Experimental Methods in Computer Science

(*Metodologias Experimentais em Informática*)

Henrique Madeira

Master in Informatics Engineering
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia da Universidade de Coimbra
2023/2024

1

Introduction to the experimental method

21

Why do we need experiments?

Henrique Madeira, DEI-FCTUC, 2018-2023

22

Why do we need experiments?

Researchers:

- Collect evidences facts about the world (or system)
- Validate hypothesis
- Support the definition, validation, parameterization of models
- Validate models
- Confirm theories
- Etc, etc…

Henrique Madeira, DEI-FCTUC, 2018-2023

23

Why do we need experiments?

Engineers (including informatics engineers):

- Tune up systems
- Compare and select among different project choices
- Verify that requirements or specifications are met
- Validate mechanisms and/or solutions
- Measure/evaluate features, e.g., to access efficiency of mechanisms
- Assess the effectiveness of processes, e.g., software development processes
- Etc, etc…

"Experimentation as the feedback step in the engineering loop"

Henrique Madeira, DEI-FCTUC, 2018-2023

24

Why do we need experiments?

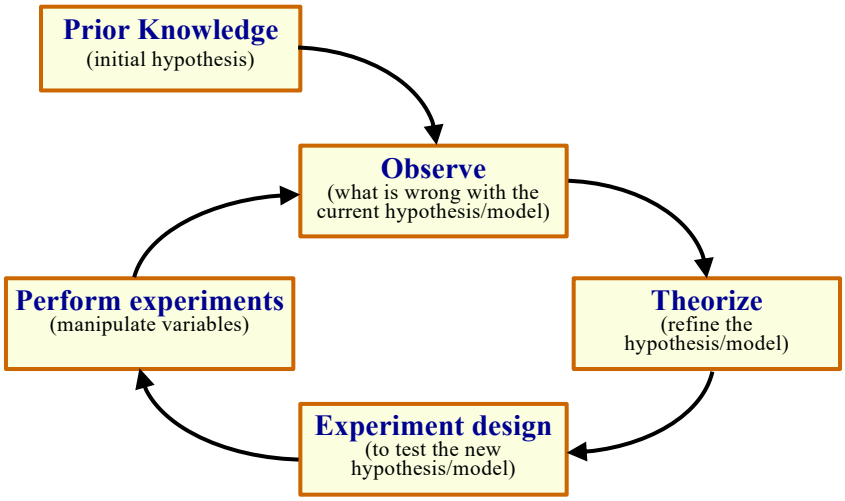
Business engineers, entrepreneurs, managers:

- Survey to confirm product market potential or to select product features (product-oriented, marketing-oriented)
- Select the best internal organization strategies for companies (production-oriented, management-oriented)

Henrique Madeira, DEI-FCTUC, 2018-2023

25

Scientific method (oversimplified)



(adapted from Steve Easterbrook, UT, slides)

Henrique Madeira, DEI-FCTUC, 2018-2023

26

Scientific method (oversimplified)

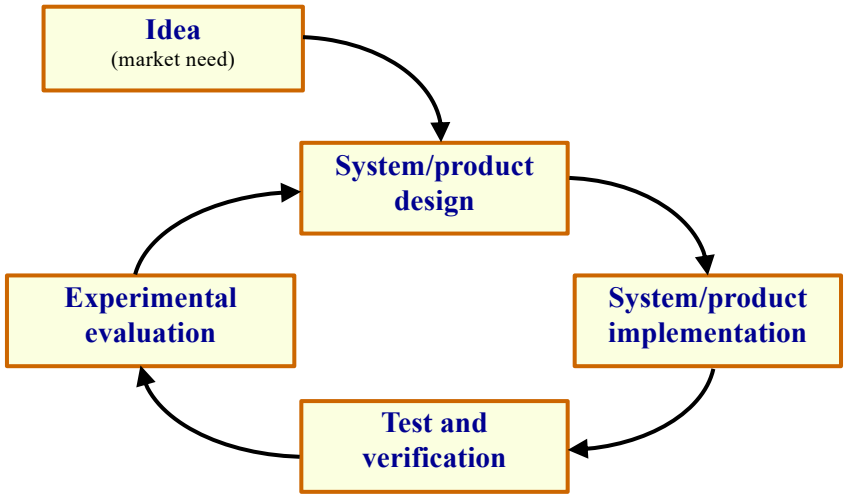
- Use previous knowledge and observations to gain insight about a phenomenon
- Formulate hypothesis
- Construct a model of the phenomenon
- Use model (hypothesis) to predict outcomes
- Test hypothesis by experimenting
- Analyze outcome of experiment
- Go back to the beginning, refine, …

Experiments (or observation of reality) may invalidate models; not the opposite.

