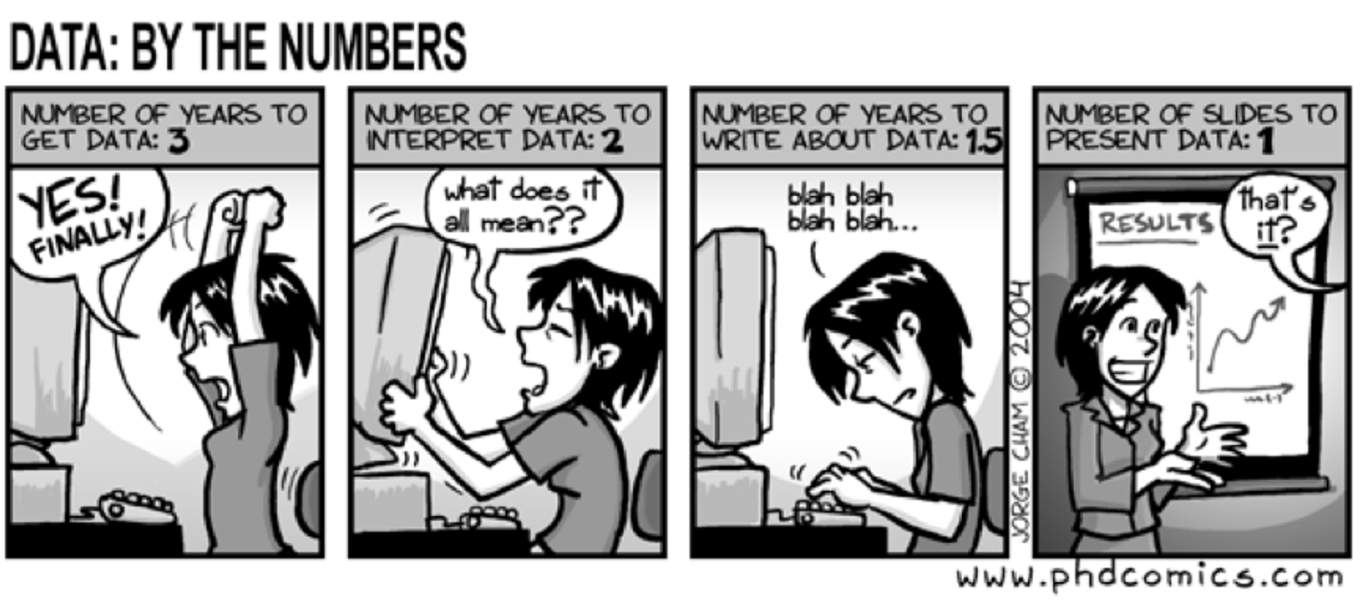
Experiment reporting

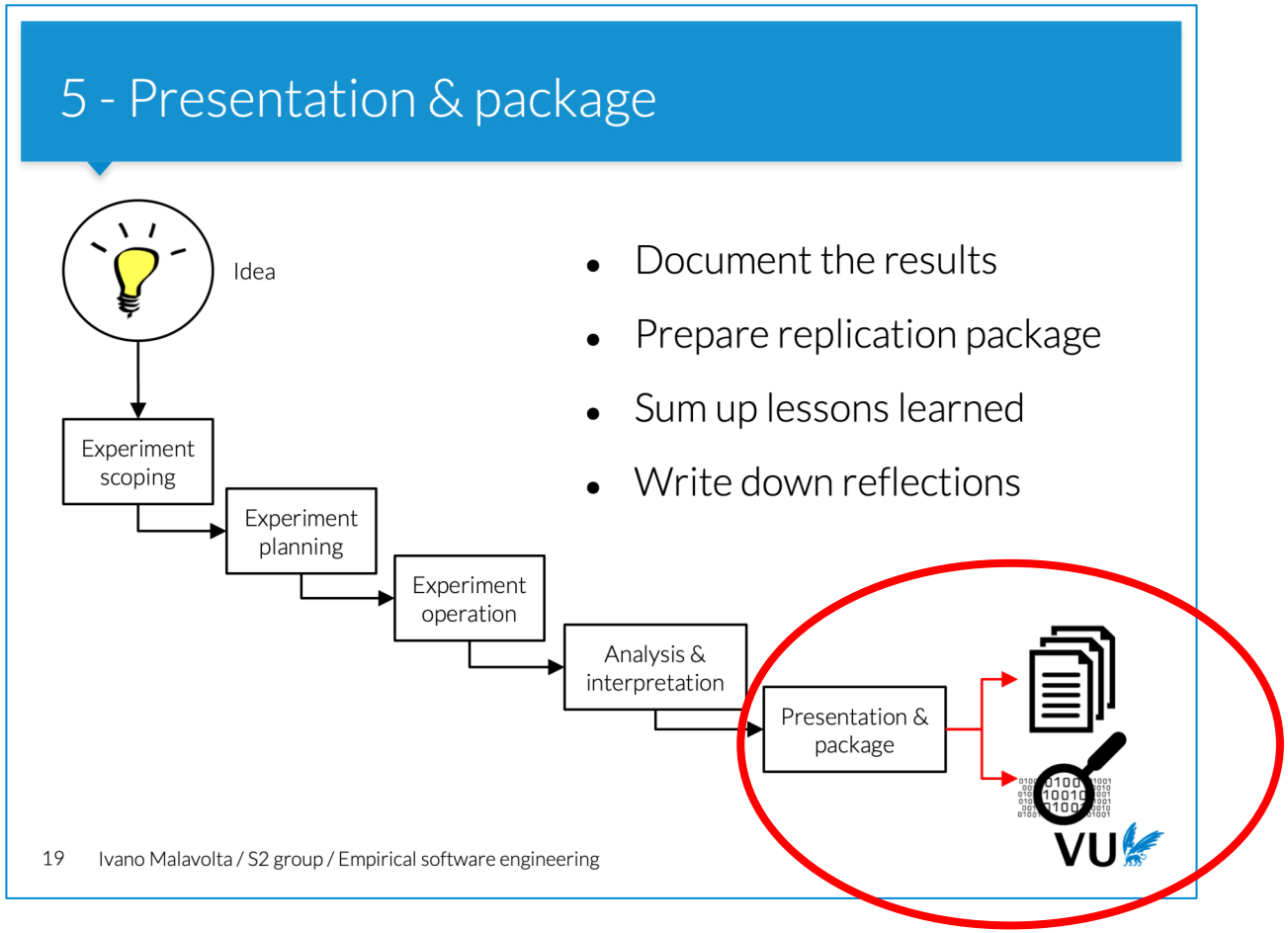
Ivano Malavolta



VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Let's take a step back



VU

Basic structure of an experiment report

(Structured) Abstract
Introduction/Motivation
(Related Work)
Experiment design
Experiment execution
Data analysis
Threats to validity
Discussion
Conclusions
References

Abstract

- People judge papers by their abstracts
  - usually decide whether to read the whole paper based on abstract
- It's important for the abstract to tell the whole story
  - background and context
  - goals
  - method
  - results
  - conclusion
- Often it is the only part freely accessible

The structured abstract

- Context
  - Brief explanation of the motivation for conducting the study
- Objective
  - Aim, objects, focus, perspective of the study (based on the GQM)
- Method
  - Experimental design, number and kind of objects/subjects, selection criteria, data collection and analysis procedures
- Results
  - Main findings
- Conclusion
  - Impact of the obtained results

Example of structured abstract

Empirical Evaluation of Two Best Practices for Energy-Efficient Software Development

Giuseppe Procaccianti*, Hector Fernandez, Patricia Lago
VU University Amsterdam, De Boelelaan 1081a, 1081 HV, Amsterdam, The Netherlands.

Abstract

Background. Energy efficiency is an increasingly important property of software. A large number of empirical studies have been conducted on the topic. However, current state-of-the-Art does not provide empirically-validated guidelines for developing energy-efficient software.
Aim. This study aims at assessing the impact, in terms of energy savings, of best practices for achieving software energy efficiency, elicited from previous work. By doing so, it identifies which resources are affected by the practices and the possible trade-offs with energy consumption.
Method. We performed an empirical experiment in a controlled environment, where we applied two different Green Software practices to two software applications, namely query optimization in MySQL Server and usage of "sleep" instruction in the Apache web server. We then performed a comparison of the energy consumption at system-level and at resource-level, before and after applying the practice.
