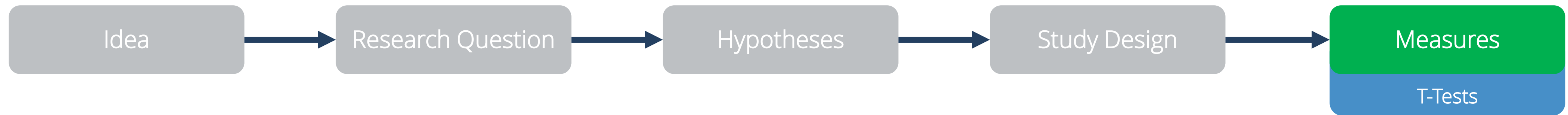
As we are dealing with probabilities, there is the **chance of making mistakes**. Error **accumulates** for multiple tests.

- **Type I error:** A Type I error (also called an  $\alpha$  error or a “false positive”) refers to the mistake of rejecting the null hypothesis when it is true and should not be rejected.  
-> *rejecting the null hypothesis when it is true.*
- **Type II error:** A Type II error (also called a  $\beta$  error or a “false negative”) refers to the mistake of not rejecting the null hypothesis when it is false and should be rejected  
-> *not rejecting the null hypothesis when it is false*

## Example: Court Hearing

*H<sub>0</sub>: The defendant is innocent.*

		Jury Decision	
		Not Guilty	Guilty
Reality	Not guilty	✓	Type I error
	Guilty	Type II error	✓



## T-Tests:

T-Tests are statistical tests used to determine if there is a **significant difference** between the means of two groups and tests whether two sample sets follow the same distribution.

What about a third condition?

**Naive idea:** Use multiple t-tests and compare pair-wise.

→ In the example, three tests (*1-2, 1-3, and 2-3*)

Participant	Condition 1	Condition 2	Condition 3
1	-0,23	-0,43	1,11
2	0,36	-0,48	1,19
3	1,3	2,8	0,97
4	-0,1	0,43	1,3
5	1,2	0,068	1,42
6	-0,097	1,1	1,18
7	-0,64	-0,69	1,4
8	1,8	1,3	1,32
9	0,17	0,66	1,03
10	-2,3	0,87	0,93
<b>Mean</b>	0,1463	0,5628	1,1850
<b>Variance</b>	1,3429	1,0939	0,0304

Problem: Type I errors ( *$\alpha$ -inflation*)

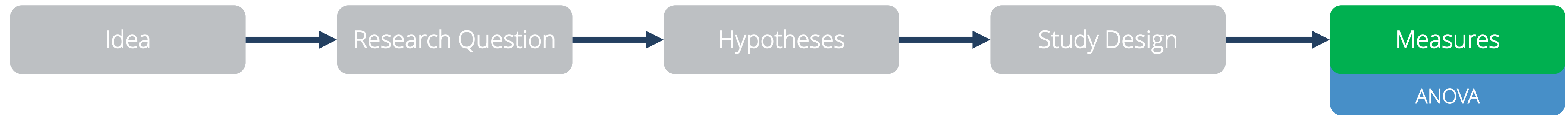
Single t-tests have a **5% error chance** that would accumulate for multiple tests:

*In the example:  $0.95 * 0.95 * 0.95 = 0.86 = 14\%$  chance of error!*

Solution: General Significance test across all sample sets

Pairwise comparisons as a second step (*that also account for errors using correction methods, e.g., Bonferroni*)





## Analysis Of Variance (ANOVA):

ANOVA is a statistical test used to compare the means of three or more groups or conditions.

Different ANOVA variants exist:

- One-way ANOVA
- Repeated-measures ANOVA
- .. etc. (depending on the study design)

Also again, Assumption have to be met: *Continuous data, Homogeneity of variance, Normal Distribution, and also Sphericity.*

Participant	Condition 1	Condition 2	Condition 3
1	-0,23	-0,43	1,11
2	0,36	-0,48	1,19
3	1,3	2,8	0,97
4	-0,1	0,43	1,3
5	1,2	0,068	1,42
6	-0,097	1,1	1,18
7	-0,64	-0,69	1,4
8	1,8	1,3	1,32
9	0,17	0,66	1,03
10	-2,3	0,87	0,93
Mean	0,1463	0,5628	1,1850
Variance	1,3429	1,0939	0,0304

### General Procedure for statistical tests:

1. Check if assumptions are met.
2. Use correct statistical method (in this case ANOVA) to indicate significant differences.
3. If significance is found: Use post-hoc tests for pairwise comparisons.
4. Use error correction methods if possible.

