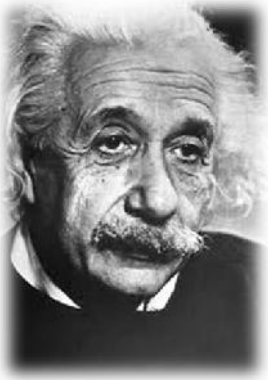

In God we trust.
Everyone else,
bring data.
Michael Bloomberg, NYC Mayor
DTU
Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet

Alternative Ways to Convince People

By force

physical force, group pressure



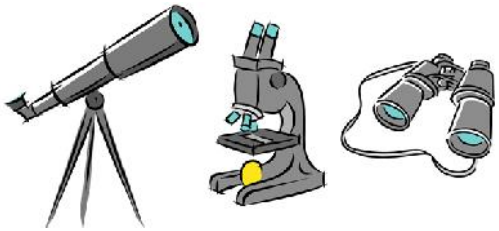
By authority

divine revelation, fame

By insight

plausibility, observation

By research



Impact on Research (& Teaching)

Early Stages
Mathematics and Electrical Engineering dominate Computer Science. "Software Engineering" is a contradiction in terms.

Contemporary SE
SE is scientifically accepted, but lacks industrial impact. Most work is conceptual, maybe implemented, but rarely validated empirically.

Future of SE
Practical work without empirical validation will be scientifically unacceptable. Industrial impact will grow due to solid evidence to support our claims.

MODELS Call for Papers

One of three categories of research papers
"Papers evaluating existing problem cases or scientifically validating proposed solutions through, for example, empirical studies, experiments, case studies, simulations, formal analyses, and mathematical proofs. [...] The research method must be sound and appropriate."

VL/HCC Call for Papers

"Research papers are expected to support their claims with appropriate evidence. [...] However, not all claims necessarily need to be supported with empirical evidence or studies with people. [...] Moreover, there are many alternatives to empirical evidence, [...] We encourage authors to think carefully about what claims their submission makes and what evidence would support them."

Overview

epistemic paradigm

deductive
→ logical proof

$\models \quad \forall$

$\Longrightarrow$

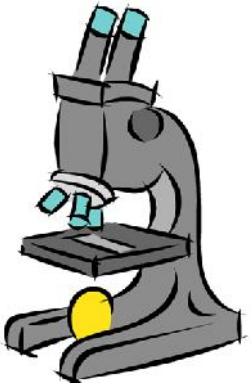
inductive
→ empirical evidence

secondary
→ Surveys

primary

qualitative
→ Field Work

quantitative
→ Experiments

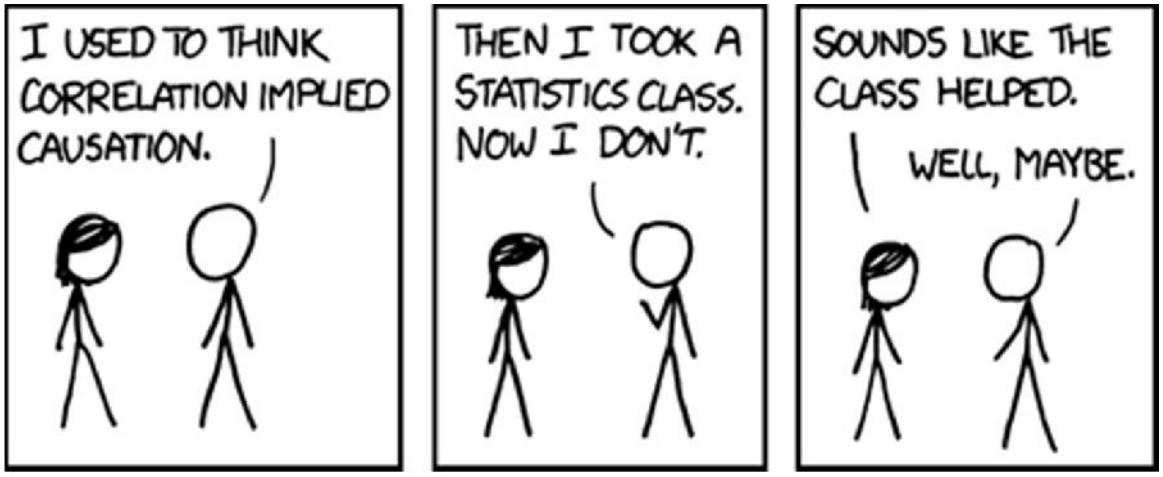


Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet (DTU)

EXPERIMENTS

Controlled Experiments

- **A controlled experiment (or just: experiment) is the "ideal type" of all (quantitative) empirical methods.**
  - Strictly speaking, all we ever get is correlation, though many people don't appreciate the difference (vaccination~autism)



- **Demanding tight control over the experimental variables, it is usually only applicable in lab situations (in vitro).**
  - In other situations (i.e., in the field, "in vivo") there is usually too much disturbance to obtain useful results.
  - However, sometimes the context is as important as the observation. Then, an experiment is pointless.