Henrique Madeira, DEI-FCTUC, 2018-2023

27

Experiments in system/product design



Henrique Madeira, DEI-FCTUC, 2018-2023

28

Typical computer science/ informatics engineering scenario

- A particular task needs to be solved by a software system
- This task is currently solved by an existing system (a baseline)
- You propose a new, in your opinion, better system
- You argue that your proposed system is better than the baseline
- You support your arguments by providing evidence that your system indeed beats the baseline
- The ultimate evidence is done by experimentation (and very often it is the easiest way as well)

Henrique Madeira, DEI-FCTUC, 2018-2023

29

Be careful while doing experiments

Some questions:

- Was the experiment designed correctly (concerning its goals)?
- Was it based on a toy scenario or in a realistic one (or in a real situations)?
- Were the measurements used appropriate for the goals of the experiment?
- Was the experiment run for long enough time?
- Is it possible to reproduce the experiment (same team) and get similar results?
- Is it possible to replicate the experiment (other team)?
- Are the conclusions correct?
- Are the results not biased?
- Is the generalization of the results credible?

Henrique Madeira, DEI-FCTUC, 2018-2023

30

Responsible skepticism

- Constantly look for
  - Failures in experimental designs
  - Failures of observations
  - Gaps in reasoning
  - Alternative explanations
- Compare new evidence against old
- Raise counter objections/hypotheses
- Question grounds for doubt as well
- Accumulate weight of evidence

Henrique Madeira, DEI-FCTUC, 2018-2023

(from Dewayne E. Perry, UT slides)

31

Key properties

- **Relevance**
  Are the goals of the experiment and the expected results important for the progress (do they have impact in science, technology, market, etc.)?
- **Representativeness**
  Is the experiment realistic and representative of real-world scenarios?
- **Repeatability**
  Is it possible to repeat the experiment and achieve same or statistically similar results?
- **Reproducibility**
  Is there enough information to allow others to reproduce the experiment?
- **Results analysis and generalization**
  Is the analysis of results sound? Is the generalization of conclusions credible?
- **Cost**
  Is the cost of the experiments compatible with the expected benefits?

Henrique Madeira, DEI-FCTUC, 2018-2023

32

Many types of experiments...

- Controlled experiments
- Field studies
- Case studies
- Pilot studies
- Benchmarks
- Simulations
- Surveys
- Rational reconstructions
- Artifact/archive analysis
- Ethnographies
- Quasi-experiments

Henrique Madeira, DEI-FCTUC, 2018-2023

33

Laboratory experiments/ controlled experiments

Experimental investigation of a testable hypothesis, in which conditions are set up to isolate the variables of interest ("independent variables") and test how they affect certain measurable outcomes (the "dependent variables")

- Good for
  - Quantitative analysis of benefits of a particular tool/technique
  - Establish cause-and-effect in a controlled setting
  - (demonstrating how scientific we are… and getting advertising benefits)
- Limitations
  - Hard to apply if you cannot simulate the right conditions in the lab
  - Limited confidence that the lab setup reflects the real situation
  - Ignores contextual factors (e.g. social/organizational/political factors)
  - Extremely time-consuming

This type is our main target. Examples?

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 34

34

Field studies

Exploratory study, used where little is currently known about a problem, or where we wish to check that our research goals are grounded in real-life settings.

Good for
- Setting a research/innovation agenda (what really matters?)
- Understanding the context for exploratory problems

Examples?

- Limitations
  - Hard to build generalizations (results may be organization specific)
  - Observers' bias

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 35

35

Case studies

A technique for detailed exploratory investigations, both prospectively and retrospectively, that attempt to understand and explain phenomenon or test theories, using primarily qualitative analysis

- Good for
  - Answering detailed how and why questions
  - Gaining deep insights into chains of cause and effect
  - Testing theories in complex settings where there is little control over the variable
- Limitations
  - Hard to find appropriate case studies
  - Hard to quantify findings

Examples?

Henrique Madeira, DEI-FCTUC, 2018-2023

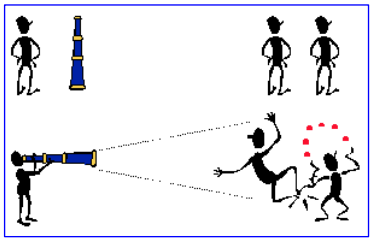
36

Pilot studies

Controlled introduction of a tool/technique into a real project, where the researcher can no longer control the context, but where the net effect can be measured (e.g. against a baseline, or against previous experience)

- Good for
  - Measuring the benefits in a real setting
  - Preparation for tech. transfer
  - Getting organizations interested in your work
- Limitations
  - Hard to get organizations to adopt unproven ideas
  - Hawthorn effect (and other bias problems)

Examples?



