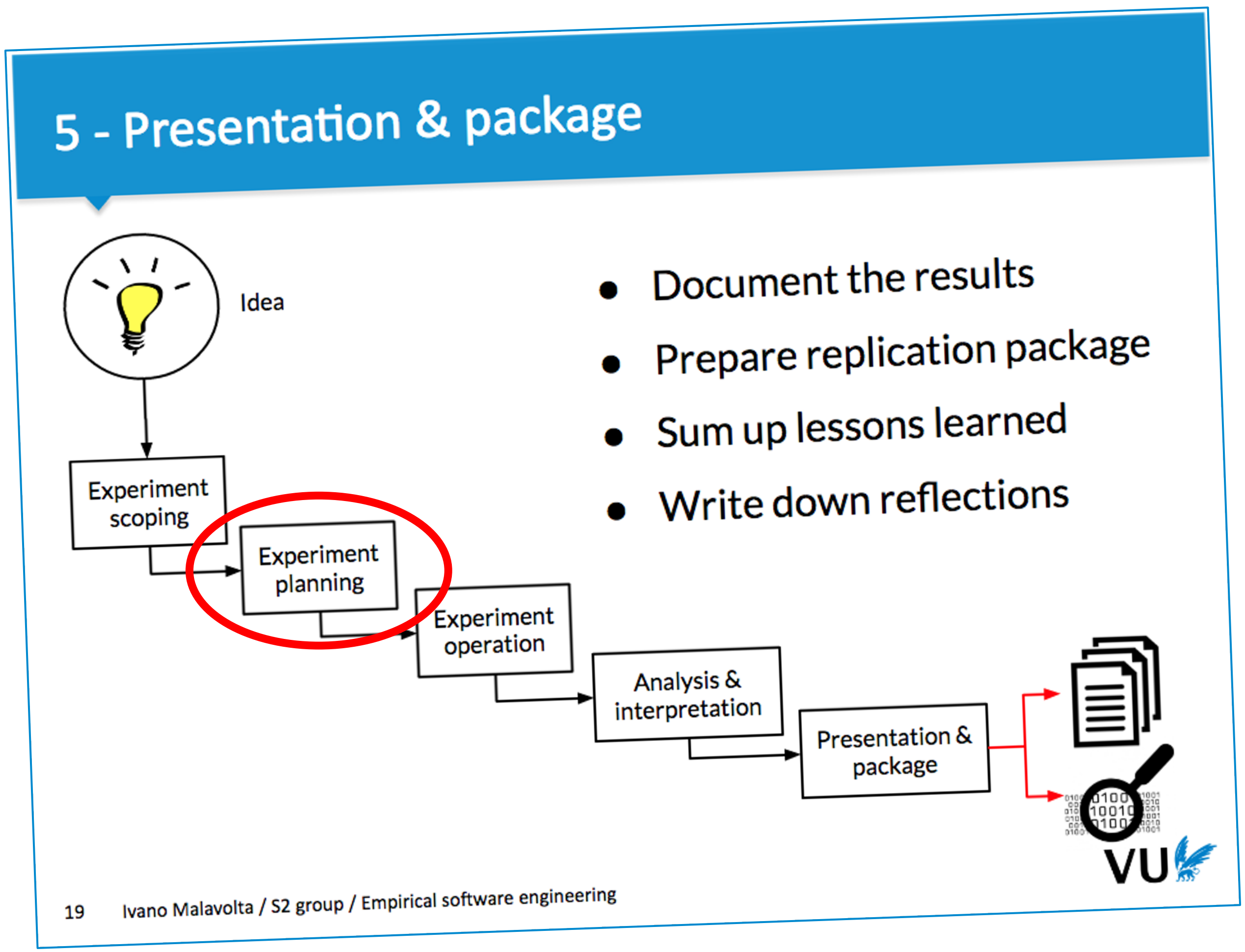
Experiment planning

Ivano Malavolta

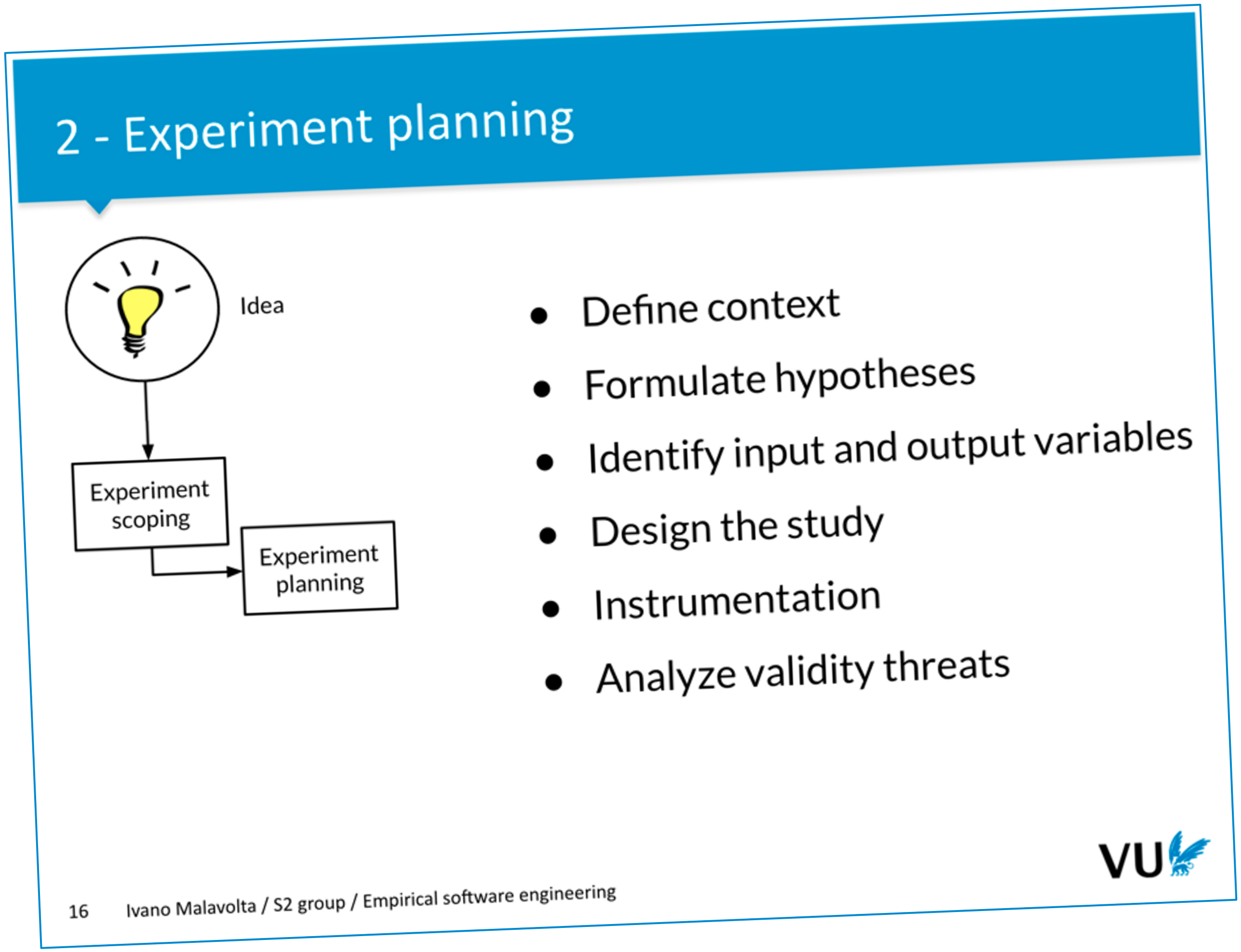
VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Recall


5 - Presentation & package
Idea
Experiment scoping
Experiment planning
Experiment operation
Analysis & interpretation
Presentation & package
Document the results
Prepare replication package
Sum up lessons learned
Write down reflections
19
Ivano Malavolta / S2 group / Empirical software engineering

Recall


2 - Experiment planning
Idea
Experiment scoping
Experiment planning
Define context
Formulate hypotheses
Identify input and output variables
Design the study
Instrumentation
Analyze validity threats
16
Ivano Malavolta / S2 group / Empirical software engineering
VU

