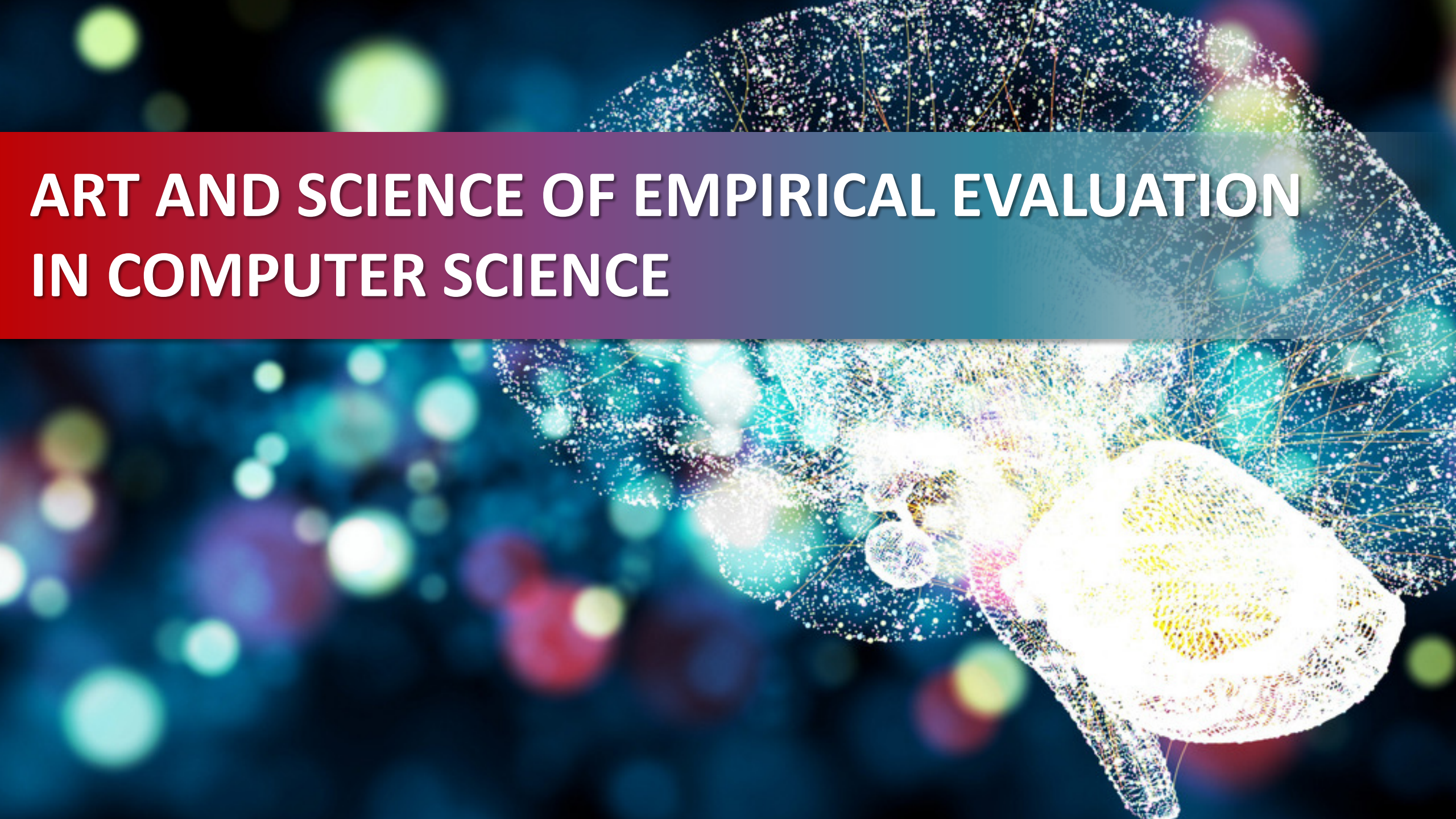
ART AND SCIENCE OF EMPIRICAL EVALUATION IN COMPUTER SCIENCE



EMPIRICAL STUDIES

Empirical studies are the collection and analysis of primary data based on direct observation or experiences in the “field”. Fields are, e.g.