Results. Our results show that both practices are effective in improving software energy efficiency, reducing consumption up to 25%. We observe that after applying the practices, resource usage is more energy-proportional i.e. increasing CPU usage increases energy consumption in an almost linear way. We also provide our reflections on empirical experimentation in software energy efficiency.
Conclusions. Our contribution shows that significant improvements in software energy efficiency can be gained by applying best practices during design

VU

Introduction and motivation

- Introduction
  - Mini-version of the whole paper
    - Attacked problem
    - Proposed solution or experiment
    - Main results
    - Main contributions (how to “use the paper”)
    - Structure of the paper

- Motivation:
  - **Scope** of the work
  - **Problem** statement
  - Research **objectives**
  - *encourages to read the paper*

Related work

- Why this paper exists?
- Complete picture about:
  - experiments with similar **goals**
  - experiments with similar **objects/subjects**
  - papers with similar goals, but **not empirical** → they lack in provided evidence
- Two main strategies:
  - Paper-by-paper comparison
  - Catalogue of related papers + overall comparison

Experiment design and execution

- **Design**: rewalk through all the experiment plan
  - GQM
  - hypotheses
  - objects and subjects
  - variables
  - measurements
  - experiment design
  - data analysis strategies (e.g., used statistical tests)
- **Execution**: how the experiment plan has been put in practice
  - preparation (eg code instrumentation)
  - data collection procedure
  - data clean up
  - any deviation from the plan
  - *reference to replication package*

VU

Analysis and threats to validity

- **Analysis** (or Results)
  - Demographics
  - Descriptive statistics and data exploration
  - For each research question:
    - hypothesis testing
    - discussion of effect size
    - quick elaboration on obtained results

- **Threats to validity**
  - External
  - Internal
  - Construct
  - Conclusion

Discussion and Conclusions

- Discussion
  - Elaborate on the obtained results
  - Impact into practice (industry)
  - Make the results actionable
    - e.g., how will app developer John use your results tomorrow?
    - e.g., what researcher Y learned about phenomenon X?
  - Lessons learnt

- Conclusions
  - Summary of the main contributions of the paper
  - Future work
    - other experiments on different aspects
    - better ways to perform the experiment
    - ...

