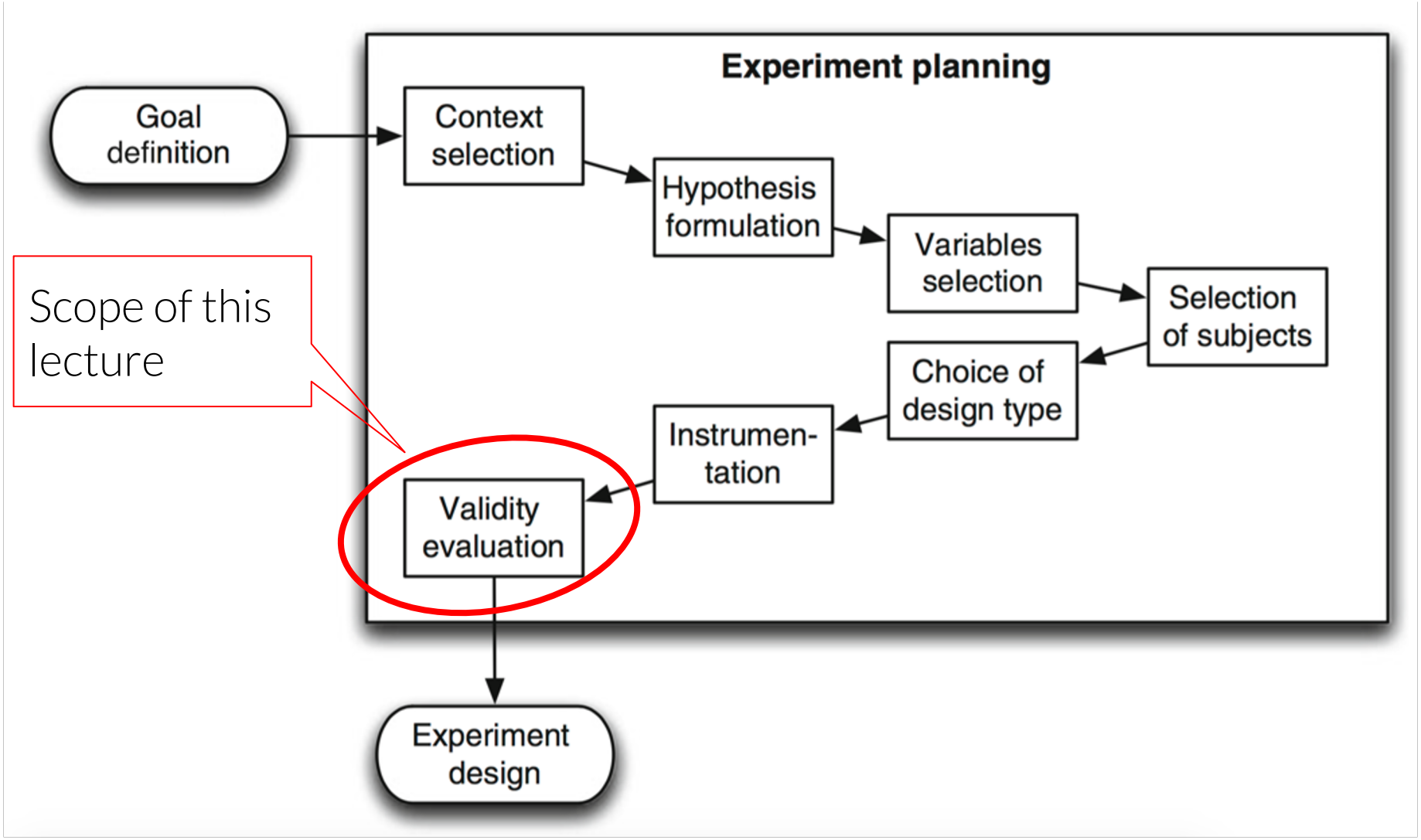
Experiment validity

Ivano Malavolta

VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Planning phases



Experiment validity

Validity is the extent to which our results are SOUND and

APPLICABLE TO THE REAL WORLD

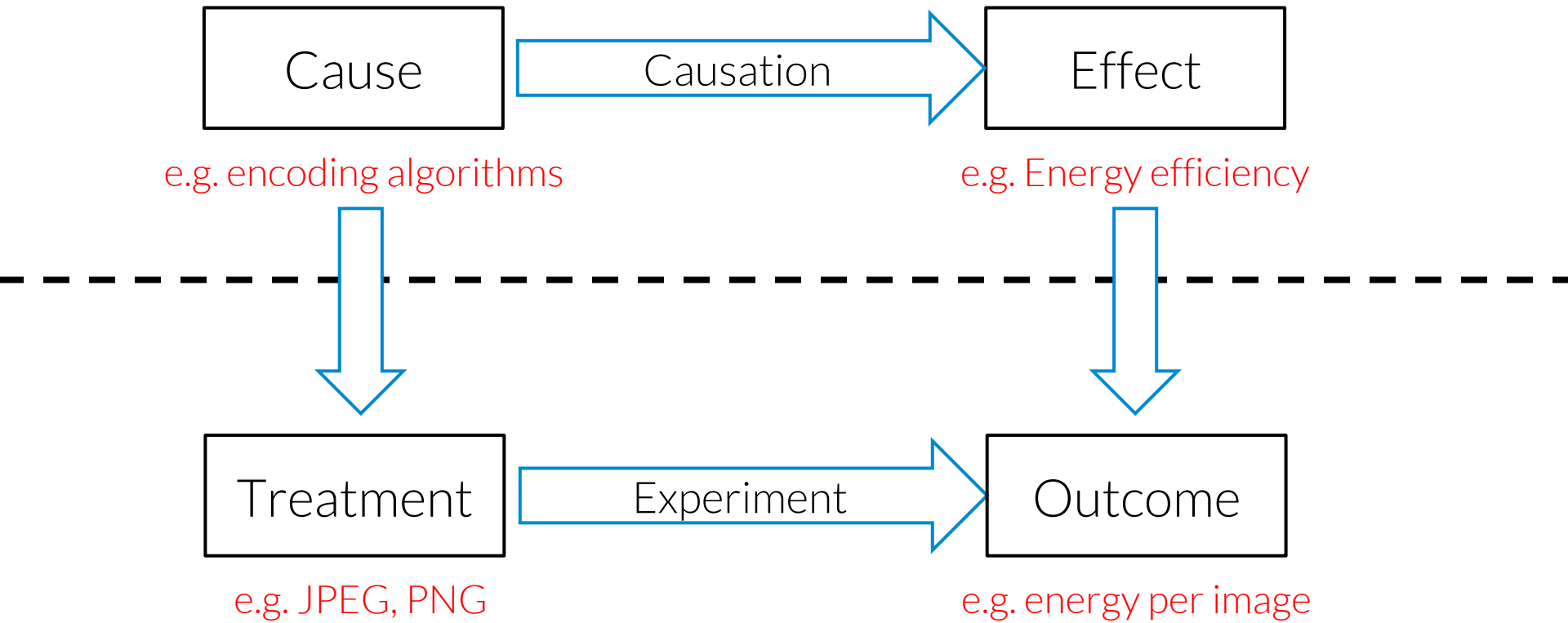
- We aim for **adequate** validity, not **universal** validity
  - What matters is our population of interest
- Validity is in *trade-off* with experiment scope

Threats Identification

- Identifying **threats** helps to plan for adequate validity
- Each threat needs appropriate **mitigation**
- Several classifications of validity threats:
  - Campbell and Stanley [1]
  - Cook and Campbell [2]

