Blekinge Institute of Technology

Claes Wohlin

Empirical Software Engineering with Human Subjects

Claes Wohlin

Blekinge Institute of Technology

Empirical research methods

Empirical research methods include:

- **Controlled experiments**
- **Case studies**
- **Surveys**

In many cases the intention is to perform a statistical analysis to determine whether one technique or method is better than another.

Engineering perspective

Software development and maintenance is an engineering discipline. Key issues:

- **Understand**
- **Monitor**
- **Estimate/predict**
- **Control/manage**
- **Evaluate/validate**
- **Improve**

To achieve this measurement and empiricism are crucial.

Empirical studies as a research method

- Empirical studies are important in software engineering because it is not solely a technical discipline.
- It is a combination of technical issues, social science and organizational aspects.
- Empirical methods are frequently used in social and human sciences.

Survey

- **A survey can be based on available literature, experiences stored in an experience base, interviews or it can be based on subjective expert judgement. The survey results can be used for a desktop evaluation.**

Experiment

- **Experiments are carefully planned and fully controlled. An experiment should be replicable, i.e. somebody else should be able to repeat it.**
  - **This type of method will be used to exemplify empirical studies.**

Case study

- **A case study is normally a study conducted in parallel with the project execution. It should be planned in advance, but we have less control over the execution than in an experiment. We are normally external observers of a ”real” software project.**

Empiricism meets engineering 1(3)

- ***Confirmation*** **of more or less accepted hypotheses. *For example*: object-orientation is good for reuse.**
- ***Evaluation*** **of methods, models, languages and tools. *For example*, whether Java produces higher quality code than C++.**
- ***Identification*** **of relationships. *For example*: find a relationship between fault-prone components and design concepts.**

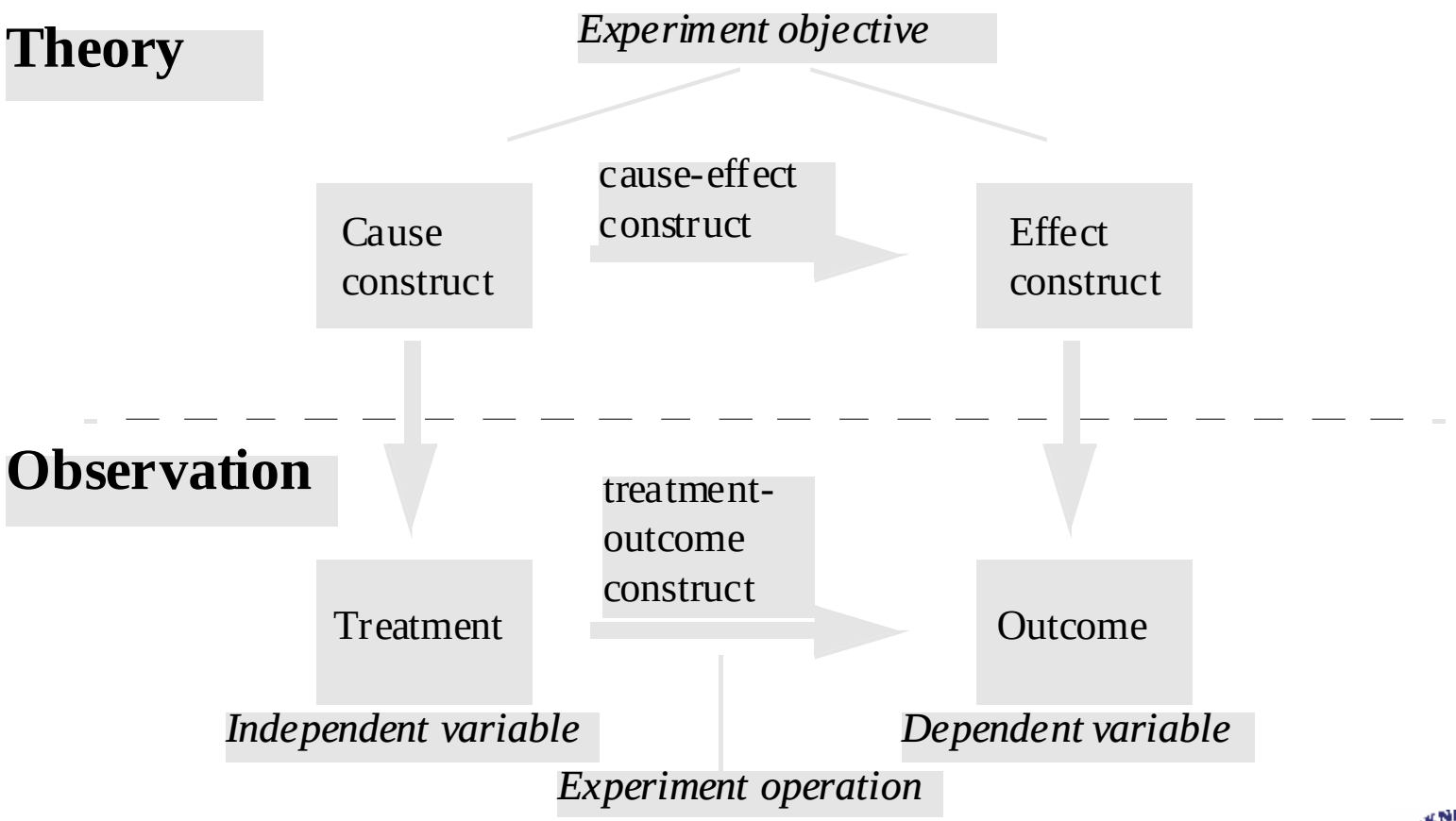
Empiricism meets engineering 2(3)