Anecdotes are not scientific evidence

The highly respected Professor Nibbowitz proved,
that octopus are more intelligent than cat,
when exposed to the same challenges and conditions.

Motivation

- **Several empirical studies have previously evaluated criteria of good diagram layout.**

- **Most of them produced inconclusive or weakly significant results.**

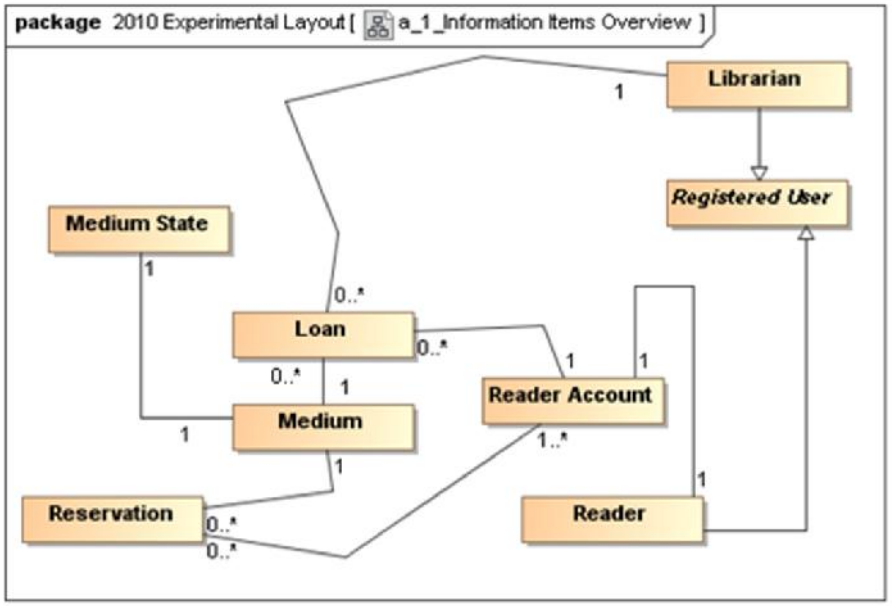
  „We could not identify a statistically significant relation between diagram quality and [understandability].“

  [Eichelberger & Schmid, J. Information & Software Technology 51 (2009) p. 1696]

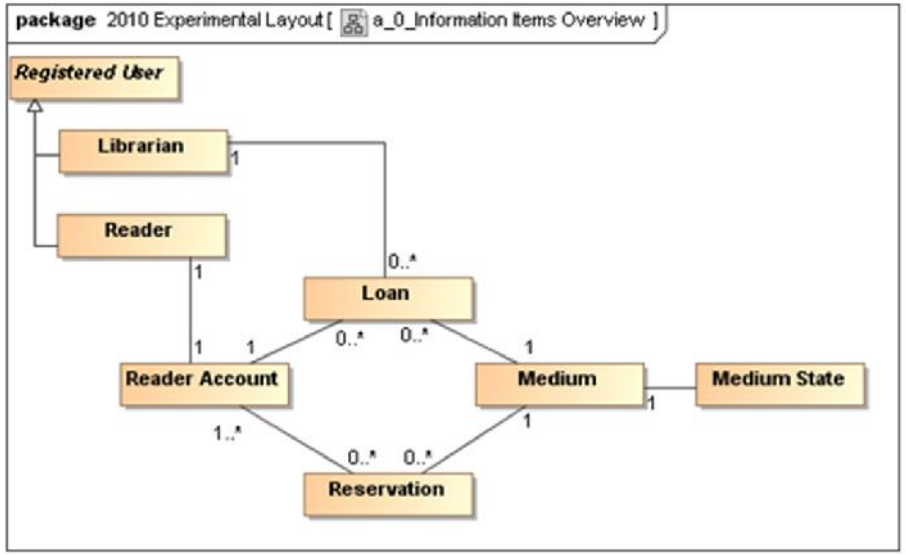
- **Limitations in previous studies motivate our research.**
  - They all studied causal effects of individual elementary layout heuristics such as „avoid line crossings and bends“ or „use consistent font types“, with a view to improve automatic diagram layout.
  - Most studies have very small numbers of participants (note the exceptions).
  - Many studies measured only objective performance, ignoring cognitive load.
  - Most of studies evaluate only simplistic/artificial UML class diagrams, some analyze interaction diagrams, but none Activity or Use Case Diagrams.