Roadmap

Experiment planning

Context selection

Research questions and hypotheses formulation

Variables selection

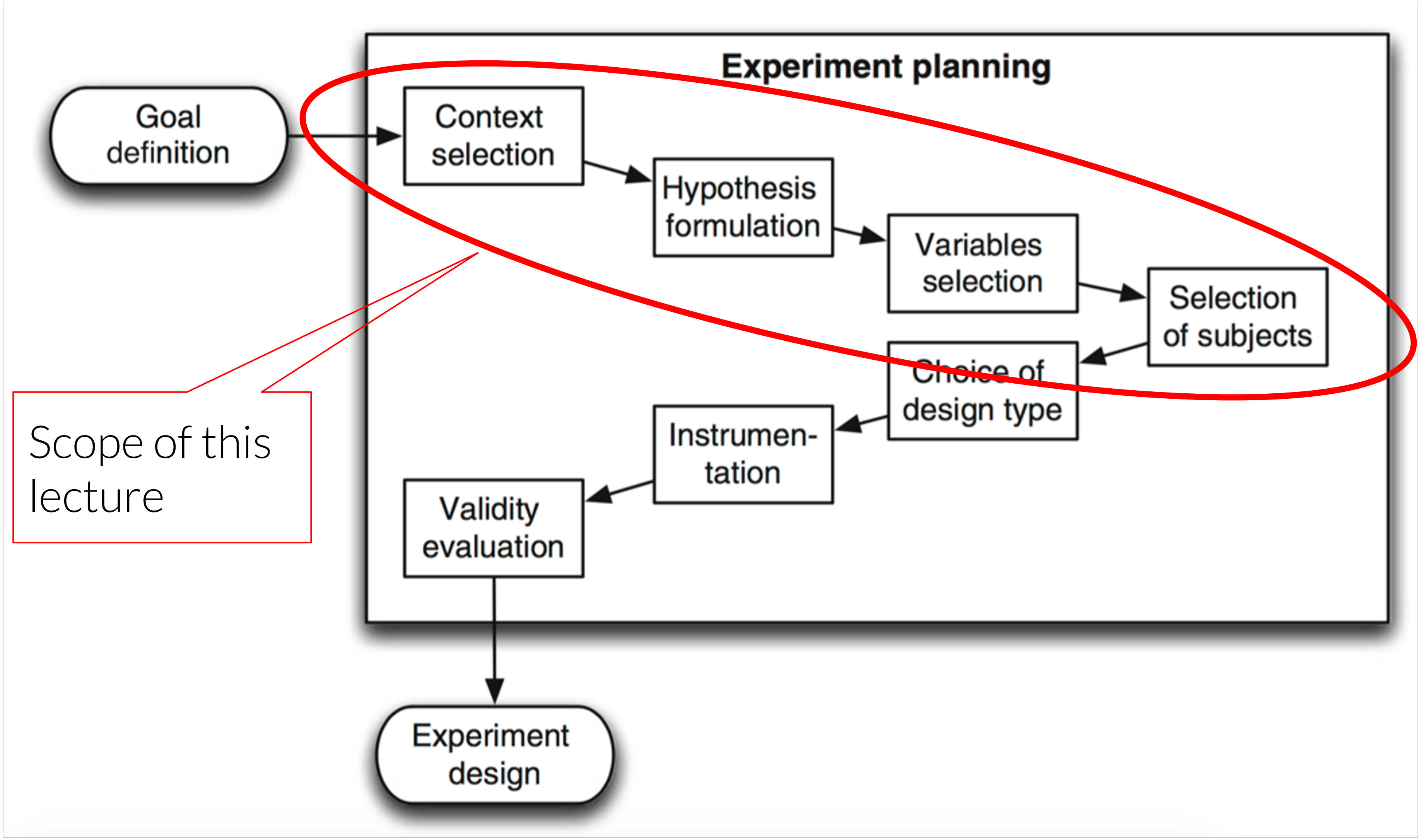
Subjects selection

Scoping VS planning

- Experiment scoping describes **WHY** we run an experiment
  - With a hint about the "how"