- ***Validation*** **of models or measures. *For example*, to validate a specific cost estimation model.**
- ***Understanding*** **of methods, techniques and models. *For example*, to understand the relationship between inspections and test.**

Empiricism meets engineering 3(3)

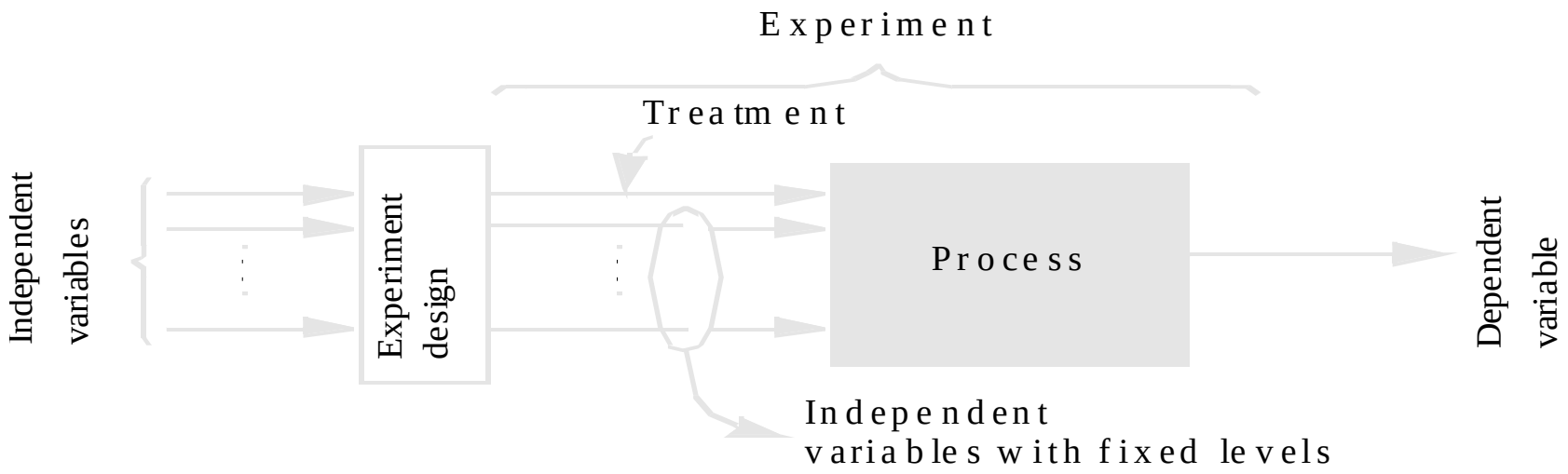
- ***Guidance/control*** **to help in management. *For example*, as input to planning of personnel to software inspections.**
- ***Change/improve*** **to support decision-making with respect to changes. *For example*, the result of a study can help us to decide whether or not to introduce a new development tool.**