Good vs. Bad (UML Diagram) Layout

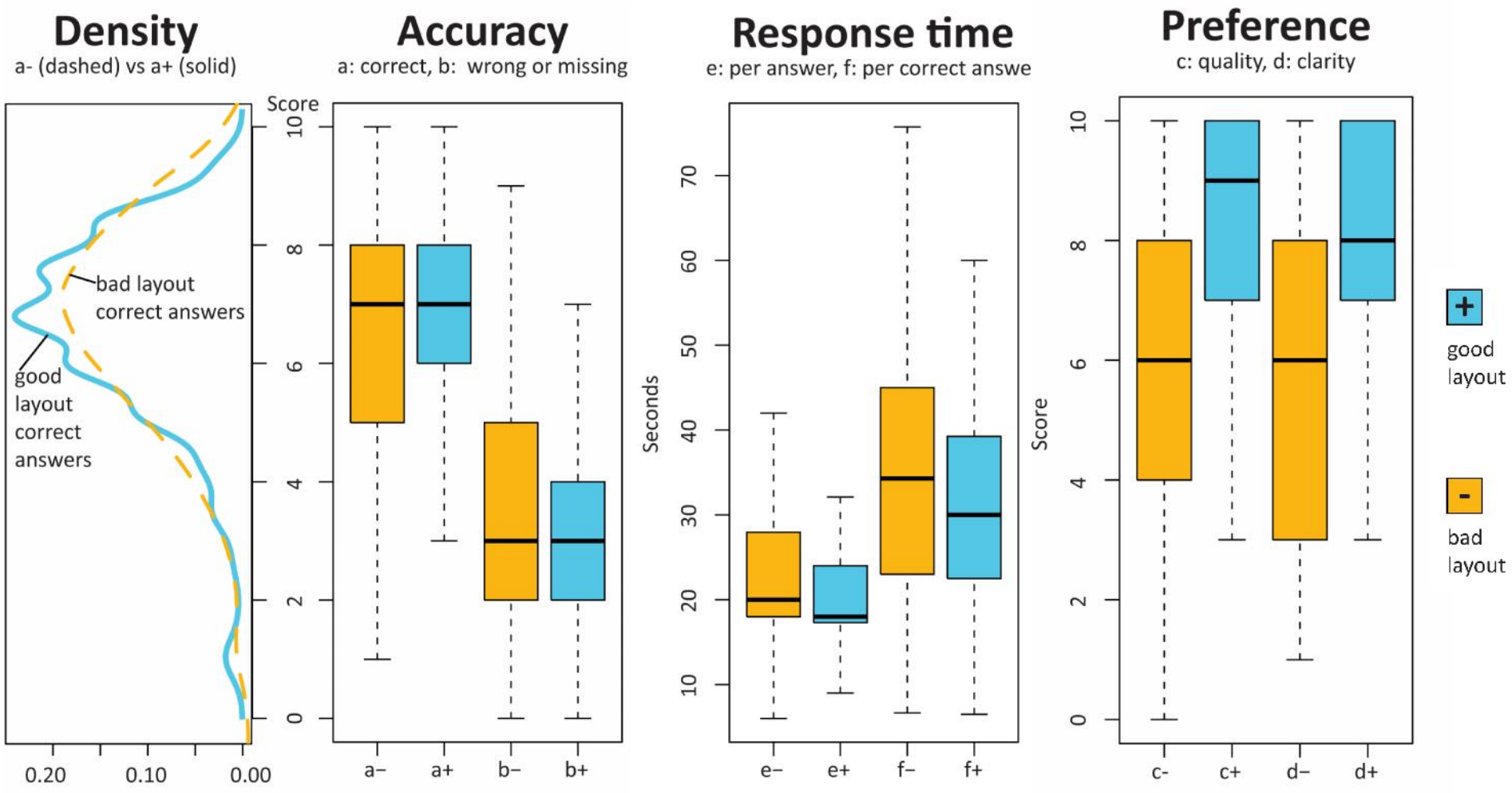
- **Elements of bad layout**
  - Edge crossings and bends
  - Overlaping/obscuring elements
  - Varying colors/sizes
  - Varying text orientation



- **Elements of good layout**
  - Join similar edges
  - Cluster similar elements
  - Orthogonal arrangement
  - Place elements in flow



Observations



Effect Size

- **In contrast to previous studies, we observe a comparatively large effect, though not necessarily in user performance.**

Accuracy (a, b)

answers	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
right	6.35	2.07	6.76	1.94	+6.5%
wrong/missing	3.65	2.07	3.24	1.94	-12.7%

Preference (c, d)

rating	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
diagram quality	5.54	2.74	8.06	2.12	+31.3%
diagram clarity	5.61	2.74	7.81	2.27	+28.2%

Response time (e, f)

s/answer	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
all answers	22.72	10.85	21.06	8.25	-7.3%
right answers	38.37	24.39	31.68	15.77	-17.4%

- **Cognitive load seems to benefit much more from good layout than objective performance indicators.**
  - This might be due to subjective coping strategies.
  - Dual stimulus experiments might shed light on this hypothesis.

Expertise ~ Impact

Initially, we found no noticeable differences between novice modelers and advanced modelers, to our surprise.

Maybe, our “experts” were no real experts?