Types of threat to validity

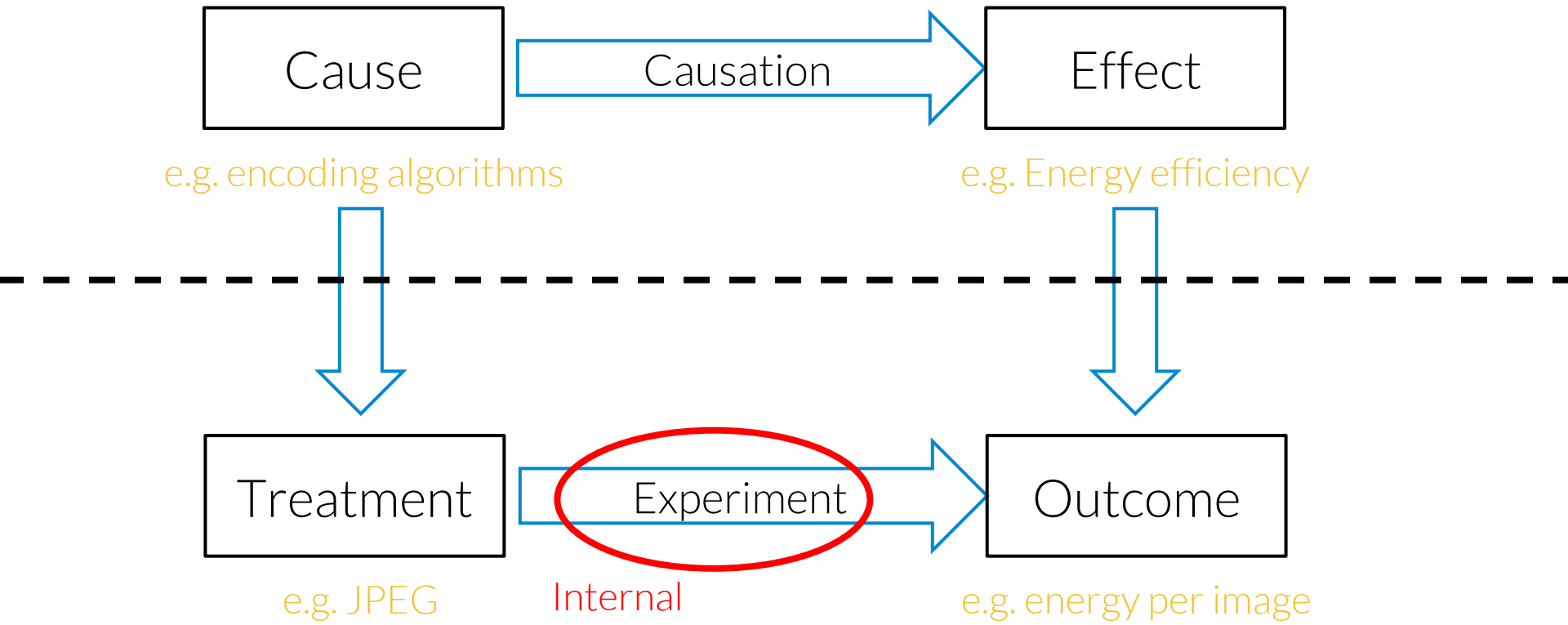
Theory



Observation

Types of threat to validity

Theory



Observation

Internal validity

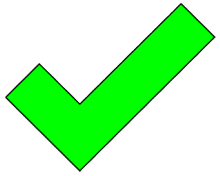
Internal Validity: causality between treatment and outcome

- Strongly related to the **experiment design and operation**
  - Are my results caused by the treatment?
  - Is my experimental environment clean enough?

Internal validity: types of threat

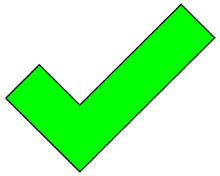
- History
  - Different trials of the experiment performed in different time frames (eg, after holidays vs normal days)
- Maturation
  - Subjects may react differently over time (eg, learning effect, tiresome, boredome)
- Selection
  - Some subjects may abandon the experiment
  - Even worse, some specific type of subjects may leave it
- Reliability of measures
  - If you repeat the measurement you should get similar results → same conclusions