Experiment principle

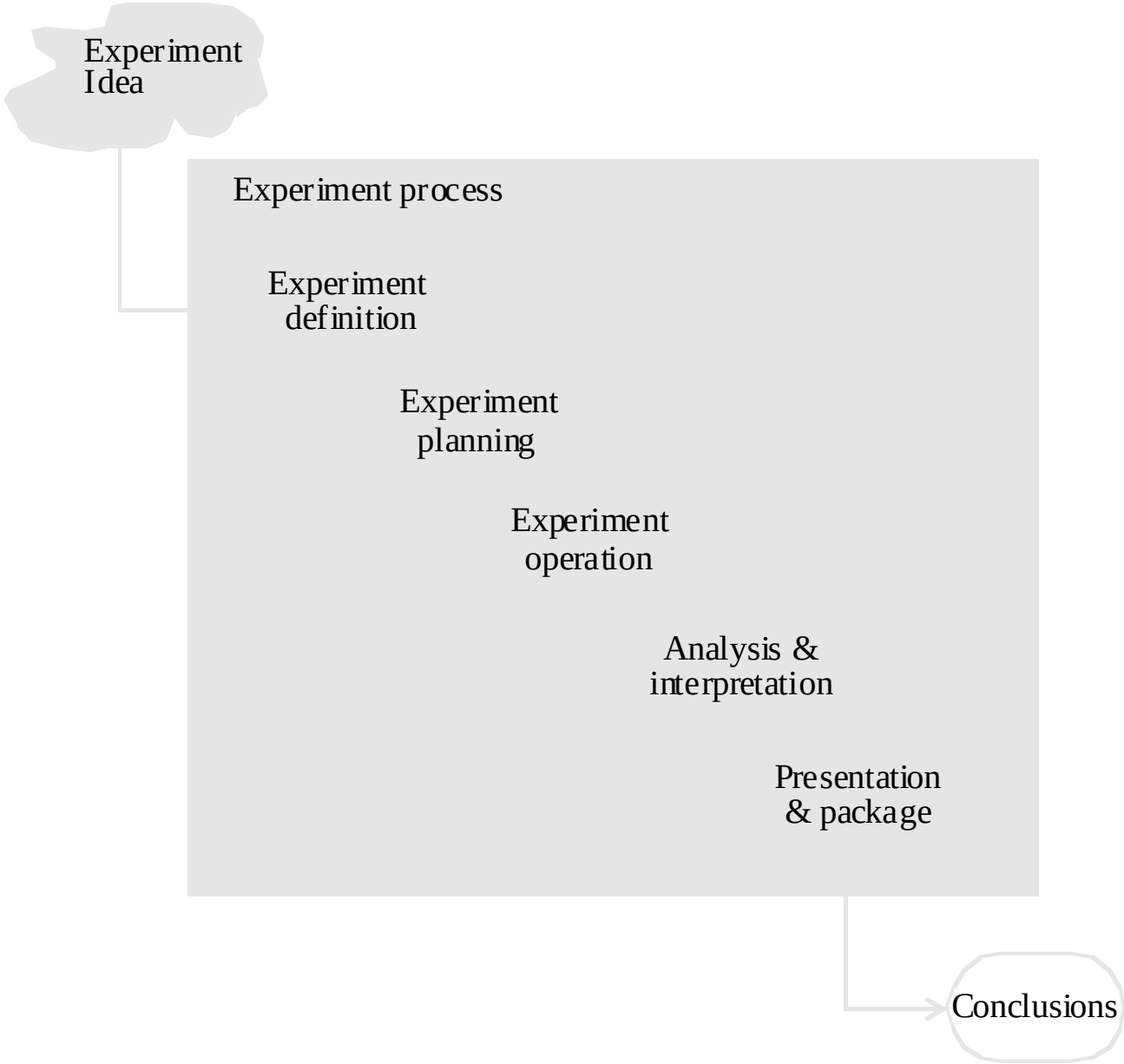


Independent and dependent

Illustration of experiment



Experiment process



Experiment definition

- The goal template is:

  Analyse <*Object(s) of study*>

  for the purpose of <*Purpose*>

  with respect to their <*Quality focus*>

  from the point of view of the <*Perspective*>

  in the context of <*Context*>.