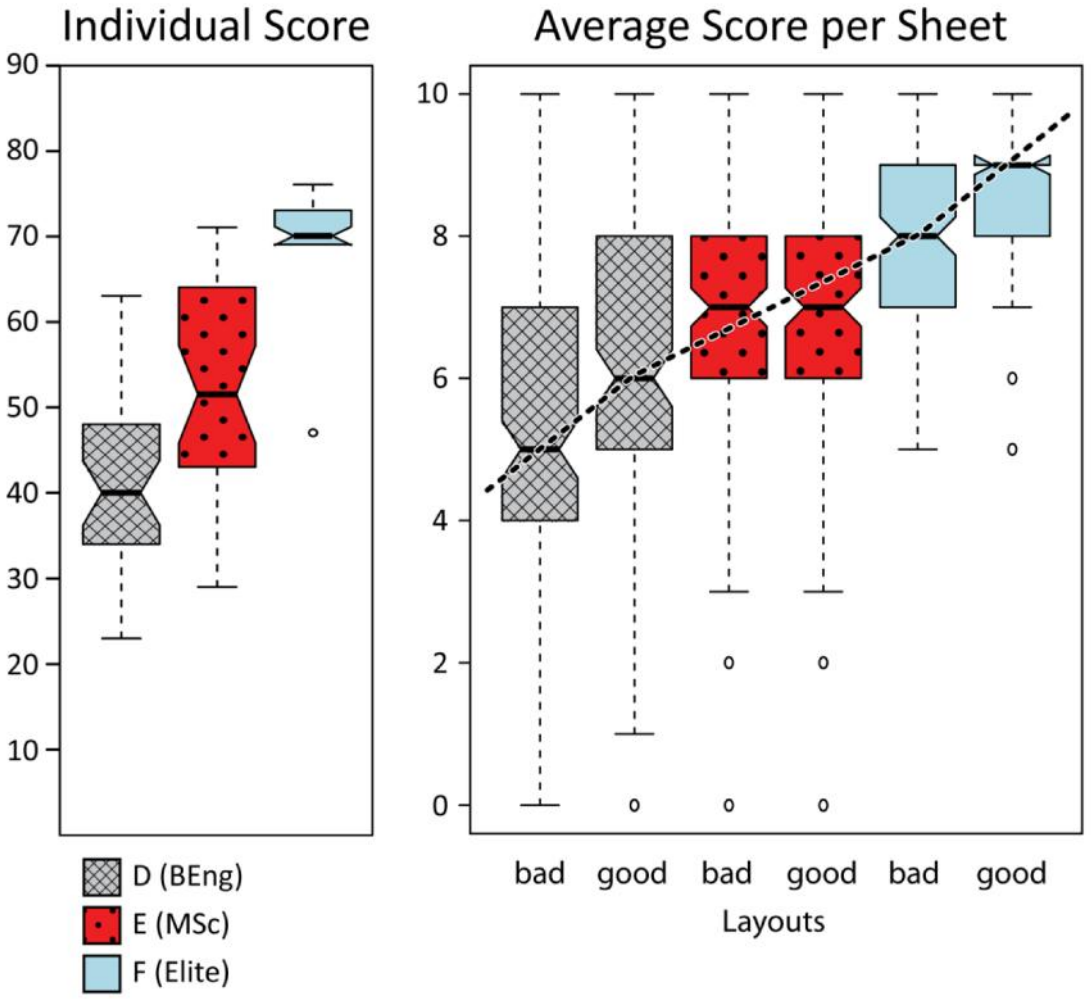
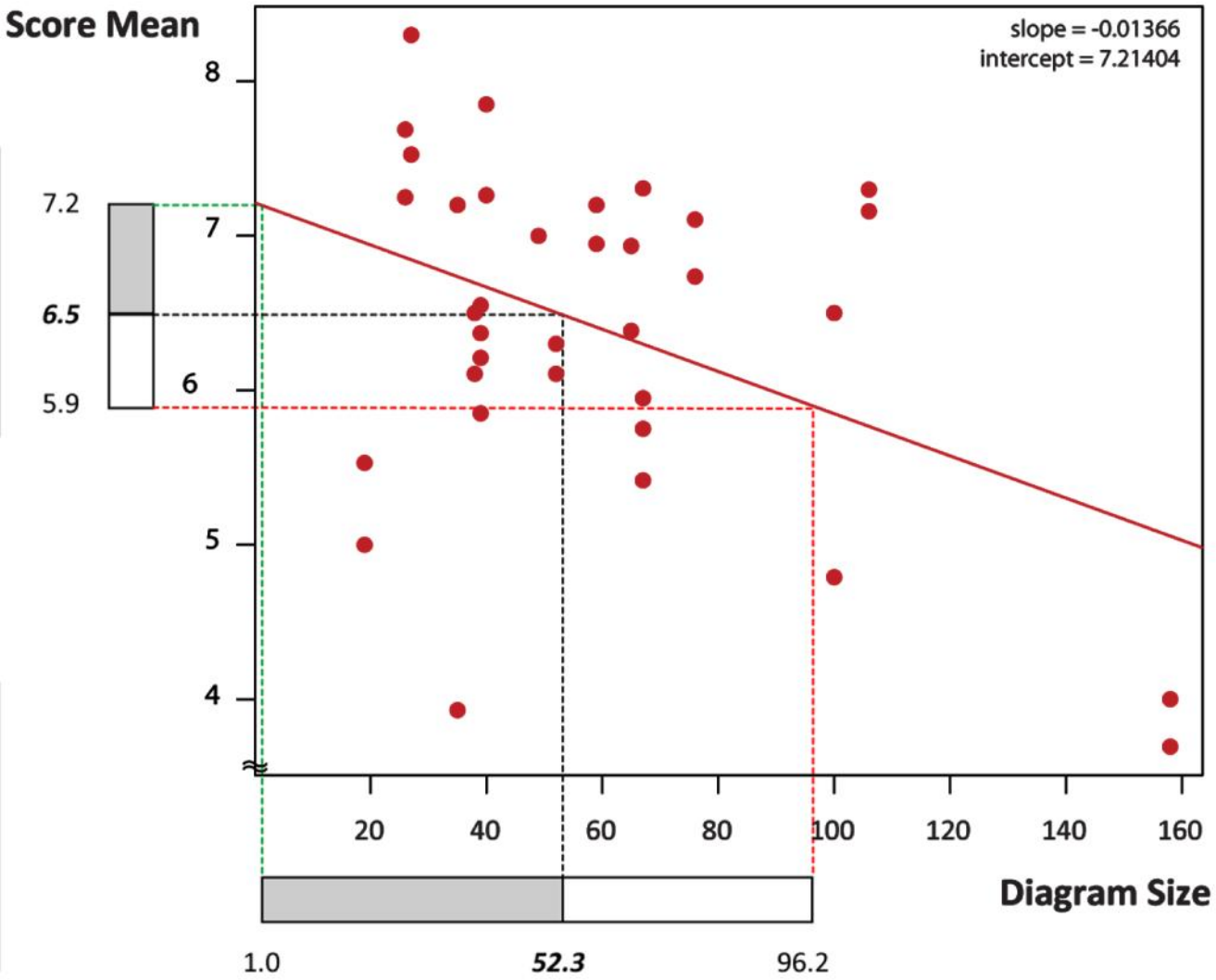