Henrique Madeira, DEI-FCTUC, 2018-2023

37

Benchmarks

A test or set of tests used to compare alternative tools or techniques. A benchmark comprises a motivating comparison, a task sample, and a set of measures (e.g., performance)

- Good for
  - Making detailed comparisons between methods/tools
  - Increasing the (scientific) maturity of a research community
  - Building consensus over the valid problems and approaches to them
- Limitations
  - Can only be applied if the community is ready
  - Become less useful / redundant as the research paradigm evolves
  - Normally quite expensive

Examples?

Henrique Madeira, DEI-FCTUC, 2018-2023

38

Simulations (SW Eng.)

An executable model of the software development process, developed from detailed data collected from past projects, used to test the effect of process innovations

- Good for
  - Preliminary test of new approaches without risk of project failure
  - Each test is relatively cheap [Once the model is built]
- Limitations
  - Expensive to build and validate the simulation model
  - Model is only as good as the data used to build it
  - Hard to assess scope of applicability of the simulation

Henrique Madeira, DEI-FCTUC, 2018-2023

39

Surveys

A comprehensive system for collecting information to describe, compare or explain knowledge, attitudes and behavior over large populations

- Good for
  - Investigating the nature of a large population
  - Testing theories where there is little control over the variables
- Limitations
  - Relies on self-reported observations
  - Difficulties of sampling and self-selection
  - Information collected tends to subjective opinion

Henrique Madeira, DEI-FCTUC, 2018-2023

40

Rational reconstructions

A demonstration of a tool or technique on data taken from a real case study, but applied after the fact to demonstrate how the tool/technique would have worked

- Good for
  - Initial validation before expensive pilot studies
  - Checking the researcher's intuitions about what the tool/technique can do
- Limitations
  - Potential bias (you knew the findings before you started)
  - Easy to ignore "signal-to-noise ratio"

Henrique Madeira, DEI-FCTUC, 2018-2023

41

Artifact/archive analysis (SW Eng.)

Investigation of the artifacts (documentation, communication logs, etc.) of a software development project after the fact, to identify patterns in the behavior of the development team.

- Good for
  - Understanding what really happens in software projects
  - Identifying problems for further research
- Limitations
  - Hard to build generalizations (results may be project specific)
  - Incomplete data

Henrique Madeira, DEI-FCTUC, 2018-2023

42

Ethnographies (SW Eng.)

Interpretive, in-depth studies in which the researcher immerses herself in a social group under study to understand phenomena though the meanings that people assign to them

- Good for
  - Understanding the intertwining of context and meaning
  - Explaining cultures and practices around tool use
- Limitations
  - No generalization, as context is critical
  - Little support for theory building

Henrique Madeira, DEI-FCTUC, 2018-2023

43

Quasi-experiments

Empirical study used to estimate the causal impact of an intervention (treatment, change, etc.) on its target population

- Good for
  - Understanding how a given population react to some change (e.g., treatment, procedure, etc.)
  - Allow some generalization of the results to the entire population
- Limitations
  - Results are subject to contamination by confounding variables
  - Limitation on the establishment of the causal impact

44

Design of experiments (a first look)

- Design of experiments (often referred as experimental design as well) is the process of systematically defining and planning experiments in such a way that the date obtained can be analyzed to draw valid and objective conclusions
- **The goal is to design and perform valid experiments that allow good technical decisions: characterize and optimize process/product decisions**
- Basic idea:
  - Introduce controlled changes to input variables in order to to study their effect on an observable variable (or variables)
  - Get the maximum amount of information on cause and effect relationships with the minimum effort

45

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify variables
3. Generate hypothesis
4. Define the experimental setup/scenario
5. Develop tools and procedures for the experiment

Design of the experiment

6. Run experiments and collect the data/measurements — Measurements
7. Perform data analysis — Analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis) — Conclusions

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 46

46

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify va
3. Generate h
4. Define the
5. Develop to
6. Run experi
7. Perform da
8. Draw conc
reformulate the problem statement or test a different hypothesis)

A good (i.e., relevant) problem statement should be focused enough to allow the clear identification of the variables of the problem but, at the same time, should be sufficiently open to allow different hypothesis to answer the problem/question.

Possible generic formulation:
How does X affect Y under conditions Z?