An example definition

Analyse *the PBR and checklist techniques*

for the purpose of *evaluation*

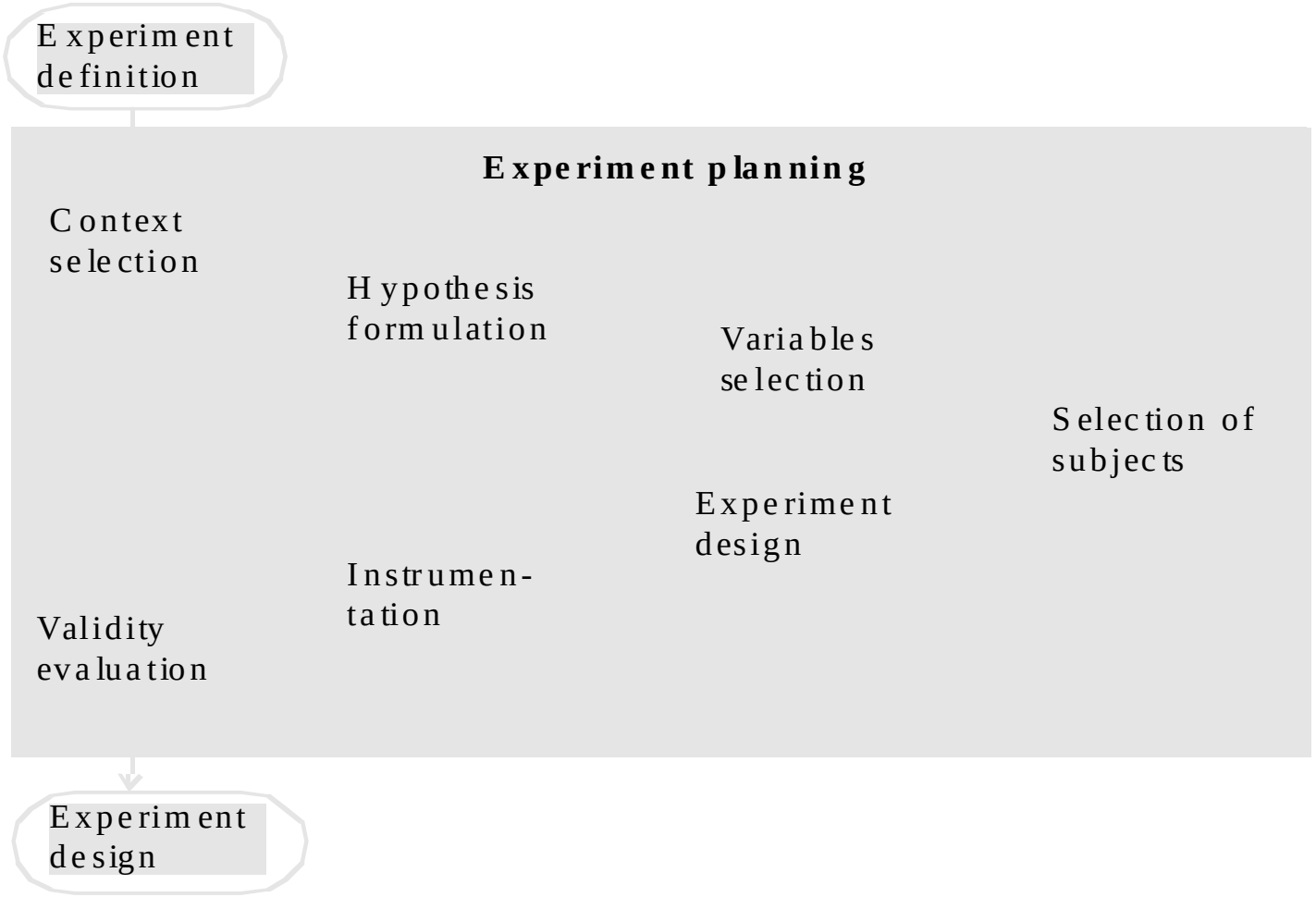
with respect to *effectiveness and efficiency*

from the point of view of *the researcher*

in the context of *students reading requirements documents*.