Internal validity: mitigation



Analyze and identify confounding factors/noise

Choose appropriate experiment design



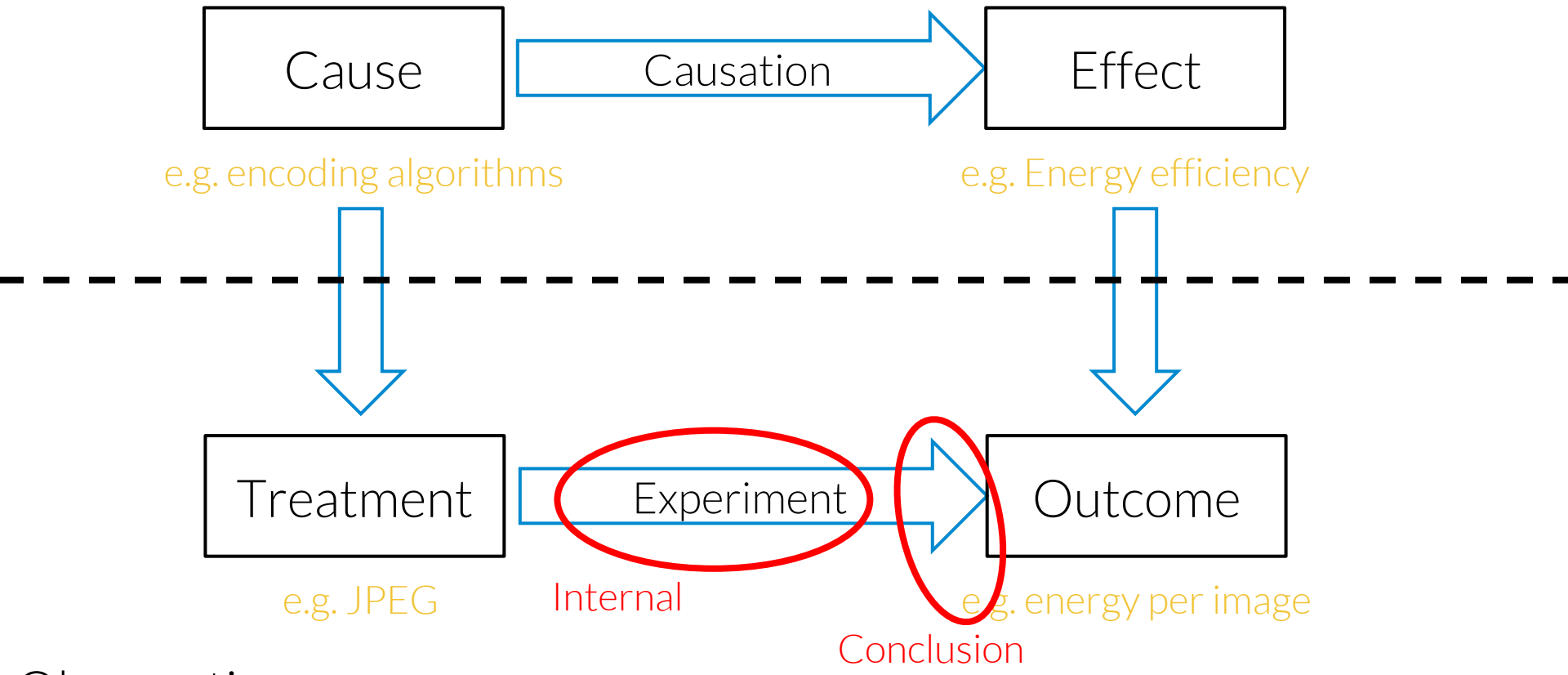
Keep environment under control

Define representative usage scenarios (if needed)

Ensure that your measures are reliable and correct

Types of threat to validity

Theory



Observation

Conclusion validity

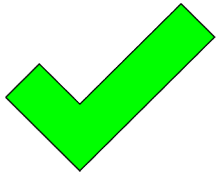
Conclusion Validity: statistical correctness and significance

- Are my conclusions correct?
- Are my results significant enough?

Conclusion validity: types of threat

- Low statistical power
  - Results not statistically significant
  - There is a significant difference but the statistical test does not reveal it due to the low number of data points
- Violated assumptions of statistical tests
  - eg, many tests assume normally distributed samples
- Fishing and error rate
  - If you are combining multiple statistical tests, also their significance should be adapted (Bonferroni, etc.)