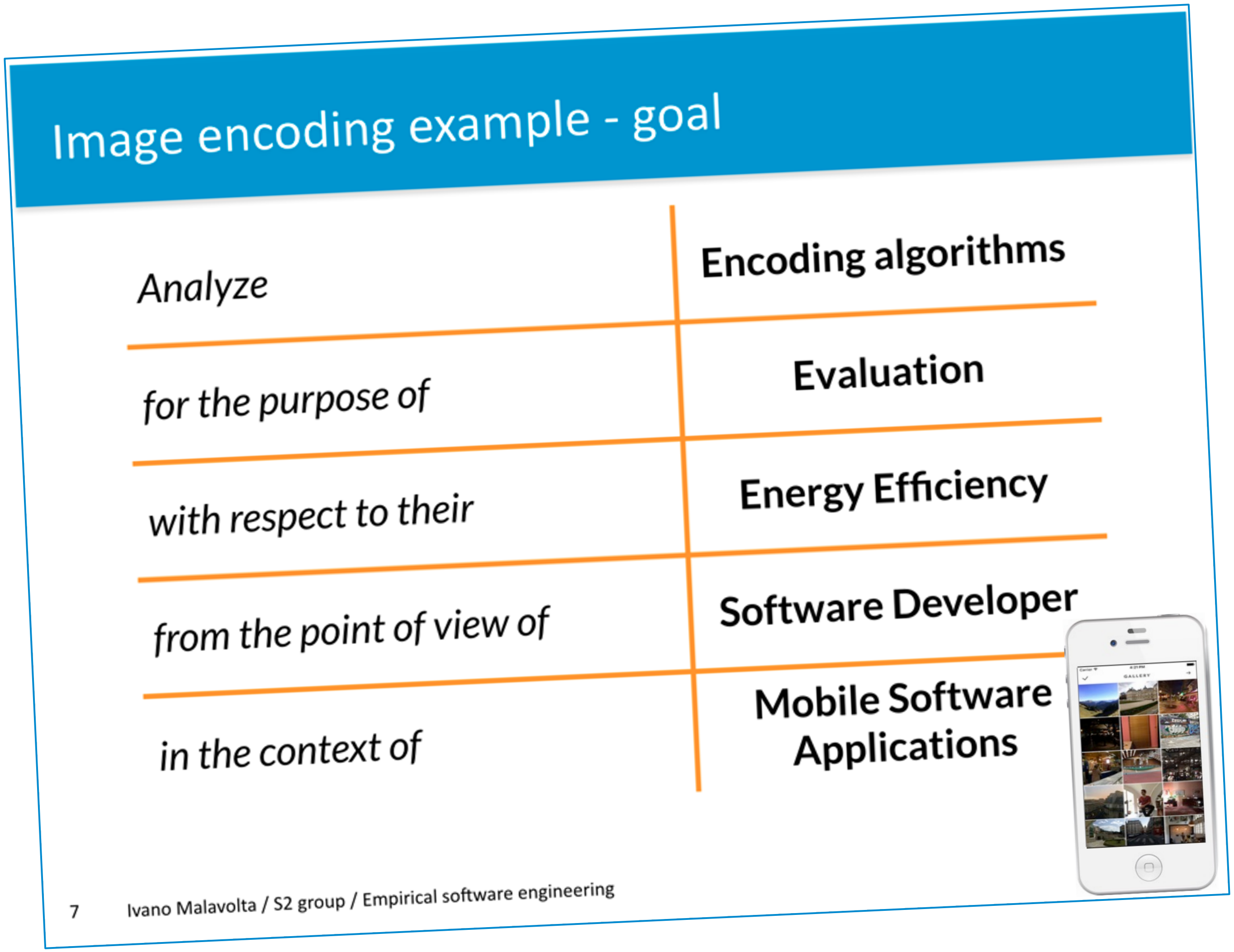
- The planning determines **HOW** the experiment will be executed
  - Be careful here → the result of the experiment can be disturbed (or even destroyed) if not planned properly

Planning phases



Context selection

We already heard about context...



Context selection

CONTEXT: the broad perspective of the experiment

Our **goal** here is to achieve the most general results
→ optimum: large real software projects, with practitioners

Many risks involved….

Context selection dimensions

Lab	Reality
Off-line	On-line
Students	Professionals
Toy objects	Real objects
Specific	General

Quick exercise

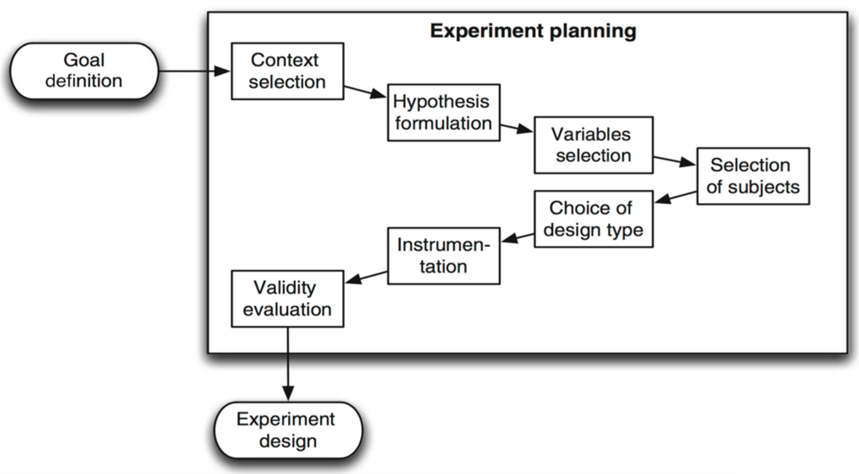
Context selection dimensions

Lab	Reality
Off-line	On-line
Students	Professionals
Toy objects	Real objects
Specific	General

10 Ivano Malavolta / S2 group / Empirical software engineering

Think about the Image encoding case study
and formulate a potential context for an experiment