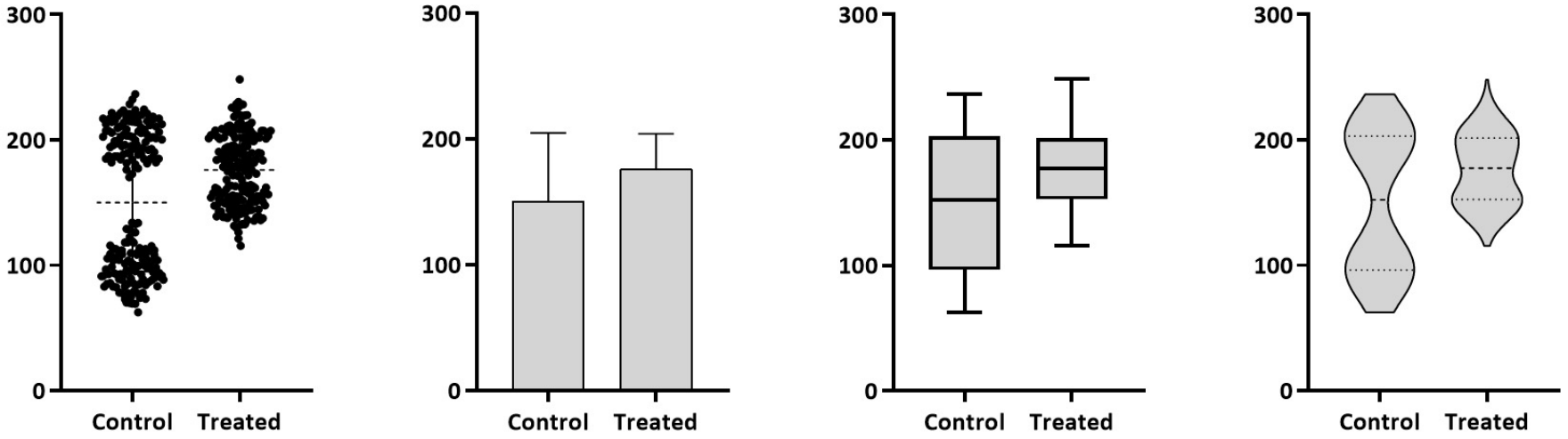
VU

Suggestion about data visualization

- No fixed rule about which plots to use
- First think about the main message you want to give with your plot, then decide which type of plot better fits your message

  e.g., comparison vs absolute values, demographics, distributions
- Get inspiration from the experiments in the *papers_experiments* folder in Canvas

Is the boxplot my silver bullet for plotting?



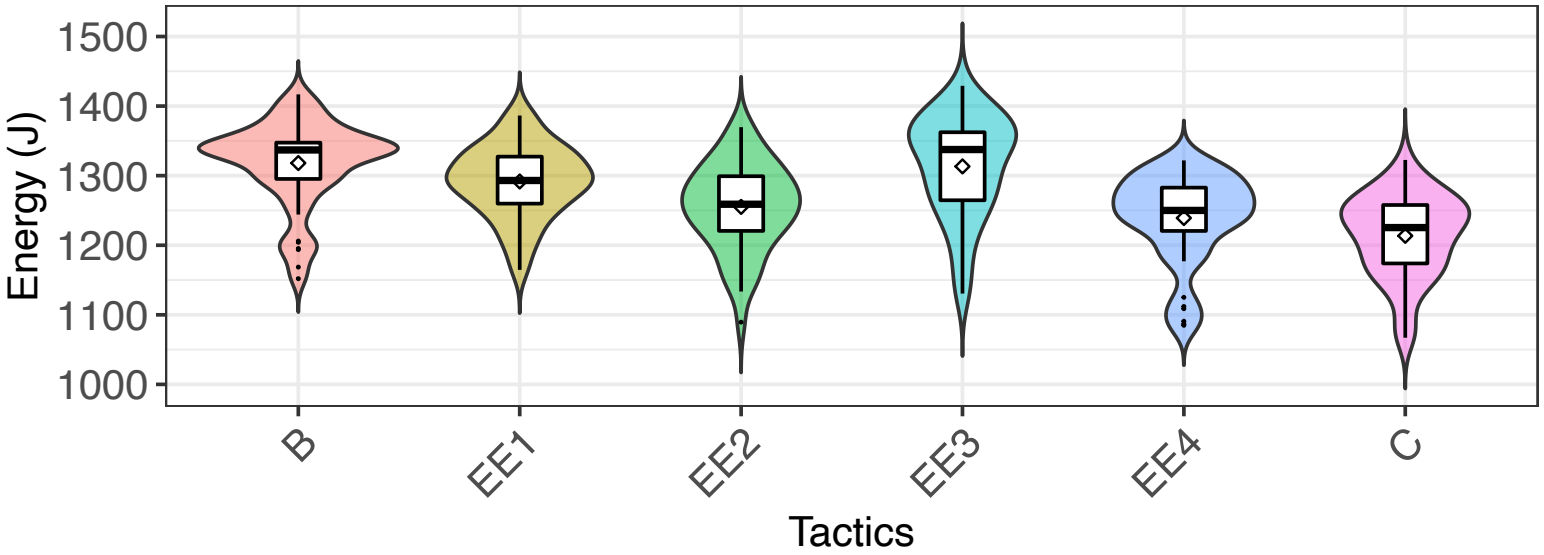
Boxplots are generally good for comparing (1) the **central tendency** and (2) **spread** of your data

What you lose with boxplots is a precise visualization of the **distribution** of your data