Conclusion validity: mitigation

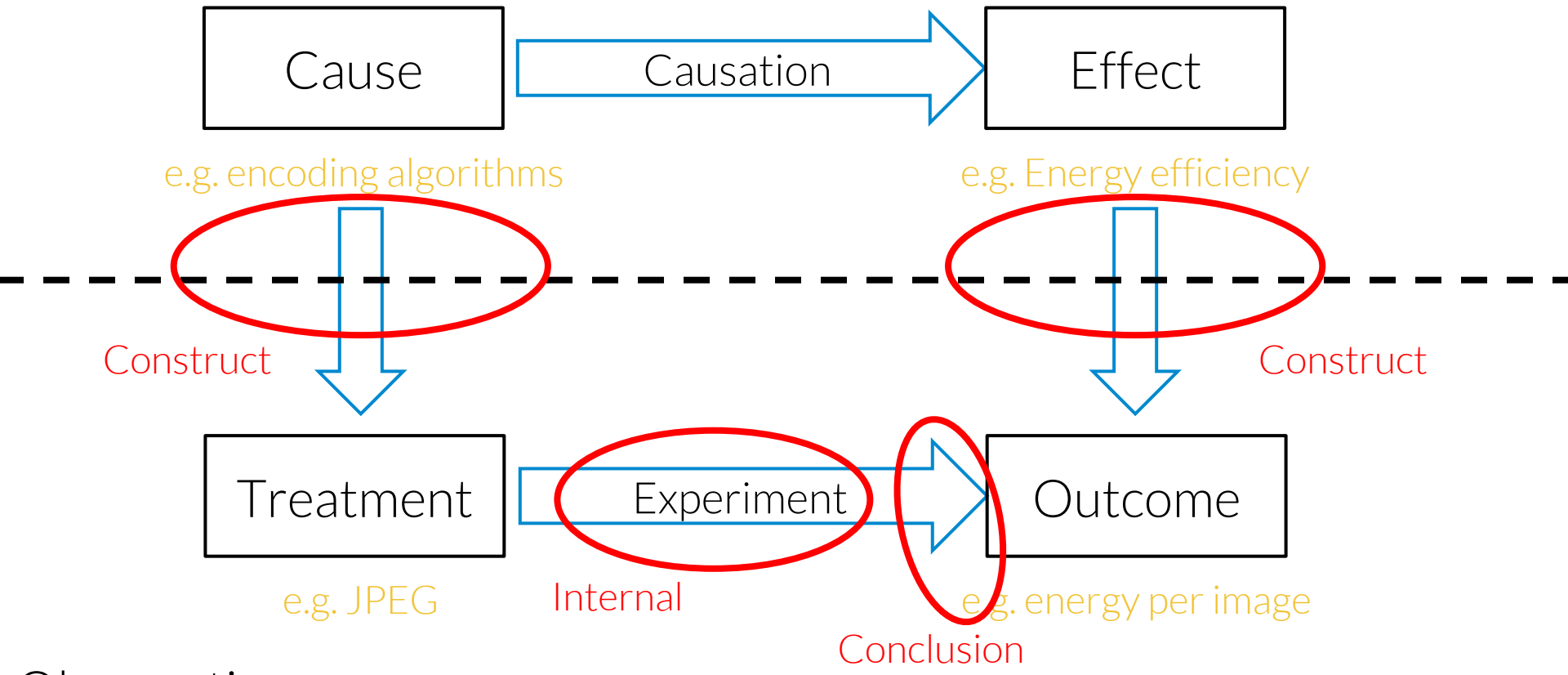


Select appropriate tests

Aim for high levels of statistical power

Types of threat to validity

Theory



Observation

Construct validity

Construct Validity: relation between theory and observation

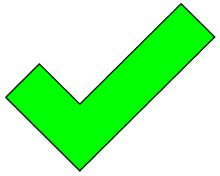
- Have I defined my constructs properly?
- Am I analyzing the correct variables for the effects?

Construct validity: types of threat

- Inadequate preoperational explication of constructs
  - construct not well defined before being translated into measures
  - Theory unclear
  - Comparing two methods, but not clear what does it mean that a method is better than another
- Mono-operation bias
  - I have one independent variable only, one single object or treatment

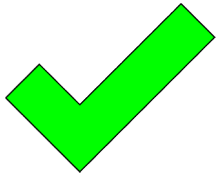
    → the experiment could not represent the theory
- Mono-method bias
  - When you use a single type of measures or observations
  - The experimenter may bias the measures

Construct validity: mitigation



Early definition of constructs (GQM)