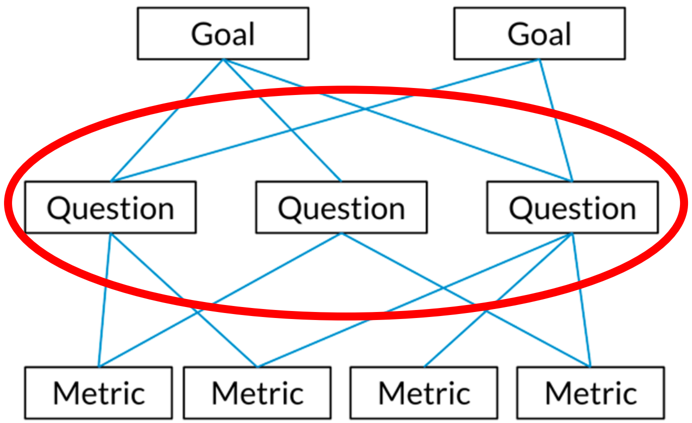
Research questions and hypotheses formulation



Research questions formulation

Research questions detail the specific objectives of the empirical study

Incepted from the study definition



Starting point to identify the variables of interest of your study

Suggestions

Research questions should be as clear as possible
→ they will guide the whole experiment

Avoid "boolean" questions!

Avoid questions you cannot answer

What is the best JavaScript framework in terms of performance?
What is the most productive programming language?
...

Golden words: to what extent..., what is the impact of X to Y, what are the traits of Xs, what are the characteristics of T...

Remind that in your report you will come back to them and explicitly answer each of them in details

From research questions to hypotheses

- When a research question is going to be addressed by applying a statistical test, it is necessary to formulate an hypothesis
- Very useful to select what kind of statistical procedure you need to use

Not needed in all cases

Hypotheses formulation

- Conjecture (P)
  - Administration of treatment has influence on some features
- Consequence (Q)
  - We observe a significant difference in terms of some features

$$P \rightarrow Q$$

Hypotheses formulation

Hypothesis: a **formal** statement about a phenomenon

- **Null** hypothesis H_0: no real trends or patterns in the experiment setting (aka ~Q)

- **Alternative** hypothesis H_a: there are real trends or patterns in the experiment setting (aka Q)

There must be at least a pair of null and alternative hypotheses **for each research question** in your GQM

Falsification (modus tollens)

- We aim at rejecting the absence of trends (~Q is false)...
  - we test the null hypothesis H_0

If we can reject the null hypothesis ~Q → **we can draw conclusions**

This comes from Popper (1959): any statement in a scientific field is true until anybody can contradict it

This is like a "guilty" verdict in a criminal trial: the evidence is sufficient to reject innocence

- Aiming at verifying Q is <u>WRONG</u>
  - "Affirming the consequent"
  - Provides no insight on the conjecture

Example

- Question:
  - What is the impact of image encoding algorithms on the energy *efficiency of mobile apps*?
- Conjecture (P):
  - using different algorithms leads to different energy consumptions
- Consequence (Q):
  - (when applying different algorithms) we observe a different energy consumption

Example

- **Null** hypothesis (¬Q): there is **no change** in terms of energy consumption

$$H_0: mean(E_{png}) = mean(E_{jpg})$$

- **Alternative** hypothesis (Q): the energy consumption when using PNG images is different then the one consumed when using JPG images

$$H_a: mean(E_{png}) \mathrel{!=} mean(E_{jpg})$$

What can happen now?

H_0: mean(E_{png}) = mean(E_{jpg})

- We <u>reject</u> the null hypothesis (~Q = false)
  - Q = true (with a certain probability)
  - our conjecture P has been corroborated → we are confident that different algorithms impact energy consumption (P)

Modus tollens:

P → Q

~Q

~P

- We <u>fail to reject</u> the null hypothesis (~Q = true)
  - ~P = true
  - our conjecture P has been falsified → no conclusions can be made

Example

Native vs Web Apps: Comparing the Energy Consumption and Performance of Android Apps and their Web Counterparts

Ruben Horn, Abdellah Lahnaoui, Edgardo Reinoso, Sicheng Peng, Vadim Isakov, Tanjina Islam, Ivano Malavolta
Vrije Universiteit Amsterdam, The Netherlands
{r.horn | a.lahnaoui | e.j.reinosocampos | s3.peng | v2.isakov}@student.vu.nl, {t.islam | i.malavolta}@vu.nl