For example, if your data is bimodal, you do not see it in a boxplot

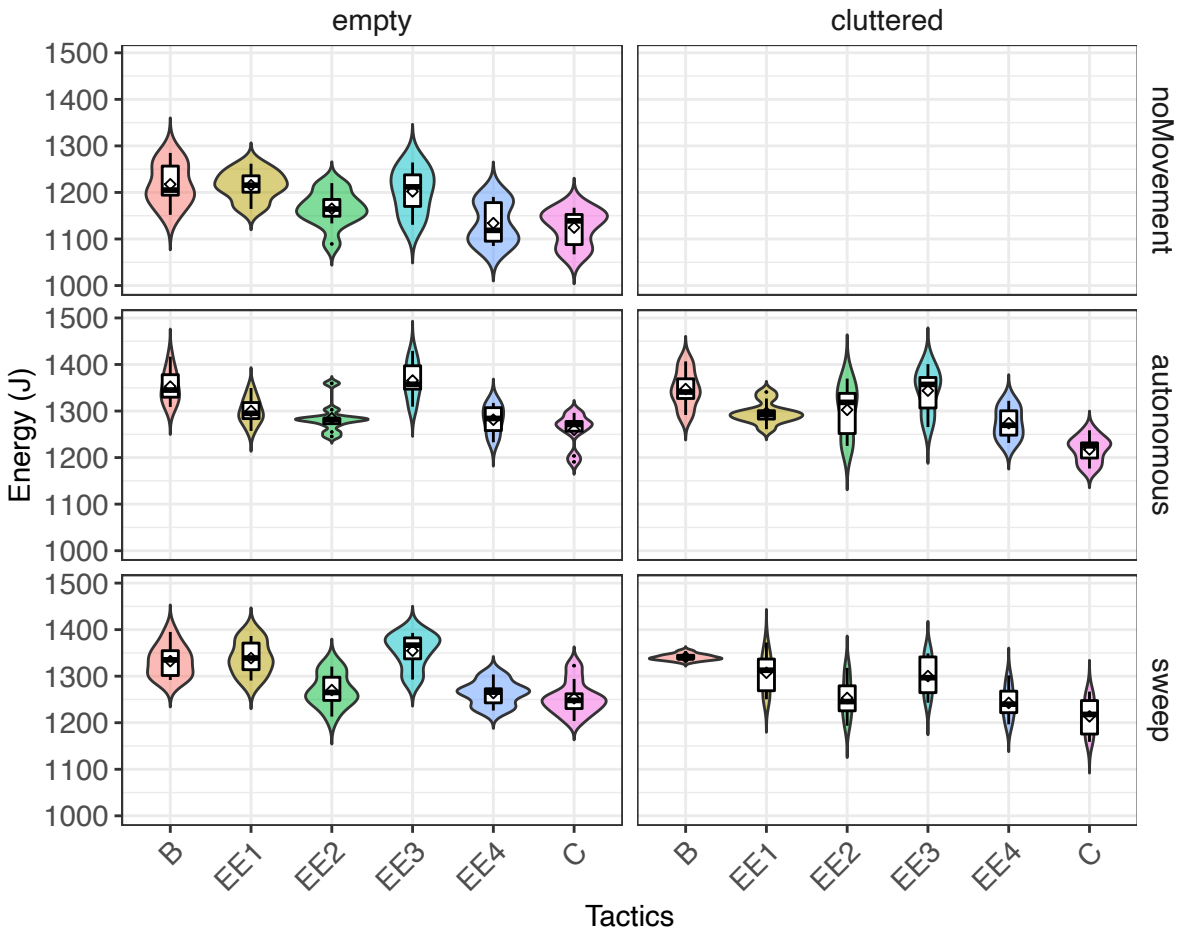
https://www.graphpad.com/guides/prism/8/user-guide/graphing-box-and-whisker-and-v.htm

Examples of visualizations



MSR_2021

Examples of visualizations



MSR_2021

Examples of visualizations

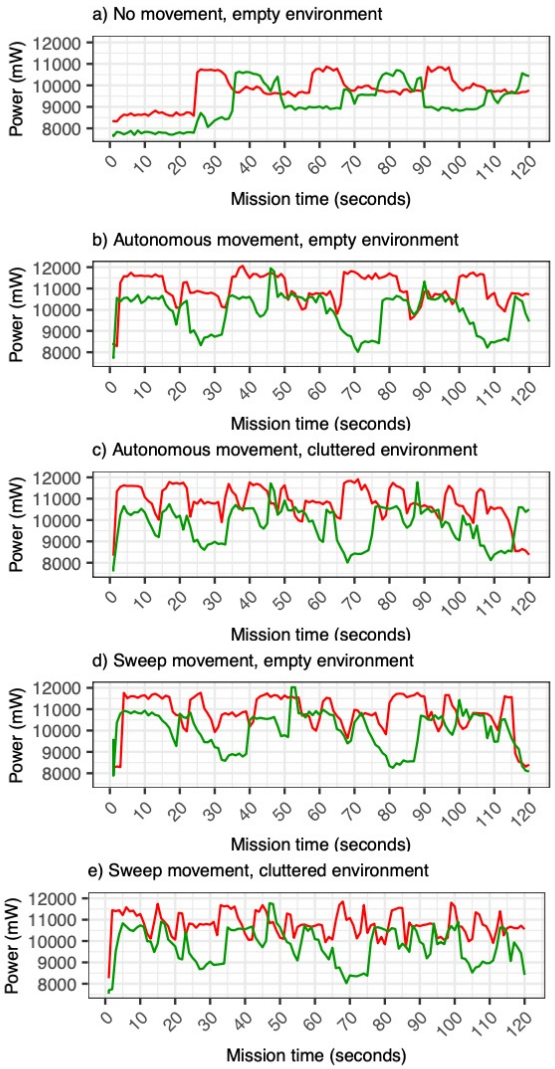
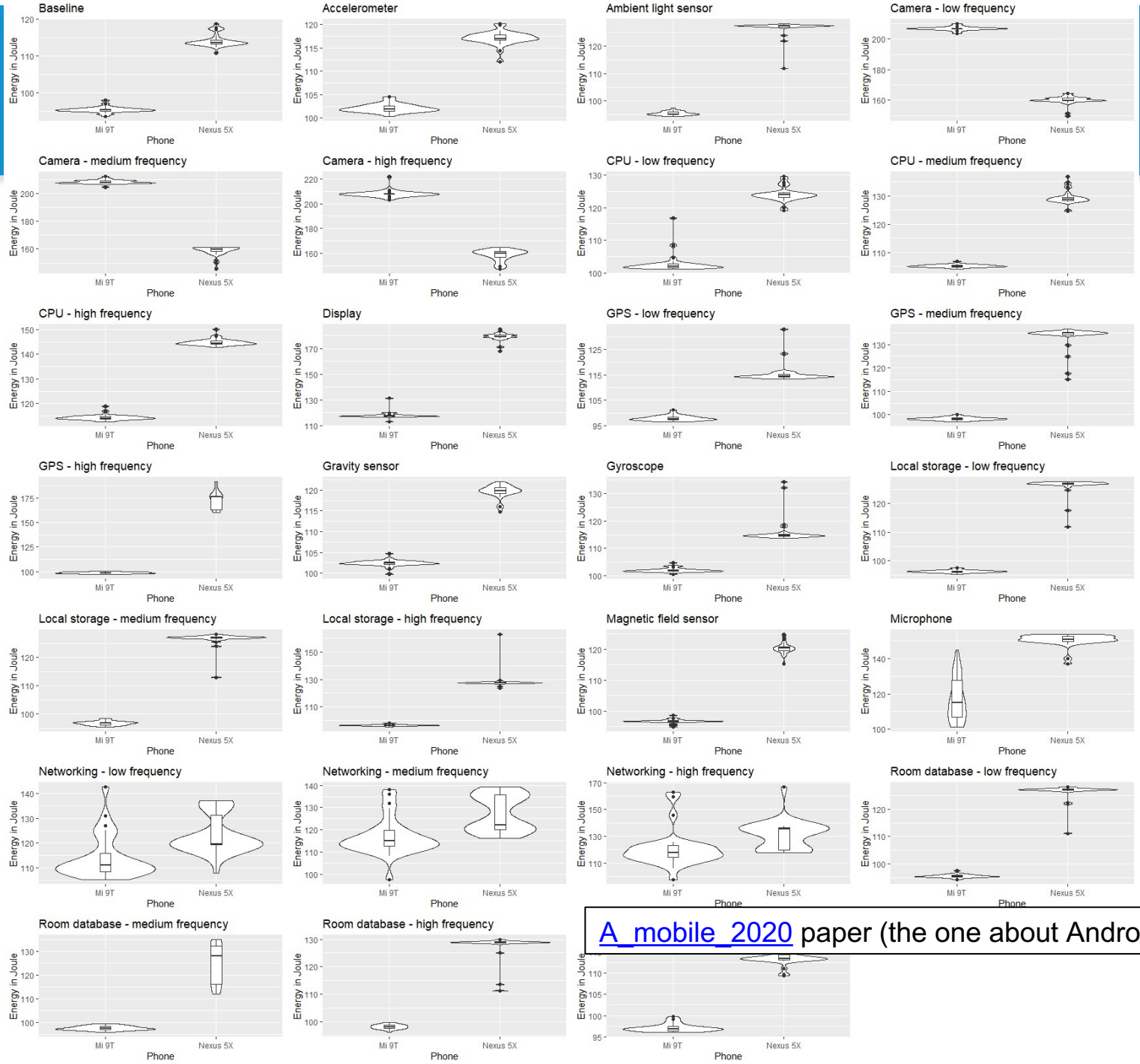