All of this is just a high-level overview and introduction. Many more tests and methods exist. More details also in the Hands-On HCI lecture.



### Limitations for Experimental Research:

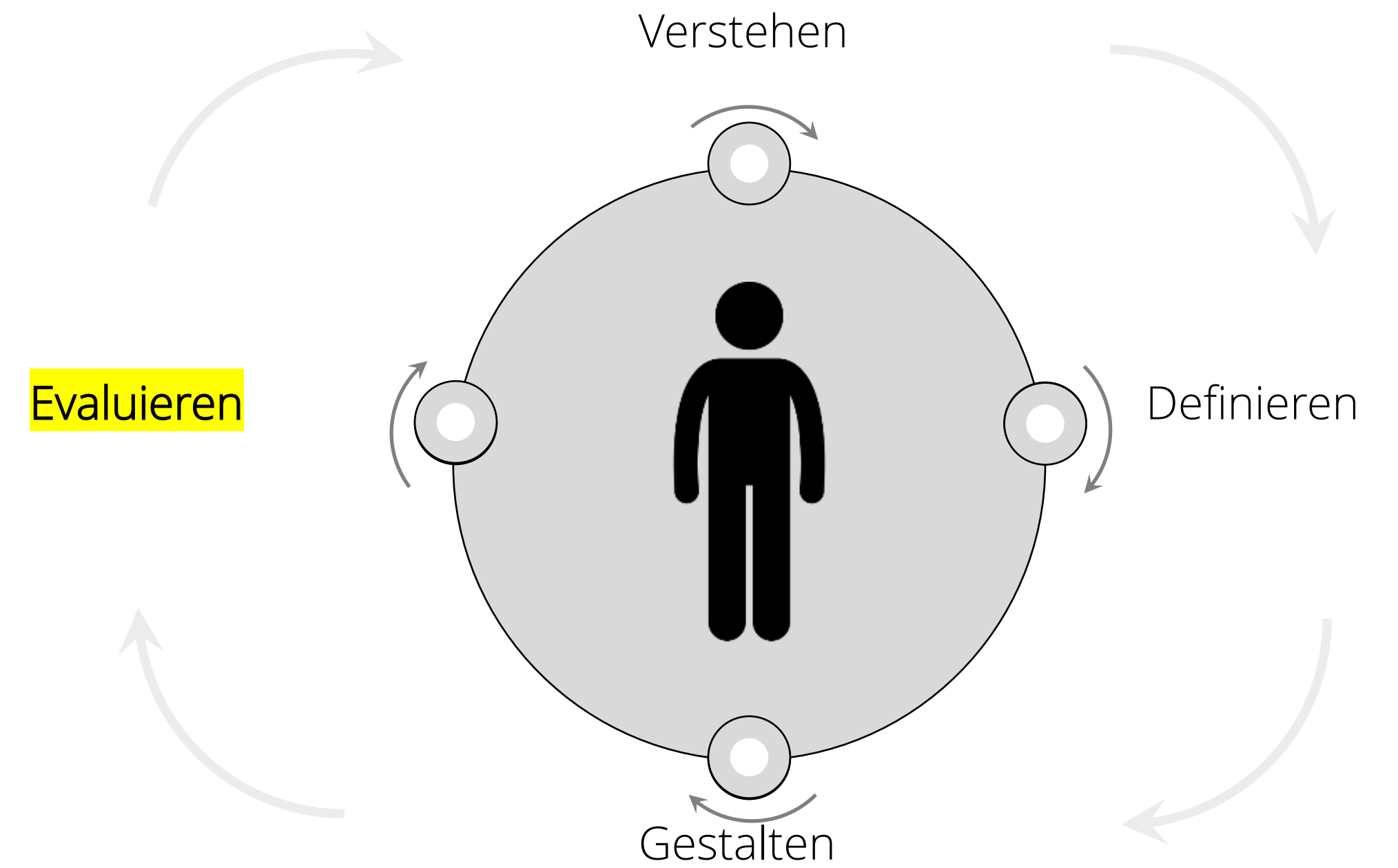
- It requires well-defined, testable hypotheses  
*With limited number of independent (IVs) and dependent variables (DVs)*
- **Strict control of factors** that influence dependent variables (*DV = measures*)  
*E.g., differences (DV) between age groups (IV)*
- Lab-based experiments **may not be a good representation of users' typical interaction behavior**  
*E.g., stress of being observed, different environment*
- No matter how well an experiment is designed and conducted, **bias and errors** can never be completely eliminated
- Needs **careful planning and time for conducting** the experiment





## Summary for Experimental Research:

- Is based on at least one testable research hypothesis and aims to validate it.
- There are *usually* at least two conditions (*a treatment condition and a control condition*) or groups (*a treatment group and a control group*).
- The **dependent variables** are often measured through quantitative measurements.
- The **results are analyzed** through various statistical significance tests.
- A true experiment should be designed and conducted with the goal of **removing potential biases**.
- A true experiment **should be replicable with different participant samples**, at different times, in different locations, and by different experimenters.