- Medicine
- Psychology
- Pharmacy
- Sociology
- Physics
- Business/Marketing
- … Computer Science?

CONTENT

- EMPIRICAL STUDIES – **WHY?**
- EMPIRICAL STUDIES – WHERE AND WHEN?
- EMPIRICAL STUDIES – HOW?

COVID-19 VACCINE DEVELOPMENT

SEARCH FOR EVIDENCE

- Acceptance of technologies = f(degree of evidence provided).
- Evidence concerned with determining what works, when and where, for whom, and how much.
- And: what does NOT work and why.
- Ongoing effort, applicable to all techniques, processes and tools in computer science and software engineering!
- Publication of results requires proof of evidence!

Cybersquare

Walter F. Tichy
University of Karlsruhe

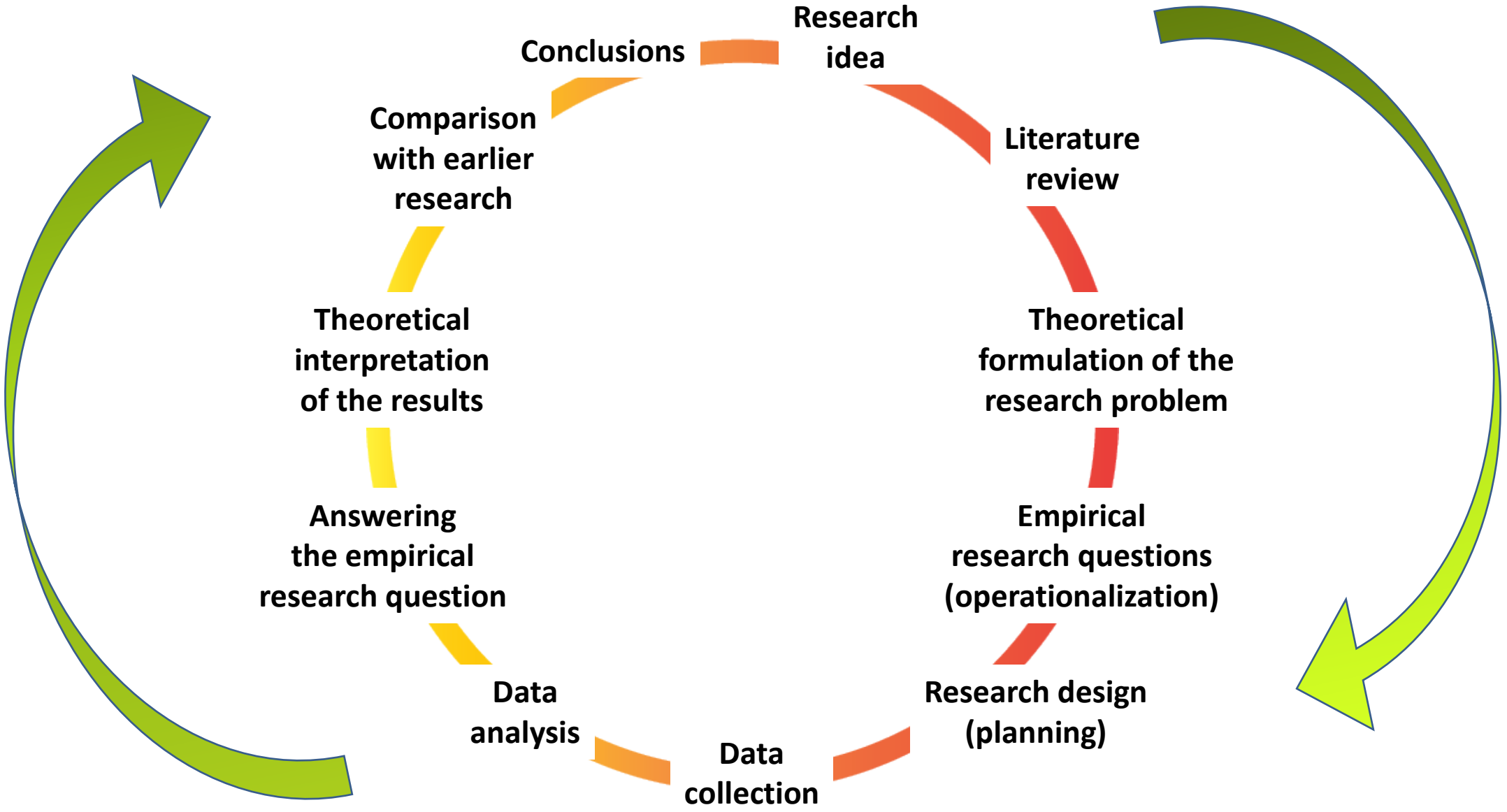
Should Computer Scientists Experiment More?