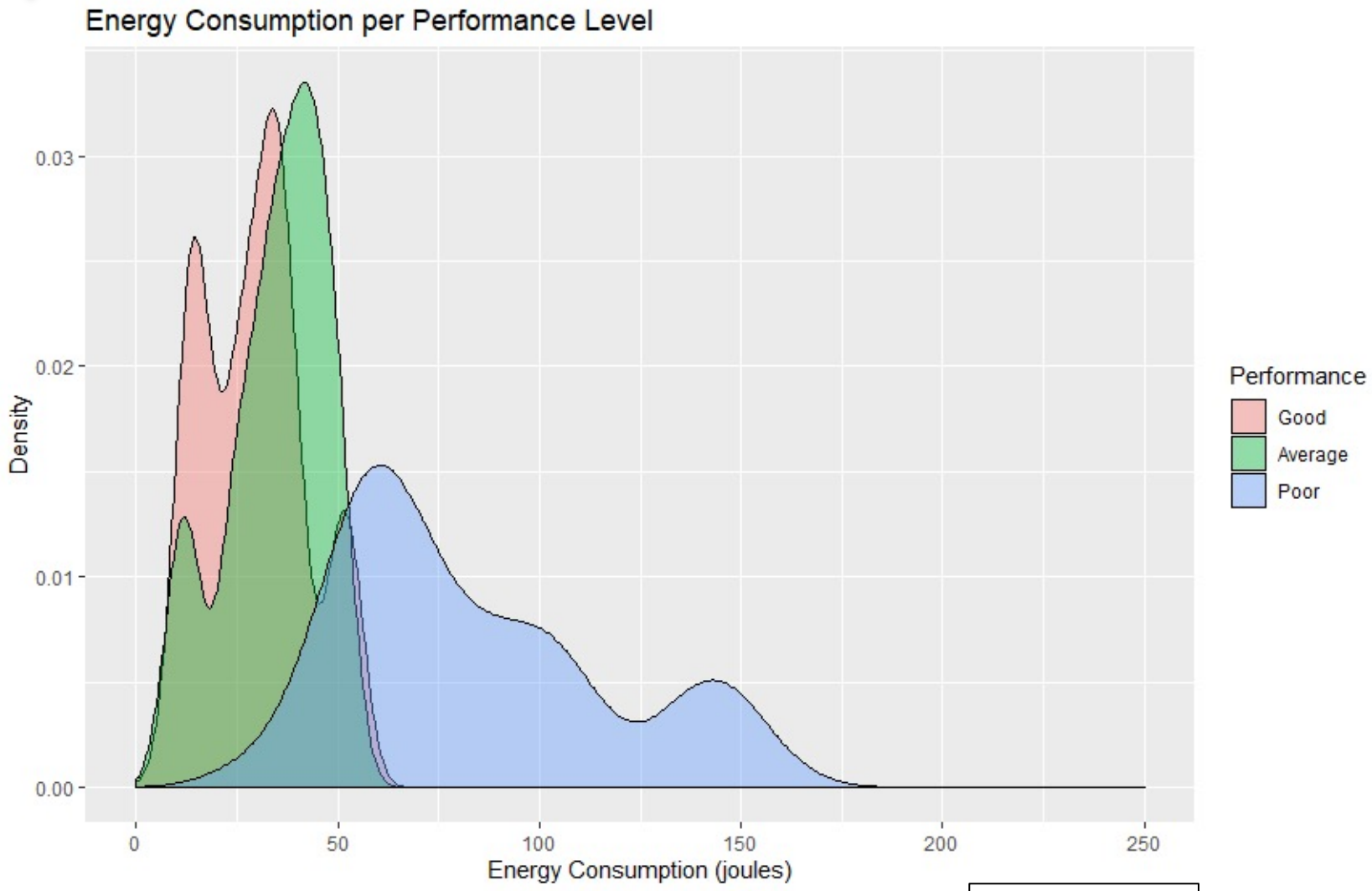
Fig. 4: Examples of power measurements across all movements and environments (baseline, combined)