PBR – Perspective-based reading

Experiment planning



Planning phase overview

Steps in planning 1(4)

- **Context:**
  - **Off-line vs. On-line**
  - **Students vs. Professionals**
  - **Toy vs. Real problems**
  - **Specific vs. General**
- **Hypothesis formulation:**
  - **Null hypothesis (no real underlying trend or pattern) and alternative hypothesis. The objective is to reject the null hypothesis with as high significance as possible.**

Steps in planning 2(4)

- **Variables:**
  - **Independent (input)**
  - **Dependent (output)**
- **Subjects**
  - **Sampling strategy**
- **Design principles**
  - **Randomisation**
  - **Blocking (e.g. on experience)**
  - **Balancing (same number of subjects in groups)**

Steps in planning 3(4)

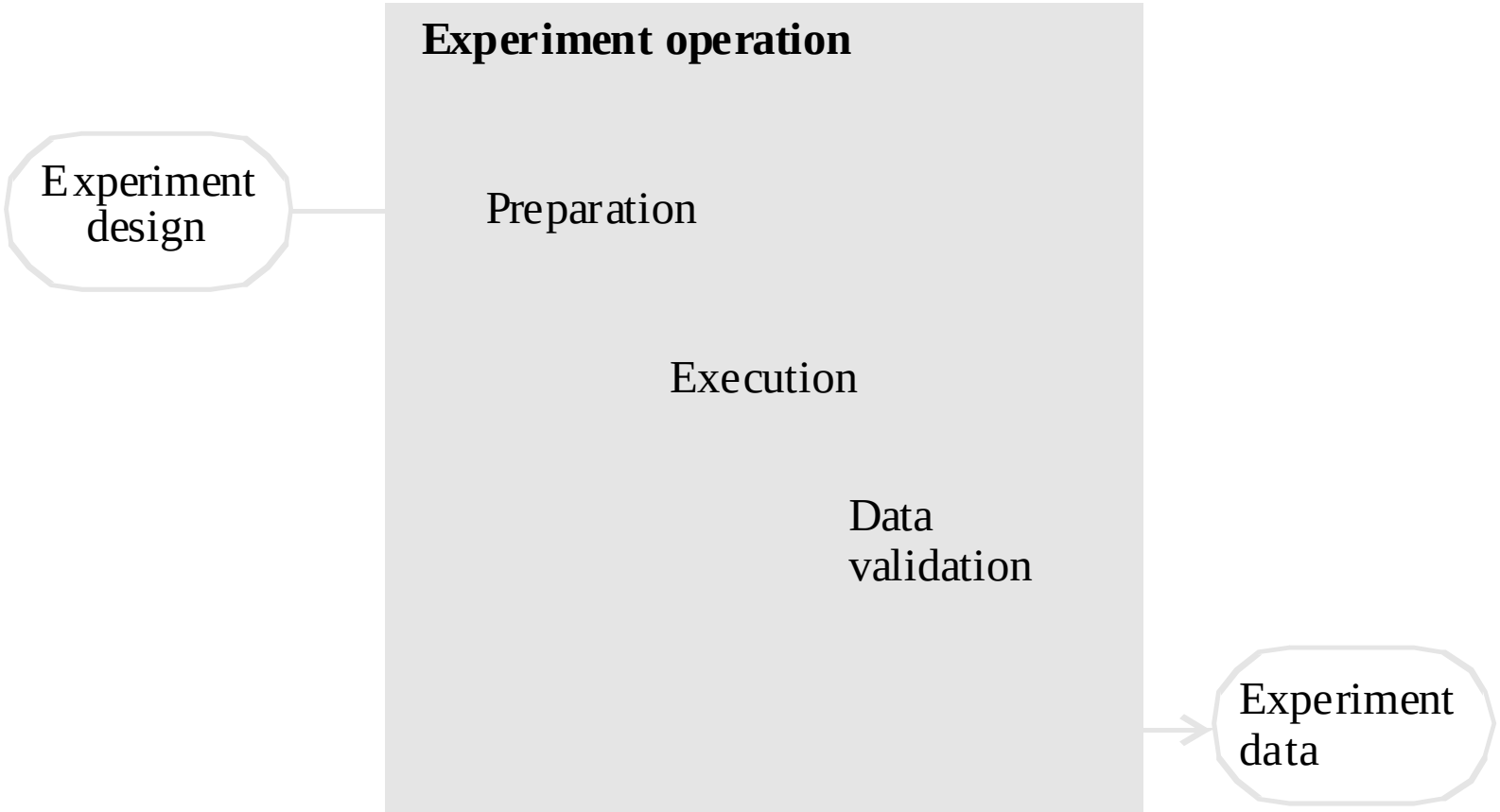
- **Design types**

  A large number of standard designs exist, and we should select an appropriate design type depending on treatments and number of subjects and of course the objective (hypothesis) of the experiment.