Diagram Size ~ Impact

Score	Q1	Median	Q3
Layouts: Good	6.2	6.9	7.2
Layouts: All	5.9	***6.5***	7.2
Layouts: Bad	5.5	6.4	7.2

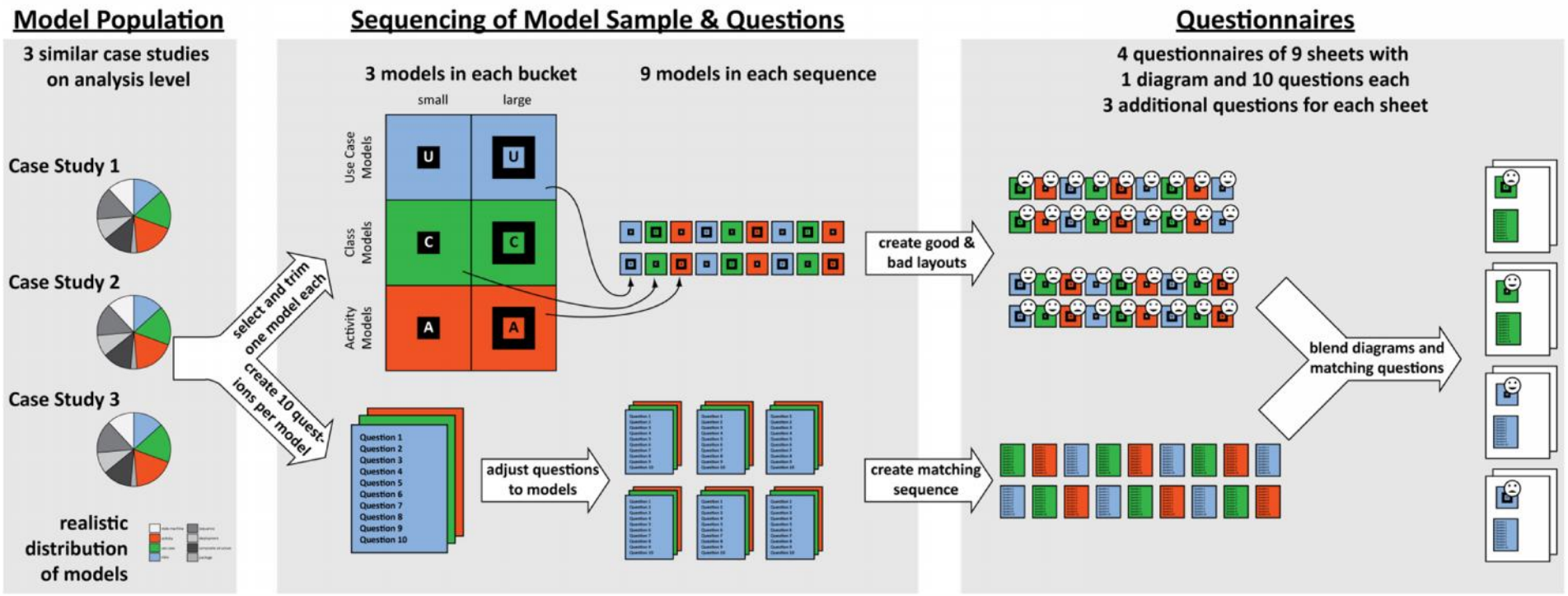
Size	Q1	Median	Q3
Layouts: Bad	1.0	23.0	74.2
Layouts: All	1.0	***52.3***	96.2
Layouts: Good	1.0	59.6	125.5