MSR_2021

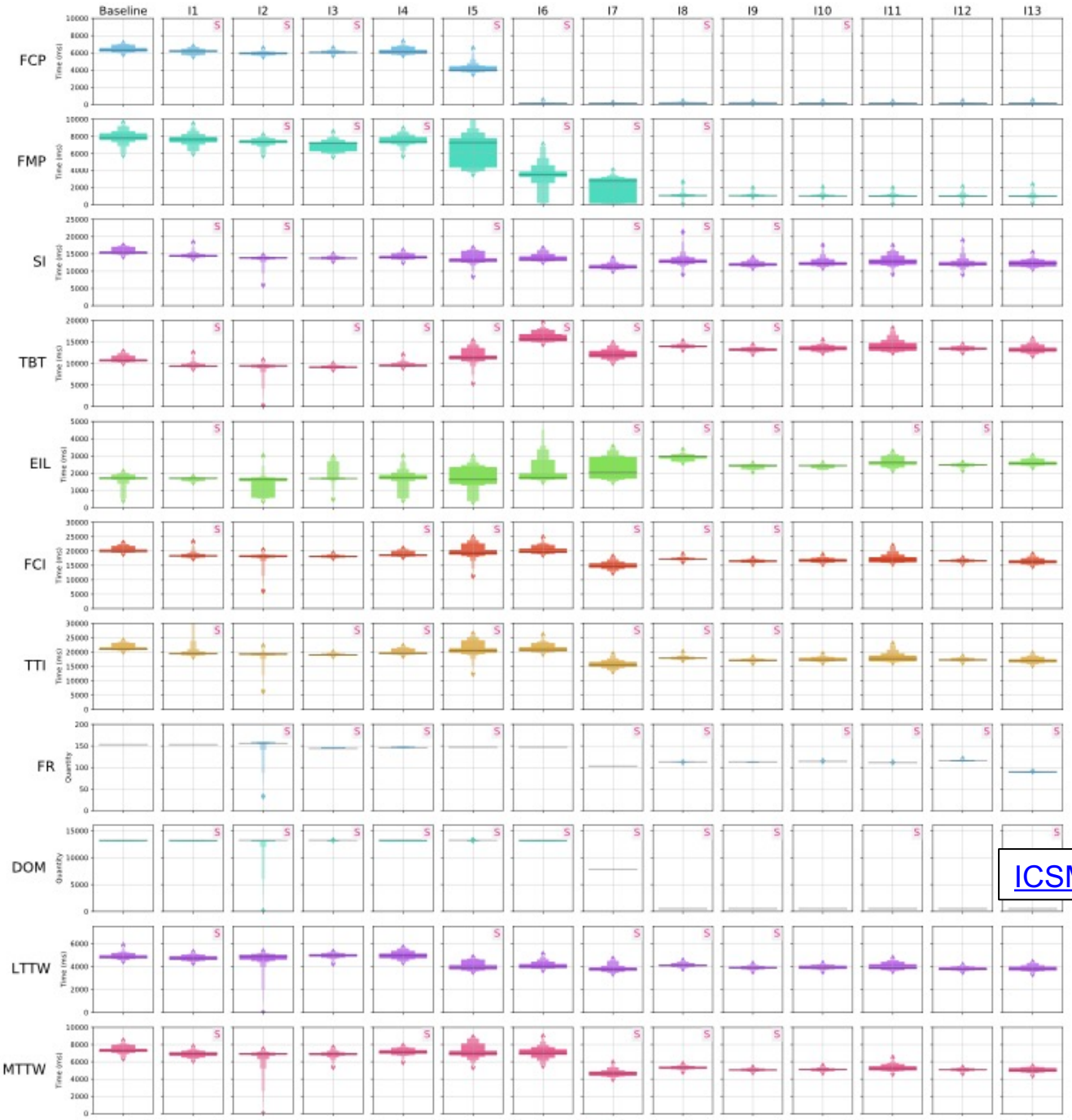


A_mobile_2020 paper (the one about Android Runner)

Examples of visualizations



EASE_2020



ICSME_2020

Fig. 2: Overview of the performance metrics across all interventions (S = statistically significant change after the intervention)

Tables can be used to complement with fine-grained data

TABLE III: Final results of the case study

Metric	Baseline	After I_{13}	Δ %	p-value
FCP	6,290.52	153.33	-97.56%	2.59e-11
FMP	7,646.05	953.33	-75.01%	2.59e-11
SI	15,310.98	12,271.07	-19.85%	1.16e-10
TBT	10,708.86	13,180.75	23.08%	1.42e-10
EIL	1,701.13	2,549.60	49.88%	2.59e-11
FCI	19,928.40	16,344.58	-17.98%	2.59e-11
TTI	21,018.35	16,958.08	-19.32%	2.59e-11
NR	153	153	-41.18%	1.97e-12
DOM	13,219	507	-96.16%	1.97e-12
LTTW	4,796.3	3,798.72	-20.80%	2.59e-11
MTTW	7,280.71	5,006.26	-31.23%	2.59e-11