Steps in planning 4(4)

- **Instrumentation**
  - **Objects**
  - **Guidelines**
  - **Measurement instruments**
- **Validity evaluation**
  - **Conclusion validity: treatment to outcome**
  - **Internal validity: treatment causes outcome**
  - **Construct validity: theory to observation**
  - **External validity: generalization**

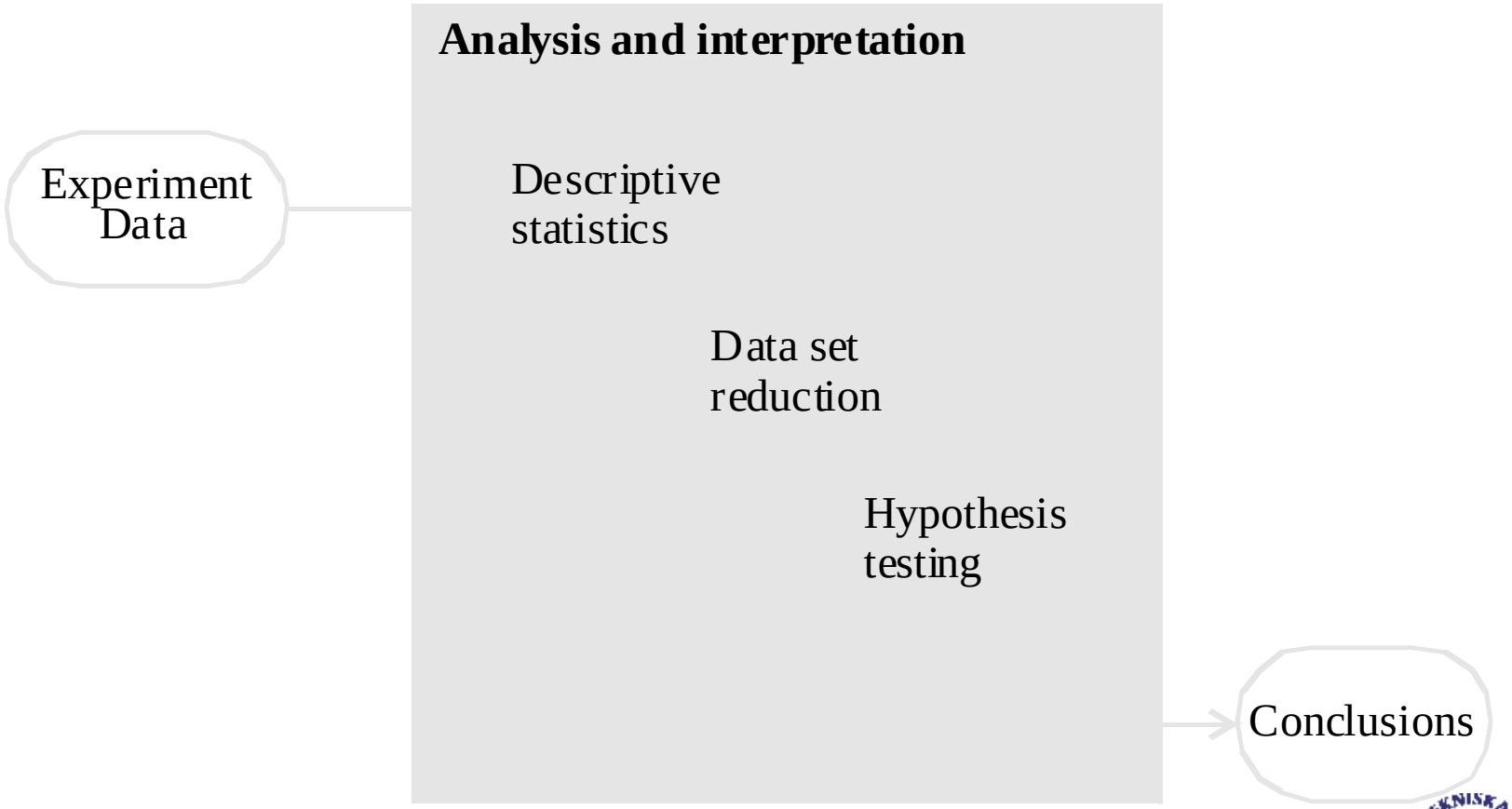
Experiment operation



Steps in operation

- **Preparation**
  - **Commit participants**
  - **Instrumentation (availability)**
- **Execution**
  - **Data collection**
  - **Experimental environment**
- **Data validation (general check)**