***Abstract—Context*. Many Internet content platforms, such as Spotify and YouTube, provide their services via both native and Web apps. Even though those apps provide similar features to the end user, using their native version or Web counterpart might lead to different levels of energy consumption and performance. *Goal*. The goal of this study is to empirically assess the energy consumption and performance of native and Web apps in the context of Internet content platforms on Android. *Method*. We select 10 Internet content platforms across 5 categories. Then, we measure them based on the energy consumption, network traffic volume, CPU load, memory load, and frame time of their native and Web versions; then, we statistically analyze the collected measures and report our results. *Results*. We confirm that native apps consume significantly less energy than their Web counterparts, with large effect size. Web apps use more CPU and memory, with statistically significant difference and large effect size. Therefore, we conclude that native apps tend to require fewer hardware resources than their corresponding Web versions. The network traffic volume exhibits statistically significant difference in favour of native apps, with small effect size. Our results do not allow us to draw any**

Even though those apps provide similar features to the end user, using their native version or Web counterpart might lead to different levels of energy consumption and performance.

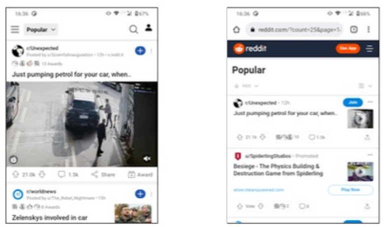


Fig. 1: Reddit native Android (left) vs Web app (right)

TABLE III: Dependent variables

Variable	Description	RQ
Energy consumption (e)	Energy consumption is measured in Joules (J) as the energy consumed by the mobile device during the experiment run	RQ1
Network traffic (n)	Amount of data in Bytes (B) sent and received by the mobile device during the experiment run	RQ2
CPU load (c)	Mean relative (%) device CPU utilization across all cores	RQ2
Memory load (m)	Mean (kB) device memory utilization	RQ2
Frame time (f)	Median time in nanoseconds (ns) between two successive frames (We use median as an aggregation measure, since we expect extreme outliers due to apps blocking on the main thread during certain operations)	RQ2

RQ1 *How does energy consumption vary between native and Web versions of the same app?*

$$H_0 : \mu_{e_{\text{native}}} = \mu_{e_{\text{Web}}}$$
$$H_a : \mu_{e_{\text{native}}} \neq \mu_{e_{\text{Web}}}$$

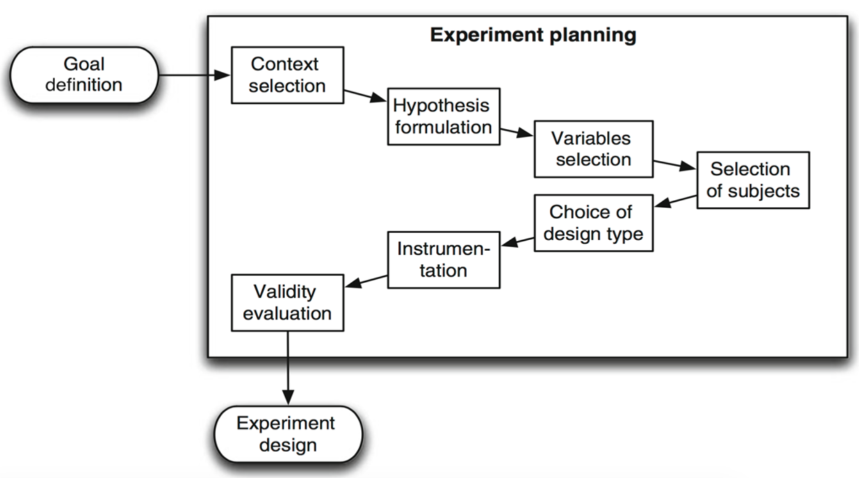
RQ2 *How does performance vary between native and Web versions of the same app?*

$$H_0 : \mu_{d_{\text{native}}} = \mu_{d_{\text{Web}}} \quad \forall d \in \{n, c, m, f\}$$
$$H_a : \mu_{d_{\text{native}}} \neq \mu_{d_{\text{Web}}} \quad \exists d \in \{n, c, m, f\}$$