ICSME_2020

Top-down approach

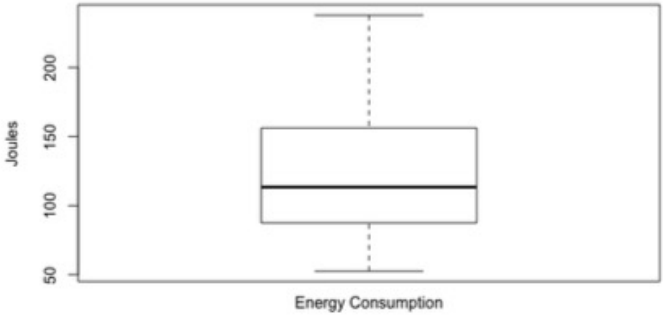


Fig. 2. Measured energy consumption values

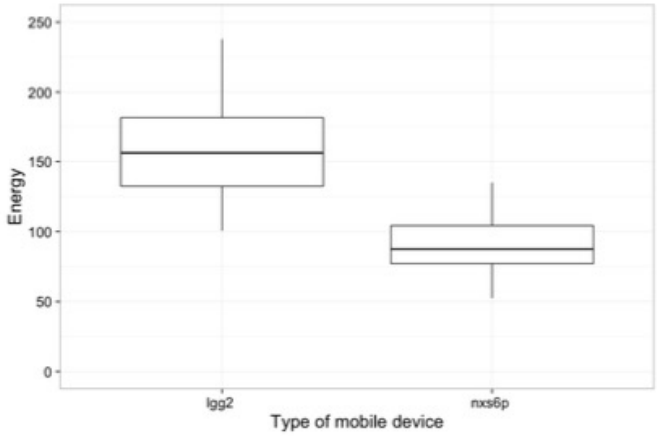


Fig. 3. Measured energy consumption values per mobile device (in Joules)

A. *Descriptive Statistics*

The energy consumption of all PWAs (two versions for each PWA, with service workers on and off) of our dataset is summarized in Table III.

Energy Consumption

Phone	Min.	Max.	Median	Mean	SD	CV
Both	52.44	237.90	113.41	124.33	45.32	0.36
LG G2	100.61	237.90	156.21	157.38	37.78	0.24
Nexus	52.44	134.97	87.44	91.28	22.14	0.24

TABLE III
DESCRIPTIVE STATISTICS FOR THE ENERGY CONSUMPTION VALUES (IN JOULES) – SD= STANDARD DEVIATION, CV = COEFFICIENT OF VARIATION)

MOBILESoft_2017

Show all subjects to spot interesting cases

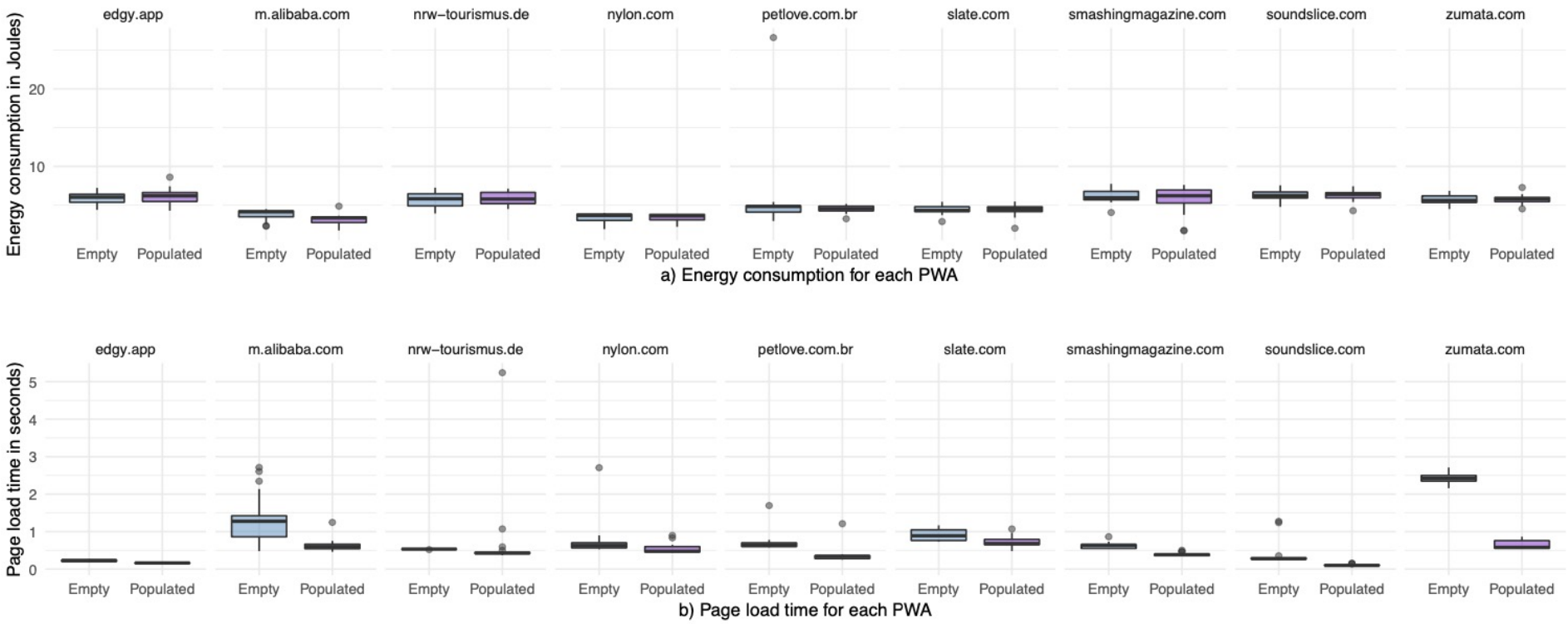
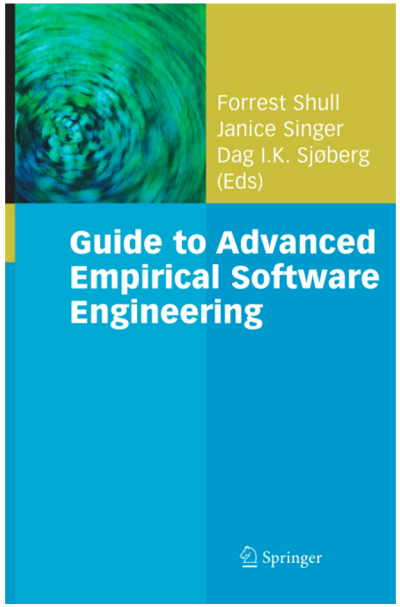


Figure 6: Energy consumption and page load time for each PWA

MOBILESoft_Caching_PWA_2020

More details? You have a full paper about experiment reporting!