Use appropriate experiment design

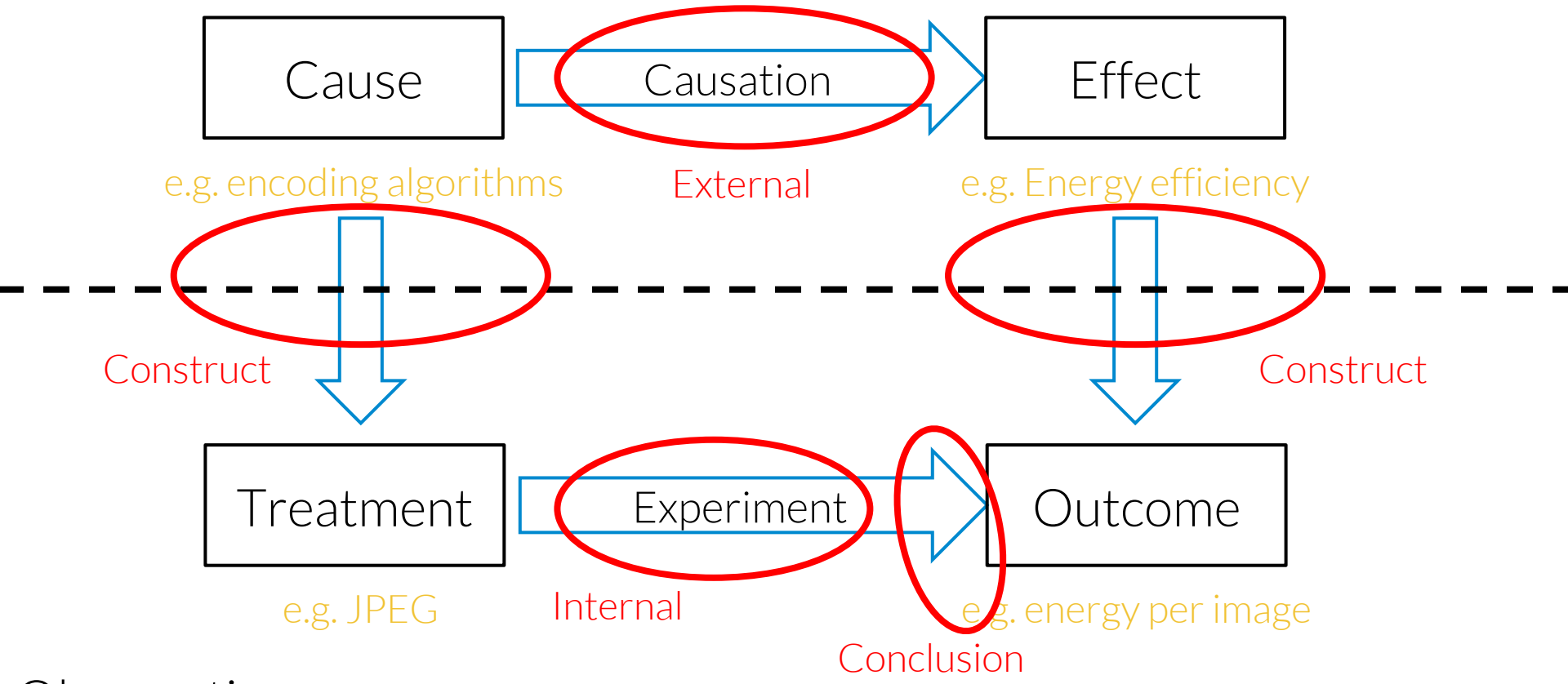


Justify your choices for factors and treatments

Introduce redundancy for cross-checks

Types of threat to validity

Theory



Observation

External validity

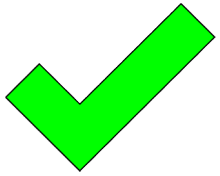
External Validity: generalizability of the results

- Are my results valid for the whole target population?
- Have I selected a representative sample?

External validity: types of threat

- Interaction of selection and treatment
  - the population of subjects **is not representative** of the one for which I would like to generalize my results
  - eg, performing experiments with toy/synthetic apps
- Interaction of setting and treatment
  - the experimental setting or the material are not representative
  - e.g. I let the subjects using tools that they don't use in the reality
  - e.g. Web development using textual editors
- Interaction of history and treatment
  - the experiment is conducted on a special time or day which affects the results
  - eg, our experiment on green software is performed after a big congress at which some subjects participated

External validity: mitigation



Use an environment as realistic as possible

Explicitly define and model your context

What this lecture means to you?

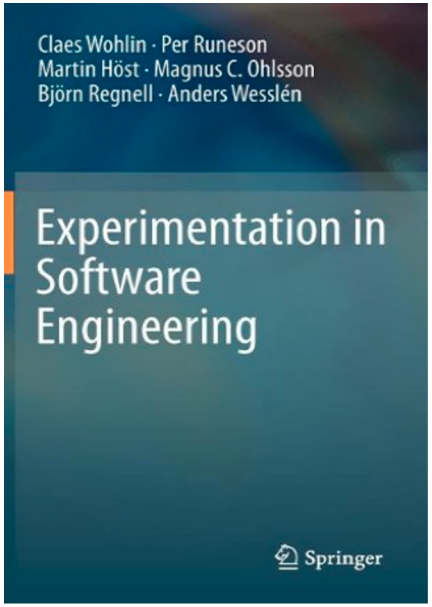
- You know that you have to **explicitly take into account** the threats to validity of your experiment
- Discussing threats actually makes your experiment stronger

  you are not showing your weaknesses, but you are improving the replicability of your study

- You will make **tradeoffs** between threats to validity in your experiment
- Consider threats to validity **as early as possible**

  Reasoning on them will make you feel more confident about the scope and design of your experiment

Readings



Chapter 8

[1] Campbell and Stanley, *Experimental and Quasi- Experimental designs for Research (1963).* (Blackboard)

[2] Cook and Campbell, *Quasi-experimentation - Design and Analysis Issues for Field Settings (1979).* Available at the VU library.

Acknowledgements

Some contents of lecture extracted from:

- Giuseppe Procaccianti's lectures at VU