Computer scientists and practitioners defend their lack of experimentation with a wide range of arguments. Some arguments suggest that experimentation is inappropriate, too difficult, useless, and even harmful. This article discusses several such arguments to illustrate the importance of experimentation for computer science.

CONTENT

- EMPIRICAL STUDIES – WHY?
- EMPIRICAL STUDIES – **WHERE AND WHEN?**
- EMPIRICAL STUDIES – HOW?

EMPIRICAL EVALUATION AS PART OF THE RESEARCH PROCESS



RESEARCH QUESTIONS [2]

- **Occurrence**
  - How often does X occur?
  - What is the average amount of occurrence?
  - How does X normally work?
  - How much of X is enough?
- **Relationship**
  - Are X and Y related/correlated?
- **Causality**
  - Does X cause Y?
  - Does X prevent Y?
  - What are all the factors that cause Y?
  - Does X cause more X than does Z?
- **Design**
  - What is an effective way to achieve X?

EXAMPLE DIRECTIONS FOR CREATING EVIDENCE

- Technology evaluation (effort, time, performance, usability)
- Find strength and limitations of tools, methods, and techniques
- Find preferred areas of applications of technologies
- Comparison between technologies
- Effectiveness and efficiency of technologies
- Impact and ROI analysis

LEARNING FROM PHARMACEUTICALS*

- Contra-indications: Conditions under which a specific technology is not recommended
- Precautions: Warnings about what to be careful about while prescribing to the method
- Overdoses: What happens when you apply too much of something?
- Adverse reactions: Unintended negative side-effects

*Alan Davies: Contraindications, Precautions, Overdoses, and Adverse Reactions: What Software Engineering Can Learn from Pharmaceuticals http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4591600&tag=1

CREATING EVIDENCE: MANY METHODS AVAILABLE

- Controlled Experiments
- Case Studies
- Surveys
- Interviews
- Artifact/Archive Analysis
- Action Research
- Simulations

CONTENT

- EMPIRICAL STUDIES – WHY?
- EMPIRICAL STUDIES – WHERE AND WHEN?
- EMPIRICAL STUDIES – **HOW?**

- Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K., Rosenberg, J.:
- Preliminary guidelines for empirical research in software engineering (2002) IEEE Transactions on Software Engineering, 28 (8), pp. 721-734

GUIDELINES – EXPERIMENTAL CONTEXT

- C1: Be sure to specify as much of the industrial context as possible. In particular, clearly define the entities, attributes and measures that are capturing the contextual information.
- C2: If a specific hypothesis is being tested, state it clearly prior to performing the study, and discuss the theory from which it is derived, so that its implications are apparent.
- C3: If the research is exploratory, state clearly and, prior to data analysis, what questions the investigation is intended to address, and how it will address them.
- C4: Describe research that is similar to, or has a bearing on, the current research and how current work relates to it.

GUIDELINES – EXPERIMENTAL DESIGN

D1: Identify the population from which the subjects and objects are drawn.

D2: Define the process by which the subjects and objects were selected.

D3: Define the process by which subjects and objects are assigned to treatments.

D4: Define the experimental unit.

GUIDELINES – DATA COLLECTION 1

- DC1: Define all software measures fully, including the entity, attribute, unit and counting rules.
- DC2: For subjective measures, present a measure of inter-rater agreement, such as the kappa statistic or the intra-class correlation coefficient for continuous measures.
- DC3: Describe any quality control method used to ensure completeness and accuracy of data collection.

GUIDELINES – DATA COLLECTION 2

- DC4: For surveys, monitor and report the response rate, and discuss the representativeness of the responses and the impact of non-response.
- DC5: For observational studies and experiments, record data about subjects who drop out from the studies.
- DC6: For observational studies and experiments, record data about other performance measures that may be adversely affected by the treatment, even if they are not the main focus of the study.