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 47

47

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify va
3. Generate h
4. Define the
5. Develop to
6. Run experi
7. Perform da
8. Draw conc
   reformulate the problem statement or test a different hypothesis)

To formulate good problem statements:

- Must know the subject area: process, system, technique, product, product market, etc.
- Must be precise and clear
- Must be sure that the problem/question is relevant

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 48

48

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify variables
3. Genera
4. Define
5. Develo
6. Run ex
7. Perfor
8. Draw c
   reformu

- **Dependent variable** (response variable)
  Measured output (e.g., response time, throughput, no. bugs, downtime, latency, error detection coverage, etc., etc.)
- **Independent variables** (factors)
  Input variables that can be changed in the experiment (e.g., memory size, clock rate, file size, channel bandwidth, etc., etc.)
- **Levels**
  Values taken by the variables. Can be (nearly) continuous (e.g., ~time, size in bytes) or discrete (type of system, type of algorithm, etc.)

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 49

49

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify variables
3. Genera
4. Define
5. Develo
6. Run ex
7. Perfor
8. Draw c
   reformu

- **Change one factor at the time**
  Simple scenario. The analysis is simple, as it is easy to understand the effect of a given factor on the independent variable.
- **Full factorial**
  Change two or more factors simultaneously. Much more complex (to be seen in detail later on in the course). Has two important advantages:
  - More efficient experiments (save time and effort)
  - Allow the study of possible interactions among factors

Henrique Madeira, DEI-FCTUC, 2018-2023

50

Design of experiments (a first look)

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify variables
3. Genera
4. Define
5. Develo
6. Run ex
7. Perfor
8. Draw c
   reformu

Terminology:

- **Baseline** (often called **golden run**)
  Set of factor values (i.e., independent variables) that represent a baseline scenario for the experiments
- **Repetition of golden run**
  Used to estimate the experimental error (noise) in the system and identify small effects that may cause variations in the results.

Henrique Madeira, DEI-FCTUC, 2018-2023

51

Design of experiments (a first look)

Laboratory

1. Problem st
2. Identify v
3. Genera
4. Define
5. Develo
6. Run ex
7. Perfor
8. Draw c

reformu

Terminology:

- **Randomization**
  Minimize potential uncontrollable biases in the experiments by randomly assigning factors to "average out" the effects of possible extraneous factors.
- **Blocking**
  The experiment is divided in homogeneous segments (blocks such as sets of machines, users, loads, etc.) to improve precision. The goal is to control the variability block to block.
- **Confounding variable**
  Extraneous variable that influences the relationship between the dependent and independent variables (i.e., correlates with both the dependent and independent variables).

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 52

52

Design of experiments (a first look)

Laboratory experim

1. Problem statement (o
2. Identify variables
3. Generate hypothesis
4. Define the experimen
5. Develop tools and pro
6. Run experiments and
7. Perform data analysis
8. Draw conclusions (oft

reformulate the problem statement or test a different hypothesis)

- Hypothesis describe provisional relationships between factors (independent variables) and the response variable (dependent). It is a interim answer to the problem statement.
- Can be directional or non-directional
- May lead to a model allowing prediction of what is going to happen in future cases.
- Quite often (in computers) the goal of the experiments is to quantify the relationship (not just confirm that exists)

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 53

53

Design of experiments (a first look)

Laboratory ex

1. Problem stat
2. Identify varia
3. Generate hyp
4. Define the experimental setup/scenario
5. Develop tools and procedures for the experiment
6. Run experiments and collect the data/measurements
7. Perform data analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis)

- Experiment complexity
- Experiment cost
- Availability of tools and frameworks that may help
- Degree of automation

54

Design of experiments (a first look)

Labora

1. Probl
2. Ident
3. Gene
4. Defi
5. Deve
6. Run experiments and collect the data/measurements
7. Perform data analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis)

- Continuous and/or discrete measurements
- Accuracy, precision, and resolution
- Basic measurements in computers…
  - Count
  - Duration
  - Size
  - Any value derived from the combination of basic measurements

55

Design of experiments (a first look)

Labor

1. Prob
2. Iden
3. Gen
4. Defi
5. Dev
6. Run experiments and collect the data/measurements
7. Perform data analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis)

- Exploratory data analysis
- Statistical data analysis
  - Tables, charts, etc., average, standard deviation
  - Coping with measurement errors
  - Confidence intervals
  - Statistical comparison of alternatives
  - Tests to check if measured data fit know distributions (chi-square, K-S tests,…)

Henrique Madeira, DEI-FCTUC, 2018-2023

56

Design of experiments (a first look)

Labor

1. Prob
2. Iden
3. Gen
4. Defi
5. Dev
6. Run
7. Perform data analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis)

- The written **report** of the experiments is quite often the **single outcome** of months or years of work
- **Quality of writing** is essential. Some relevant attributes of the report:
  - Clear (in the goals, approach, setup, steps, analysis, discussion, conclusions)
  - Credible (in the data reported, conclusion, etc.)
  - Self-contained

Henrique Madeira, DEI-FCTUC, 2018-2023

57