Validity of Method & Results

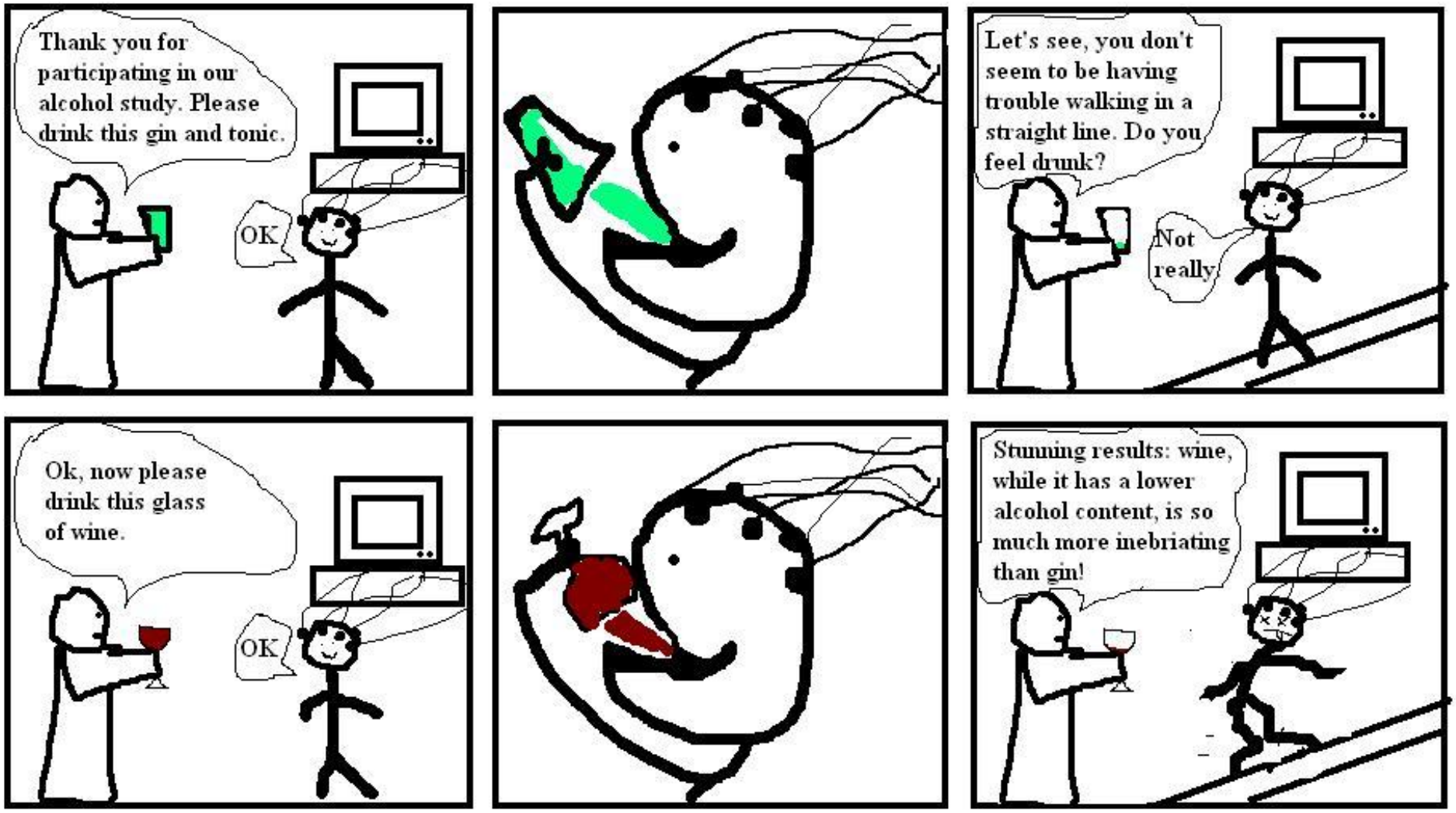
- **Online survey vs. Paper**
  - Online surveys may reach a larger audience, but often achieve poor completion (~20%), high noise, and little control of participants.
  - Also, paper is more realistic, as domain experts and decision makers will usually be faced with printed model reports rather than modeling tools.
- **Questionnaires vs. Eye tracking**
  - Maletic et al. have used eye-tracking in UML CD comprehension studies to validate questionnaire-based results yielding similar observations.
  - Even modern eye tracking equipment imposes substantial difficulty and effort, thus severely restricting the number of participants.
- **Cognitive load measures**
  - Subjective assessments have been shown to be as reliable as physiological indicators such as skin conductivity, heart rate, pupillary dilatation.
- **Internal Validity**
  - Low p-values control Type I errors.
  - High n controls Type II errors.
- **External Validity**
  - Case studies and models are realistic, but not real.
  - Number of models might be too small for general conclusions.
- **Construct Validity**
  - Measuring cognitive load leads to different conclusions and a higher degree of construct validity than previous work.
- **Conclusion Validity**
  - Consistent observations over multiple measures, measurements, models, tasks, and populations.
  - Consistent results over a series of 7 experiments.
  - No independent replication yet.