GUIDELINES – ANALYSIS

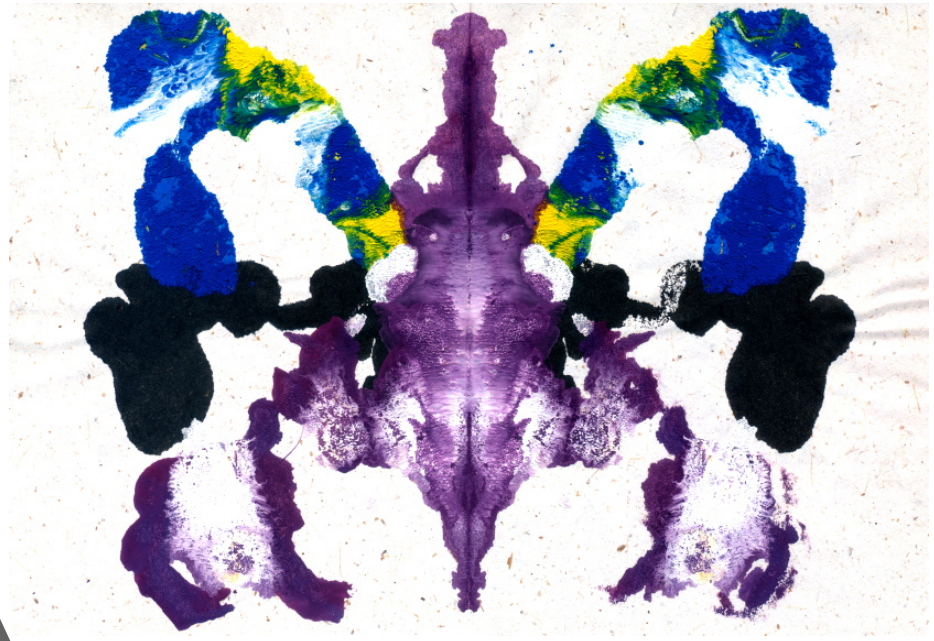
- A1: Specify any procedures used to control for multiple testing.
- A2: Consider using blind analysis.
- A3: Perform sensitivity analyses.
- A4: Ensure that the data do not violate the assumptions of the tests used on them.
- A5: Apply appropriate quality control procedures to verify your results.

GUIDELINES – PRESENTATION

- P1: Describe or cite a reference for all statistical procedures used.
- P2: Report the statistical package used.
- P3: Present quantitative results as well as significance levels.
- P4: Present the raw data whenever possible.
- P5: Provide appropriate descriptive statistics.
- P6: Make appropriate use of graphics.