# Human-Computer Interaction SoSe 25

## *Evaluation - Part 1*



2. In einem Experiment messen wir, das Vergnügen auf einer Skala von 1-6 (geringes Vergnügen - hohes Vergnügen). Sind dies qualitative oder quantitative Daten und warum? (1 Punkt)  
*We measure in an experiment the rate enjoyment on a scale from 1-6 (low enjoyment - high enjoyment). Is this qualitative or quantitative data and explain why (1 point)?*

3. Unter Berücksichtigung des Fitts'schen Gesetzes, wo auf dem Bildschirm sollten wir einen wichtigen Button (der mehrmals pro Stunde verwendet wird) für eine PC-Anwendung platzieren und warum? (1 Punkt für Platzierung, 2 Punkte für Erklärung) (3 Punkte)  
*Knowing Fitts' Law, where on the screen do we want to position an important button (used multiple times per hour) for a PC application, and why? (1 point for placement, 2 points for explanation) (3 points)*

Name: \_\_\_\_\_

Matrikelnummer: \_\_\_\_\_

4. Wird diese Button Position auch auf Smartphones den selben Vorteil bringen? Warum oder warum nicht? (1 Punkt)  
*Will this position of the button also have the same benefit on smartphones? Why or why not? (1 point)*



Name: \_\_\_\_\_

Matrikelnummer: \_\_\_\_\_

7 Experiment (21 Punkte)

Wir haben ein neues Layout für eine physische Tastatur für den PC entwickelt, bei dem die Anordnung der Tasten basierend auf der Häufigkeit der Buchstaben im alltäglichen Chatten optimiert wurde (wir nennen es ChattBoard). Dieses neue Layout basiert auf Daten, die wir in den letzten 10 Jahren von 10.000 Menschen aus Deutschland gesammelt haben. Unser Werbeteam möchte die Tastatur bewerben und behauptet, dass Benutzer schneller tippen können und weniger Fingerbeschwerden erleben werden als mit alternativen Layouts. Ihre Aufgabe ist es, das Experiment zu entwerfen.

*We designed a new layout for a physical keyboard for a PC in which the arrangement of the keys is optimized based on the frequency of letters used in casual chatting (we call it ChattBoard). This new layout is grounded in data collected over the last 10 years from 10,000 people in Germany. Our advertising team wants to promote the keyboard, claiming that users will be able to type faster and experience less finger pain as with alternative layouts. Your task is to design the experiment.*

1. Nennen Sie unsere unabhängige Variable (IV) (1 Punkt) und schlagen Sie drei Stufen dieser IV vor (mit Erklärung warum), die wir im Experiment verwenden sollten (3 Punkte). (4 Punkte)  
*Name our independent variable (IV) (1 point) and propose (with explanation why) three levels of this IV that we should use in the experiment (3 points). (4 points)*
2. Nennen Sie beide abhängigen Variablen (DV) (1 Punkt) und wie Sie diese messen würden (2 Punkte). (3 Punkte)  
*Name our two dependent variables (1 point) and explain how you would measure them (2 points). (3 points)*

Name: \_\_\_\_\_

Matrikelnummer: \_\_\_\_\_

3. Erklären Sie die Aufgabe (was wird ein Teilnehmer während der Studie tun) (2 Punkte), die Sie im Experiment zur Messung unserer abhängigen Variablen verwenden würden. Achten Sie darauf welche Texte in der Studie getippt werden sollen! (2 Punkte)  
*Explain the task (what will a participant do during the study) (2 points) that you would use in the experiment to measure our dependent variables. Pay attention to that type of texts should be typed in the study. (2 points)*

4. Wir haben zwei Studiendesigns (Between-Subjects, Within-Subjects) im Unterricht gelernt. Erklären Sie jedes Design (2 Punkte) und nennen Sie einen Nachteil für jedes Design (2 Punkte). (4 Punkte)  
*We learned two study designs (Between Subjects, Within Subjects) in class. Explain each (2 points) and name one disadvantage for each (2 points). (4 points)*