Analysis and interpretation



Steps in analysis 1(2)

- **Descriptive statistics**
  - **Scale types (nominal, ordinal, interval and ratio)**
  - **Measures of central tendency, dispersion and dependency**
  - **Graphical visualization**
- **Data set reduction**
  - **Outliers**

Steps in analysis 2(2)

- **Hypothesis testing**
  - **Parametric tests (assumes a specific distribution, usually a normal distribution)**
  - **Non-parametric tests (no assumption on distributions)**

The different types of tests are related to the standard design types. The intention is to be able to reject the null hypothesis with a statistical significance.

Interpretation

The statistical analysis forms the basis for interpretation.

The interpretation is the foundation for decision-making based on engineering principles.

Packaging

Report outline:

- **Introduction**
- **Problem statement**
- **Experiment planning**
- **Experiment operation**
- **Data analysis**
- **Interpretation of results**
- **Discussion and conclusions**
- **Appendix**

Additional concerns

- **Triangulation**
- **Replication**
- **Lab packages**
- **Meta-analysis**

Simplistic example: experiment

- **Problem: We want to evaluate reading techniques for inspections.**
- **We have two competing methods. State a hypothesis, for example method A supports defect detection in requirements specifications better than method B.**
- **Let people inspect a requirements specification with a known number of defects.**
- **Use a statistical method to evaluate the hypothesis.**
- **Determine which method is the best.**
- **Decide whether or not to start using the method.**

Example: case study

- **Hypothesis: Defects in testing can be estimated from design measures.**
- **Collect failure data from testing.**
- **Collect design metrics.**
- **Build a model using statistical techniques.**
- **Validate the model in the next release.**
- **Use the model in the following release.**

Some conclusions

- There is a lack of validated results in the field,
- Measurement is a key issue to success,
- Empirical studies is needed in software engineering research,
- Empirical studies mean that the human dimension in software development can be included in the analysis.

Some sources of information

- **ESERNET: network of excellence for experimental software engineering (www.esernet.org)**
- **ISERN: International Software Engineering Research Network (www.iese.fhg.de/ISERN/)**
- **CeBASE: NSF-funded initiative in the US for empirically based software engineering (www.cebase.org)**

Selected references

- Basili, Selby and Hutchens, ”Experimentation in Software Engineering”, IEEE Transactions on Software Engineering, Vol. SE-12, No. 7, pp. 733-743, 1986.
- Wohlin et al., ”Experimentation in Software Engineering – An Introduction”, Kluwer Academic Publishers, 1999.
- Juristo and Moreno, ”Basics of Software Engineering Experimentation”, Kluwer Academic Publishers, 2001.
- Fenton and Pfleeger, ”Software Metrics. A Rigorous and Practical Approach”, International Thomson Computer Press, 1996.

Examples of experiments

The following slides present some of the controlled experiment we have conducted over the years.

Inspection effectiveness

Question: Is it possible to draw general conclusions from several published inspection studies?

Study: Variant of meta-analysis

- **Individual performance**
- **Number of participants**
- **Benchmarking**

Outcome: Some interesting results can be found, for example, in terms of what can be expected for a group of inspectors.

Reading techniques

Question: Is it possible to identify reading techniques for inspections that are more effective than others?

Study: Several inspection experiments with two groups

Outcome: There are differences, but it is hard to find a pattern. Usage-oriented inspections seems promising.

Capture-Recapture

Question: Is it possible to estimate the number of remaining faults after an inspection and which is the best estimation method?

Study: We have run a series of experiments in this area, including the first controlled experiment.

Outcome: It is possible to make estimations. However, the estimation error is likely to be around 20%.

Design rationale

Question: Does a design rationale make it faster and more correct to make changes in an existing system?

Study: Two groups

Outcome: Yes, but no significant results. Qualitatively the participants viewed the rationale as important.

Effort estimation

Question: What is the best way to conduct subjective effort estimations? How do we combine estimates from individuals?

Study: An effort estimation study conducted within the context of the PSP.

Outcome: The method of combination is more important than the way the subjective estimation is done.

Students vs. professionals

Question: Are the results from studies using students and professional the same?

Study: A study of different aspects influence on development time was conducted with students and professionals.

Outcome: The results are surprisingly similar, which indicate that in some cases (at least) using students as subjects do make sense.