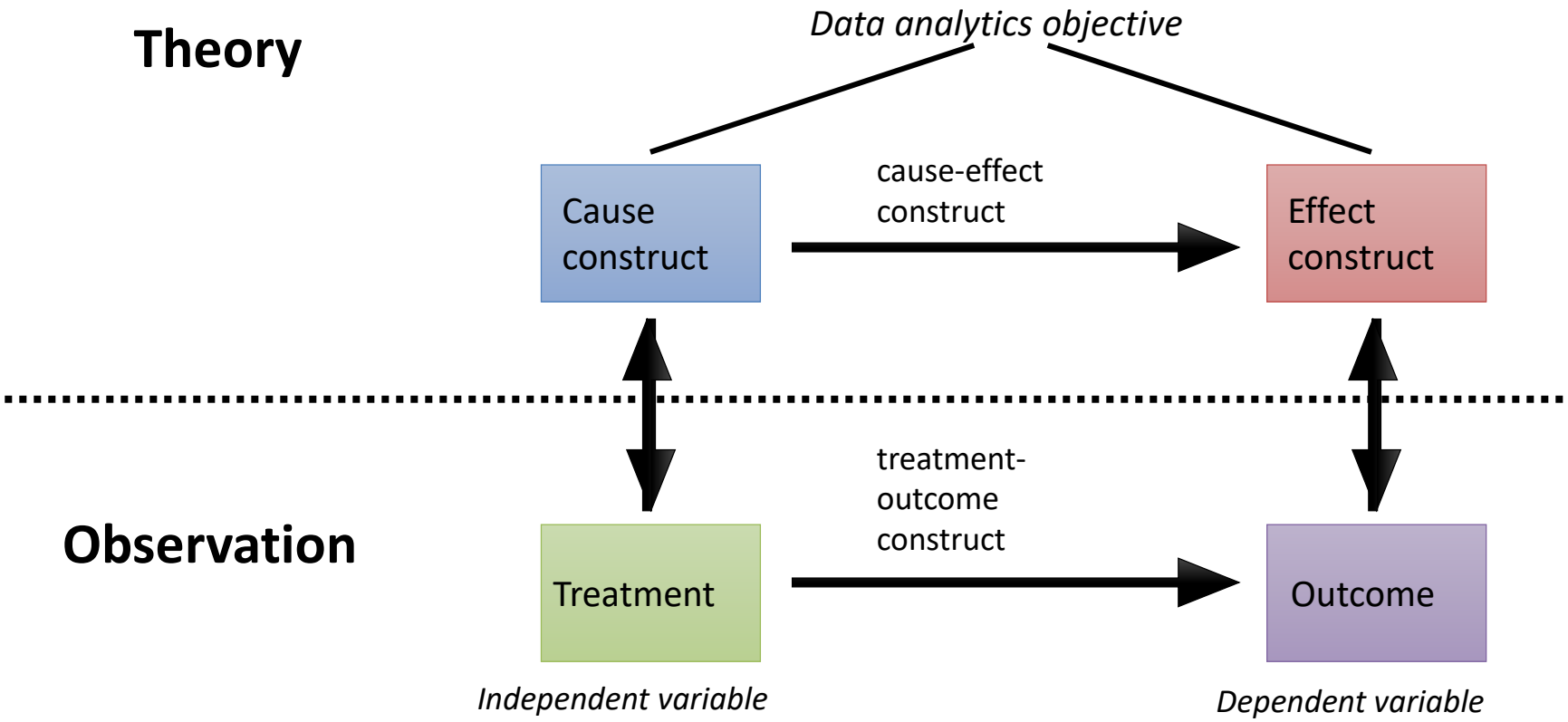
GUIDELINES – INTERPRETATION

- I1: Define the population to which inferential statistics and predictive models apply.
- I2: Differentiate between statistical significance and practical importance.
- I3: Define the type of study.
- I4: Specify any limitations of the study.



http://silvieandmaryl.com/2010/07/dream-interpretation

THREATS TO VALIDITY



Sources: Experimentation in Software Engineering, Wohlin et al. (2000)

MAIN TYPES OF VALIDITY THREATS DISCUSSED IN THE LITERATURE [4]

Validity threat type	Example of typical questions to be answered
Conclusion validity	Does the treatment/change we introduced have statistically significant effect on the outcome we measure?
Internal validity	Did the treatment/change we introduced cause the effect on the outcome? Can other factors also have had an effect?
Construct validity	Does the treatment correspond to the actual cause we are interested in? Does the outcome correspond to the effect we are interested in?
External validity, Transferability	In the cause and effect relationship we have shown valid in other situations? Can we generalize our results? Do the results apply in other contexts?
Credibility	Are we confident that the findings are true? Why?
Dependability	Are the findings consistent? Can they be repeated?
Confirmability	Are the findings shaped by the respondents and not by the research?

REFERENCES

[1] Tichy, W.F., 1998. Should computer scientists experiment more? Computer, 31(5), pp.32-40

[2] Shaw, M.: Writing good Software Engineering Research Papers. In 25th International Conference on Software Engineering, 2003. Proceedings, pp. 726-736

[3] Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K. and Rosenberg, J., 2002. Preliminary guidelines for empirical research in software engineering. IEEE Transactions on Software Engineering, 28(8), pp. 721-734

[4] Feldt, R. and Magazinius, A. Validity threats in empirical software engineering research - An initial survey. In Proceedings SEKE 2010, pp. 374-379

Questions

CONTACT: GUENTHER RUHE

PHONE: 1.403.220.7692

EMAIL: ruhe@ucalgary.ca