Chapter 8

Table 2 Quick reference

Section	Content	Scope	Priority
3.1 Title			
3.2 Authorship			
3.3 Structured abstract	Backgro		
	Objectiv		
	Methods		
	Results		
	Limitati		
	Conclus		
3.4 Keywords			
3.5 Introduction	Problem state		
	Researcl objec		
	Context		
3.6 Background	Technol unde tigati		
	Alternat techr gies		
	Related		
	Relevan pract		
3.7 Experiment planning	Goals		

Table 2 (continued)

Section	Content	Scope
	Experimental units	From which popu sample be dra groups be forn treatments)? A omization and described
	Experimental material	Which objects are
	Tasks	Which tasks have the subjects?
	Hypotheses, parameters, and vari-ables	What are the cons operationalizat traceable deriv question respe experiment
	Design	What type of expe been chosen?
	Procedure	How will the expe collection) be instruments, m will be used a
	Analysis proce-dure	How will the data
3.8 Execution	Preparation	What has been do execution of th schedule, train
	Deviations	Describe any devi plan, e.g., how collection actu
3.9 Analysis	Descriptive statistics	What are the resul statistics?
	Data set prepa-ration	What was done to why, and how?
	Hypothesis testing	How was the data the analysis m
3.10 Discussion	Evaluation of results and implications	Explain the result of the results t especially thos *Background* se
	Threats to validity	How is validity of results assured actually valida

Table 2 (continued)

Section	Content	Scope	Priority
		Threats that might have an impact on the validity of the results as such (threats to internal validity, e.g., confound-ing variables, bias), and, further-more, on the extent to which the hypothesis captures the objectives and the generalizability of the find-ings (threats to external validity, e.g., participants, materials) have to be discussed	
	Inferences	Inferences drawn from the data to more general conditions	Required
	Lessons learned	Which experience was collected during the course of the experiment	Nice to have
3.11 Conclusions and future work	Summary	The purpose of this section is to provide a concise summary of the research and its results as presented in the former sections	Required
	Impact	Description of impacts with regard to cost, schedule, and quality, circumstances under which the approach presumably will not yield the expected benefit	
	Future work	What other experiments could be run to further investigate the results yielded or evolve the Body of Knowledge	
3.12 Acknowled-gements		Sponsors, participants, and contributors who do not fulfil the requirements for authorship should be mentioned	If appropriate
3.13 References		All cited literature has to be presented in the format requested by the publisher	Absolutely required
3.14 Appendices		Experimental materials, raw data, and detailed analyses, which might be helpful for others to build upon the reported work should be provided	Might be made available trough technical reports or web site

(continued)

Replication package

1. Data
   - raw data (usually CSV)
   - processed data (CSV)
   - (name it properly!)
2. Source code (commented!)
   - measurement scripts (Python)
   - data processing scripts (R)
   - analysis scripts (R)
3. Figures
   - exploratory figures
   - final figures
4. Text
   - README file
   - (optional) experiment notes

You have to follow this structure when providing the replication package of your final assignment!

VU

What this lecture means to you?

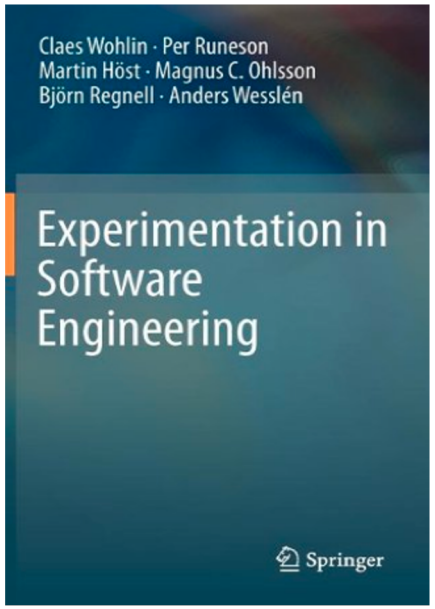
- You know how to **present** your experiment
  - push for clarity, soundness, and completeness
- **Replication package** needed
  - helps community building
  - enables replications
  - gives you also *visibility*!

Examples of replication package

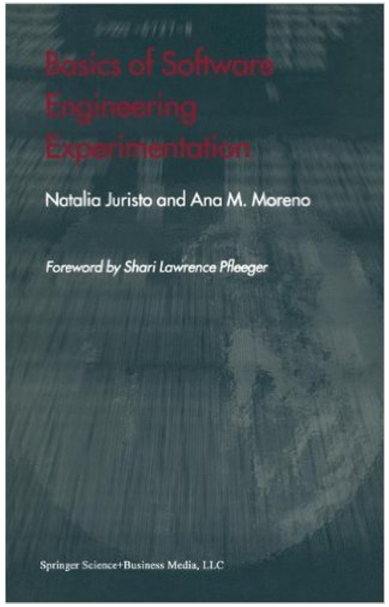
- https://github.com/S2-group/ease-2023-wasm-iot-rep-pkg
- https://github.com/S2-group/ICSME2018ReplicationPackage
- https://github.com/search?q=topic%3Areplication-package+org%3AS2-group&type=repositories&s=updated&o=desc

VU

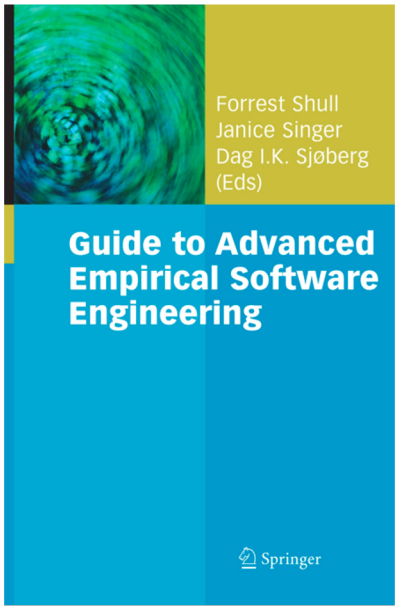
Readings



Chapter 11



Chapter 16



Chapter 8

BONUS → on-line book on data visualization
https://serialmentor.com/dataviz/