5. Welches Design würden Sie in unserem Experiment für ChattBoard verwenden (und warum!)? (2 Punkte)  
*Which design would you use in our experiment for ChattBoard (and why!)? (2 points)*

Name: \_\_\_\_\_

Matrikelnummer: \_\_\_\_\_

6. Formulieren Sie eine Hypothese für jede unserer abhängigen Variablen (2 Punkte).  
*Write down a hypothesis for each of our dependent variables (2 points).*

7. Nennen Sie eine kontrollierte, eine zufällige und eine störende Variable in unserem Experiment mit ChattBoard. (3 Punkte)  
*Name one controlled, one random, and one confounding variable in our experiment with ChattBoard. (3 points)*



✨ a Paid AR Study with a Friend! ✨

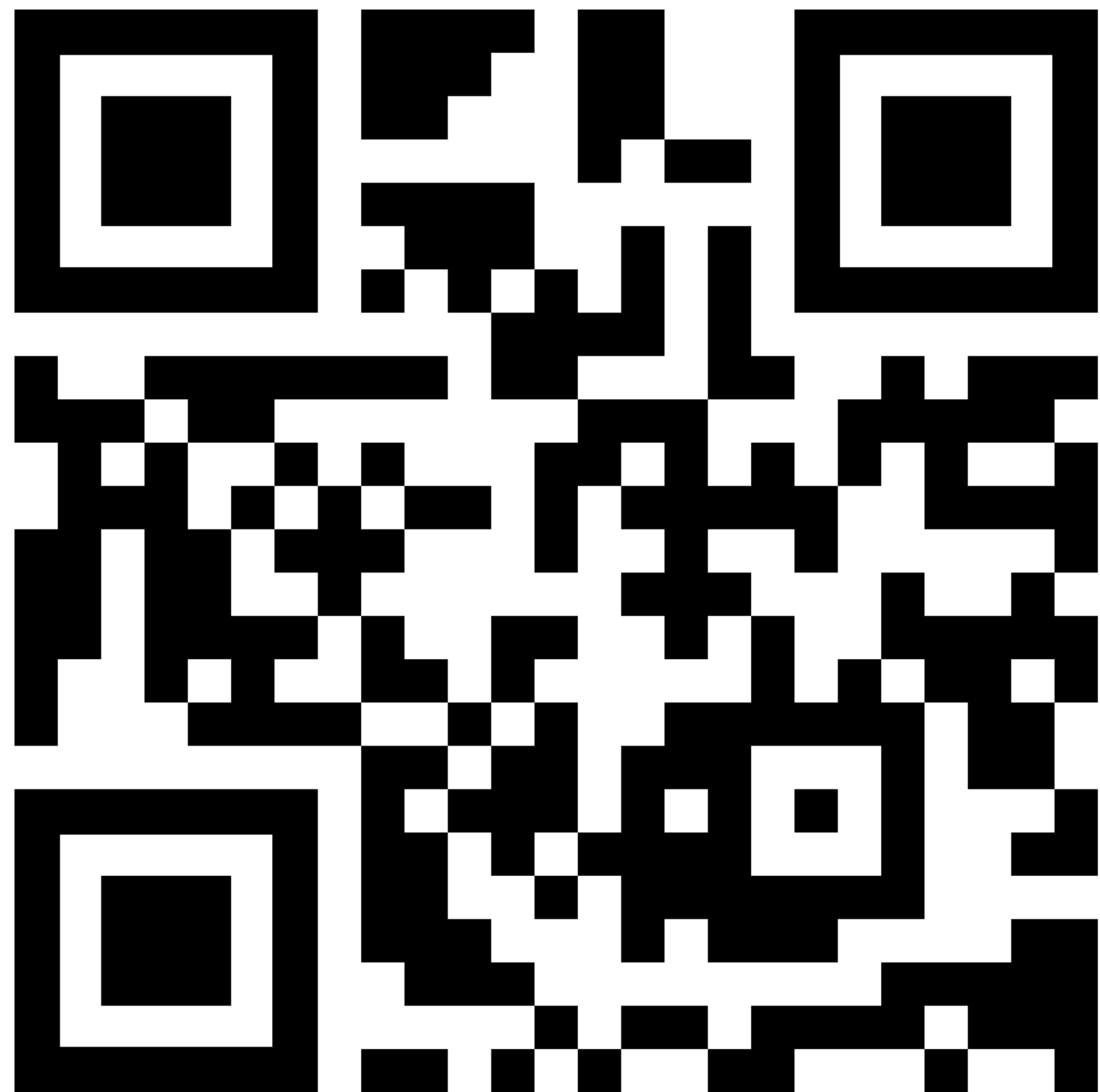
**What if we can share and visualise Memes in F2F conversation in AR**