Study Design



- **Within-Subjects design reduces impact of individual variance.**
- **Models from different case studies reduce semantic inferences.**
- **Systematic variation of independent variables cancels out learning.**

Learning Effects



Validity of Method

- **We chose a rather conventional setup: tasks were presented as A4 sized printed pages containing layouts and tick-off-questions.**

- **Online survey vs. Paper**
  - Online surveys may reach a larger audience, but often achieve poor completion (~20%), high noise, and little control of participants.
  - Also, paper is more realistic, as domain experts and decision makers will usually be faced with printed model reports rather than modeling tools.
- **Questionnaires vs. Eye tracking**
  - Maletic et al. have used eye-tracking in UML CD comprehension studies to validate questionnaire-based results yielding similar observations.
  - Even modern eye tracking equipment imposes substantial difficulty and effort, thus severely restricting the number of participants.
- **Cognitive load measures**
  - Subjective assessments have been shown to be as reliable as physiological indicators such as skin conductivity, heart rate, pupillary dilatation.

Validity of Results

- **Internal Validity**
  - Low p-values control Type I errors.
  - High n controls Type II errors.
- **External Validity**
  - Case studies and models are realistic, but not real.
  - Number of models might be too small for general conclusions.
- **Construct Validity**
  - Measuring cognitive load leads to different conclusions and a higher degree of construct validity than previous work.
- **Conclusion Validity**
  - Consistent observations over multiple measures, measurements, models, tasks.

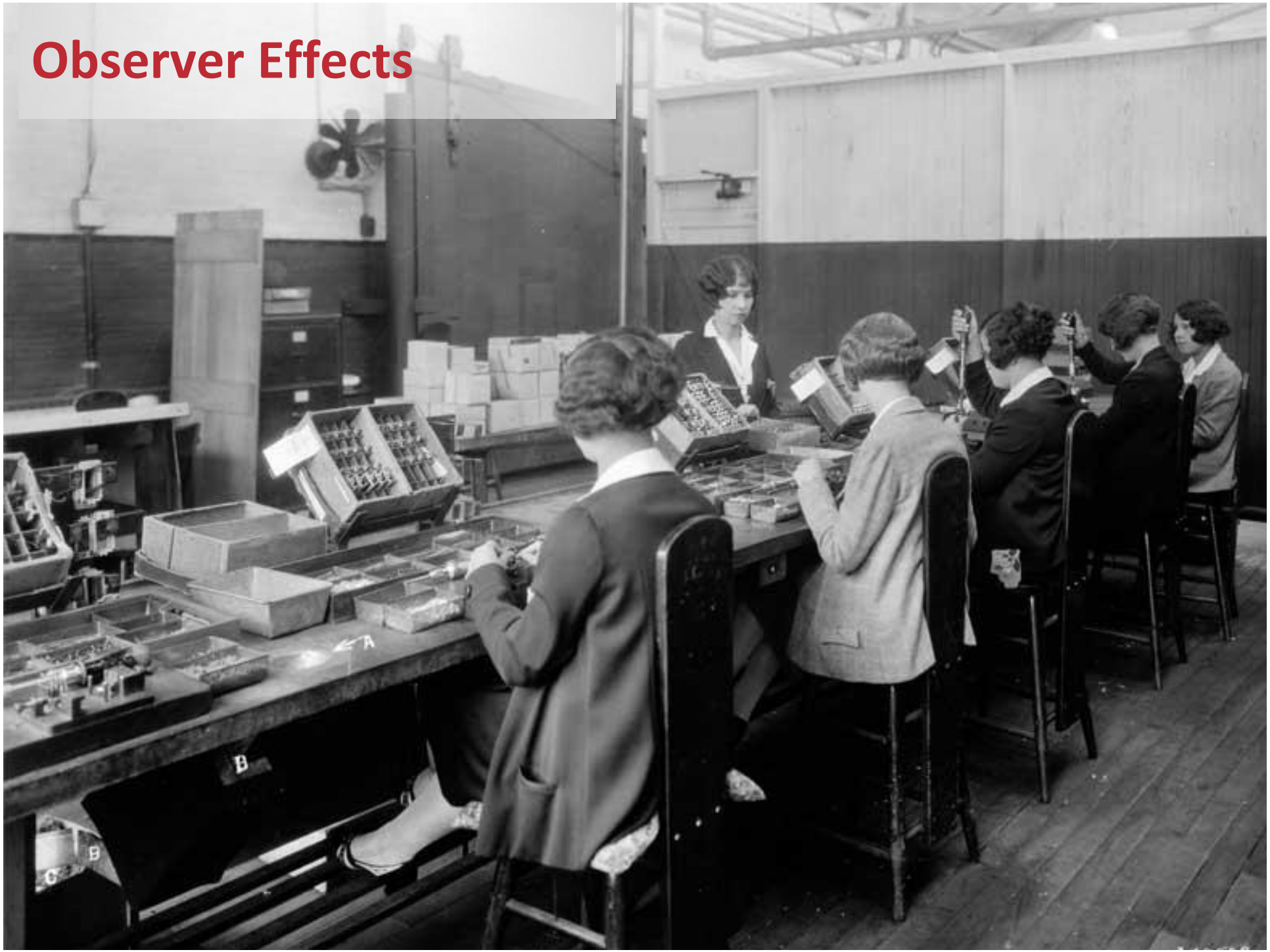
Participants

	male	female	all	completion rate (core questions)
novices	40	3	43	80.0 %
experts	30	4	34	84.4 %
all	70	7	77	81.9 %