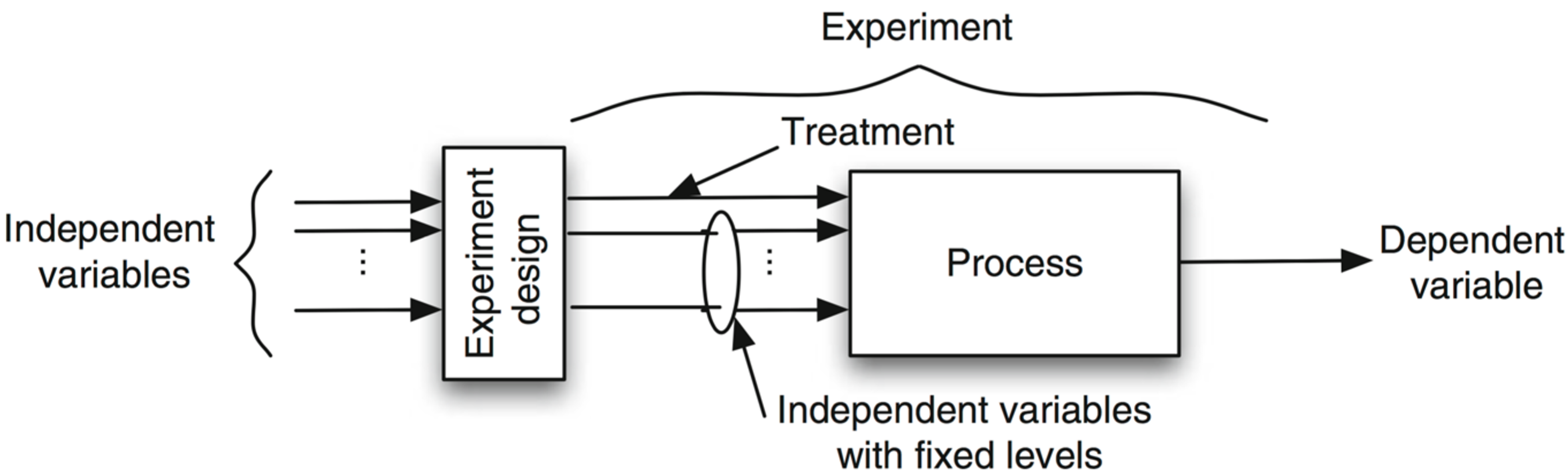
Other examples of hypotheses

Null Hypothesis	Alternative Hypothesis	Experiment
H_0: there is no difference in defect detection rates of teams applying the PBR inspection technique as compared to teams applying the usual technique	H_1: the defect detection rates of teams applying PBR are higher compared to teams using the usual technique	(Basili, 1996)
H_0: classes declared as friends of other classes have the same inheritance as other system classes	H_1: classes declared as friends of other classes have less inheritance than other system classes	(Counsell, 1999)
H_0: there is no difference between the different inspection techniques with respect to the team scores on defect detection rate	H_1: there is a difference between the various techniques with respect to the team scores on defect detection rate	(Fusaro, 1997)
H_0: there is no difference in intervals neither in number of defects detected between inspections with large teams and with smaller teams	H_1: inspections with large teams have longer intervals, but find no more defects than smaller teams	(Porter, 1997)
H_0: there is no difference in effectiveness in teams who begin an implementation using an existing example and in teams who begin implementing from scratch	H_1: teams who begin an implementation using an existing example for guidance are more effective than those who begin implementing from scratch are	(Shull, 2000)

Variables selection



Recap



Terminology

- Dependent variables: quantities observed in the study (a.k.a. *response*, *output* variables)
  - e.g. energy consumption, image quality
- Independent variables: quantities that we are able to manipulate/control (a.k.a. *input* variables)
  - e.g. encoding algorithm, size of image, operating system

Variables selection

- The choice of independent and dependent variables is usually done **in parallel**
- Some variables cannot be measured directly (e.g. productivity, code quality, effort...)
  - We use **proxies** to estimate them
    - proxies may introduce a *construct validity threat:* is what we are measuring a good representation of our variable?

Variables selection

- Independent variables should have some effect on the dependent ones

  → do not choose variables randomly, think about your RQs

- After choosing the variables you have to define their types, scales, ranges → this is part of measurement theory

Hypotheses formulation

- There is always only 1 dependent variable
  - e.g., power consumption
- ... and 1 independent variable (the main factor)

  - often one level for the **control group**

    e.g. use of traditional technique

  - one or more levels for **experimental groups**

    e.g. use of new technique(s) tool(s)

Other independent variables are the co-factors

Co-factors

Our main factor is not the only variable influencing the dependent variable(s)

- e.g., network instability of your experimental environment, usage patterns of the analysed websites, skills of subjects, experience of developers, ...