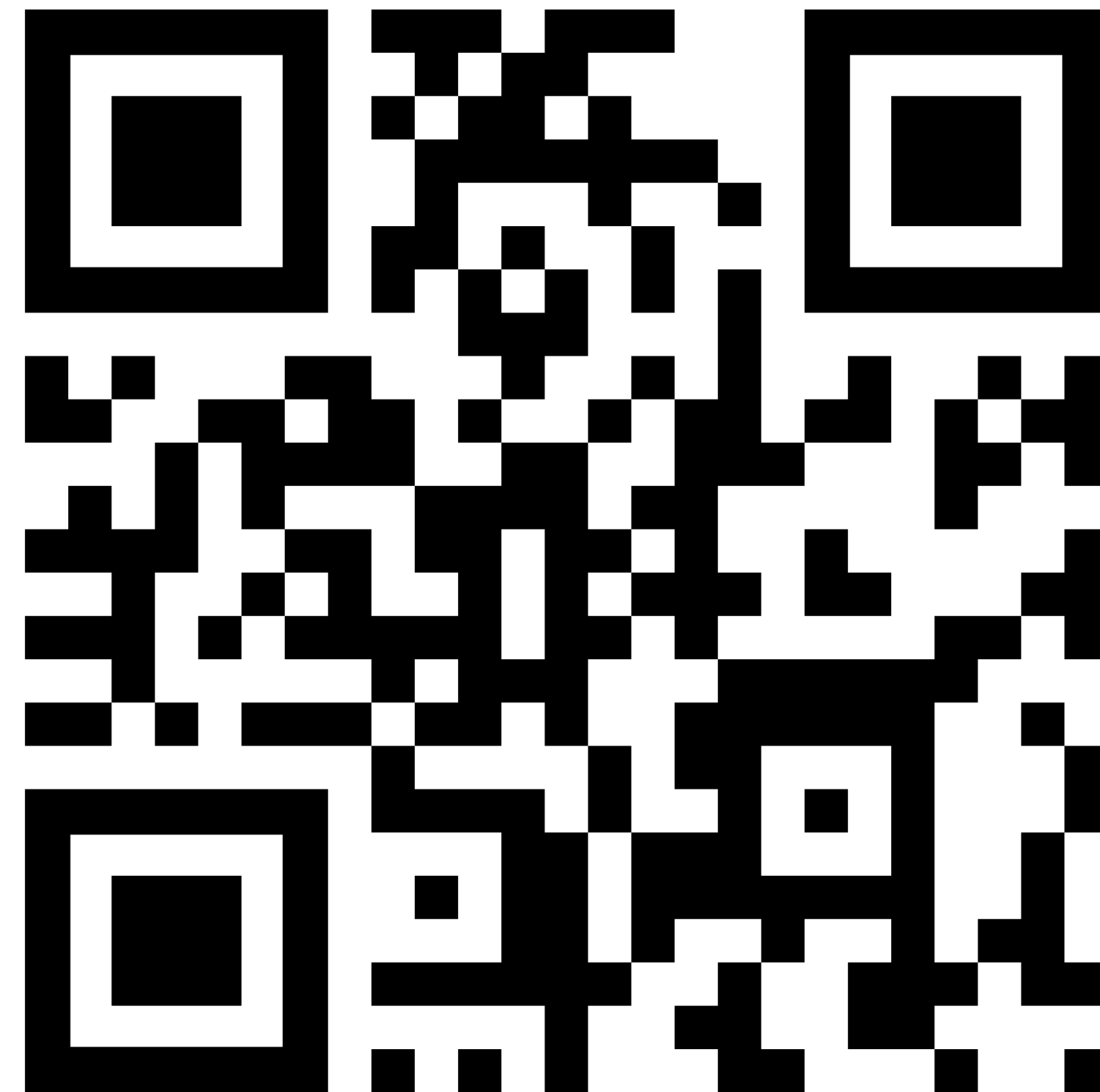






**Website with all info**

<https://shorturl.at/kReSF>



**Scheduling a timeslot**

<https://shorturl.at/Tm4lg>

**Contact: [yanni.mei@tu-darmstadt.de](mailto:yanni.mei@tu-darmstadt.de)**

When: From 21.07.2025  
How long: approx. 1h  
How much: 20 Euro PER PERSON