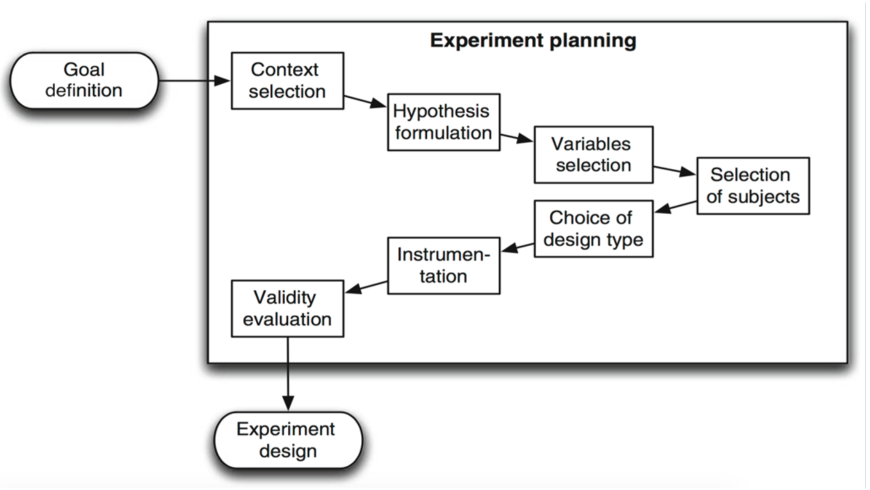
We will never account for all possible co-factors

Your best friend here is randomization

In a good experiment:

- You **limit** the effect of co-factors through a good experimental design
- You are able to **separate** the effect of co-factors from main factors
- You analyze the **interaction** of co-factors with main factor

Subjects selection



Subjects selection

- **Population:** the complete set of items of interest for our experiment
  - e.g. open-source software applications
  - e.g., all existing progressive web apps
- **Sample:** representative selection of individuals for that population
  - e.g. Apache, MySQL
  - e.g. progressive web apps mined from the Tranco list

Sampling techniques

- **Probability Sampling:** the probability of selecting each subject in the population is *known*
    - *Simple Random Sampling:* random selection from the population, probability is 1/total
    - *Stratified random sampling*: the population is divided into groups with a known distribution between the groups. Random sampling is then applied within each group
- **Non-probability sampling:** the probability of selecting each subject out of the population is *unknown*
    - *Convenience:* the most convenient (cost/distance/ complexity) subjects are selected *[usually it is the only way to go]*
    - *Quota:* you select (usually by convenience) samples from groups of subjects (e.g. male vs females, open-source vs closed source)

How big should be a sample?

- **Sample size:** the larger, the better (more general results)
- If the population has a high *variation*, a **larger** sample size is needed
- **Data analysis** may influence sample size
    - some statistical tests have meaning only on large samples

Example of subjects selection

Selected Subjects and Usage Scenarios

Category	Subject	Usage scenario (looped)
News	• ESPN	Open news article, scroll down, continue with next article
	▪ The Weather Channel	Check hourly forecast, check 10-day forecast, check radar
Social media	• LinkedIn	Scroll personal feed, scroll jobs
	▪ Pinterest	Scroll posts, open post, go back
E-Commerce	• Coupang	Open category, scroll products, open product page, check comments
	▪ Shopee	Open category, scroll products, open product page, check comments
Audio streaming	• SoundCloud	Listen to promoted song
	▪ Spotify	Search for a playlist, listen to playlist
Video streaming	• Twitch	Search for channel, watch channel
	▪ YouTube	Search for video, watch video

Read the details in Section 3a in the MobileSoft 2023 paper on Canvas

What this lecture means to you?

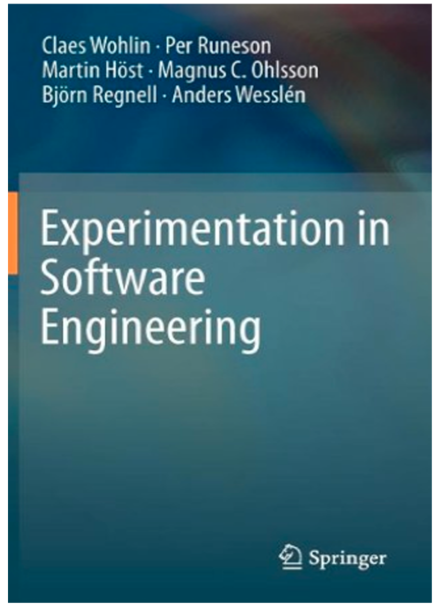
You know how to:

- define the context of your experiment
- define research questions and hypotheses
- define independent and dependent variables
- strategies for selecting subjects

Next step

Measurement theory → how to define the "type" of variables

Readings



Chapter 8

+ All papers in the “Articles on performed experiments” folder in Canvas

(only the part related to subjects selection and variables definition)

+ Example of assignment in Canvas

Acknowledgements

Some contents of this part of lecture extracted from:

- Giuseppe Procaccianti's lectures at VU
- Massimiliano Di Penta's lectures at GSSI (Italy)