Significance

HYPOTHESIS	P-VALUE	SIGNIFICANCE
$H_{0,1}$: same user performance for good/bad layouts wrt.		
...correct answers	0.003	**
...wrong answers	0.002	**
...time per answer	0.061	*
...time per correct answer	< 0.001	***
$H_{0,2}$:same user assessment of good/bad layouts wrt.		
...layout quality	$< 10^{-15}$	***
...diagram clarity	$< 10^{-15}$	***
$H_{0,3}$ same performance for good/bad layouts by experts/novices wrt.		
...correct answers	< 0.0001	***
...wrong answers	< 0.0001	***
$H_{0,4}$: novices benefit more than experts from good layouts		
...correct answers	0.39	-
...wrong answers	0.24	-

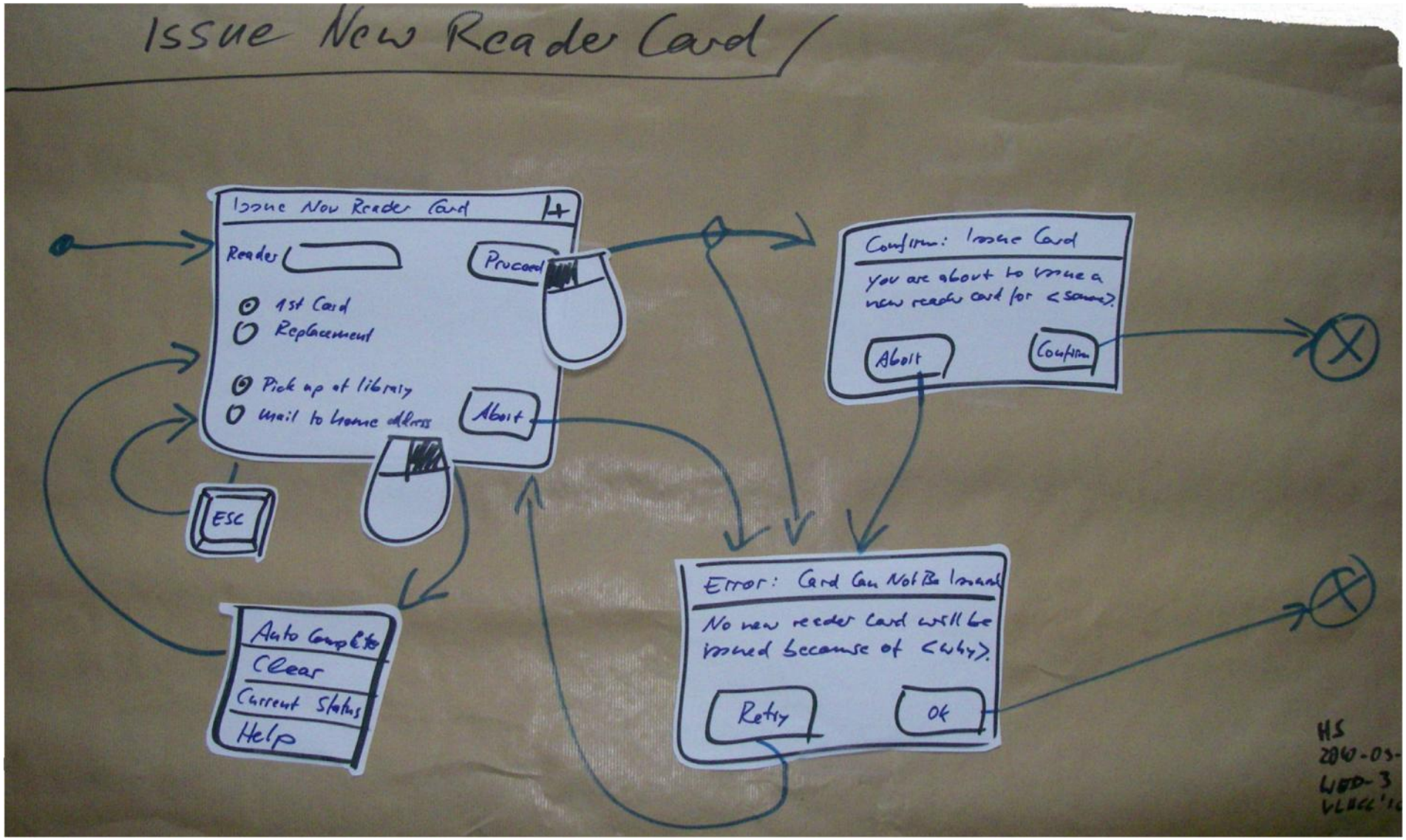
Observer Effects



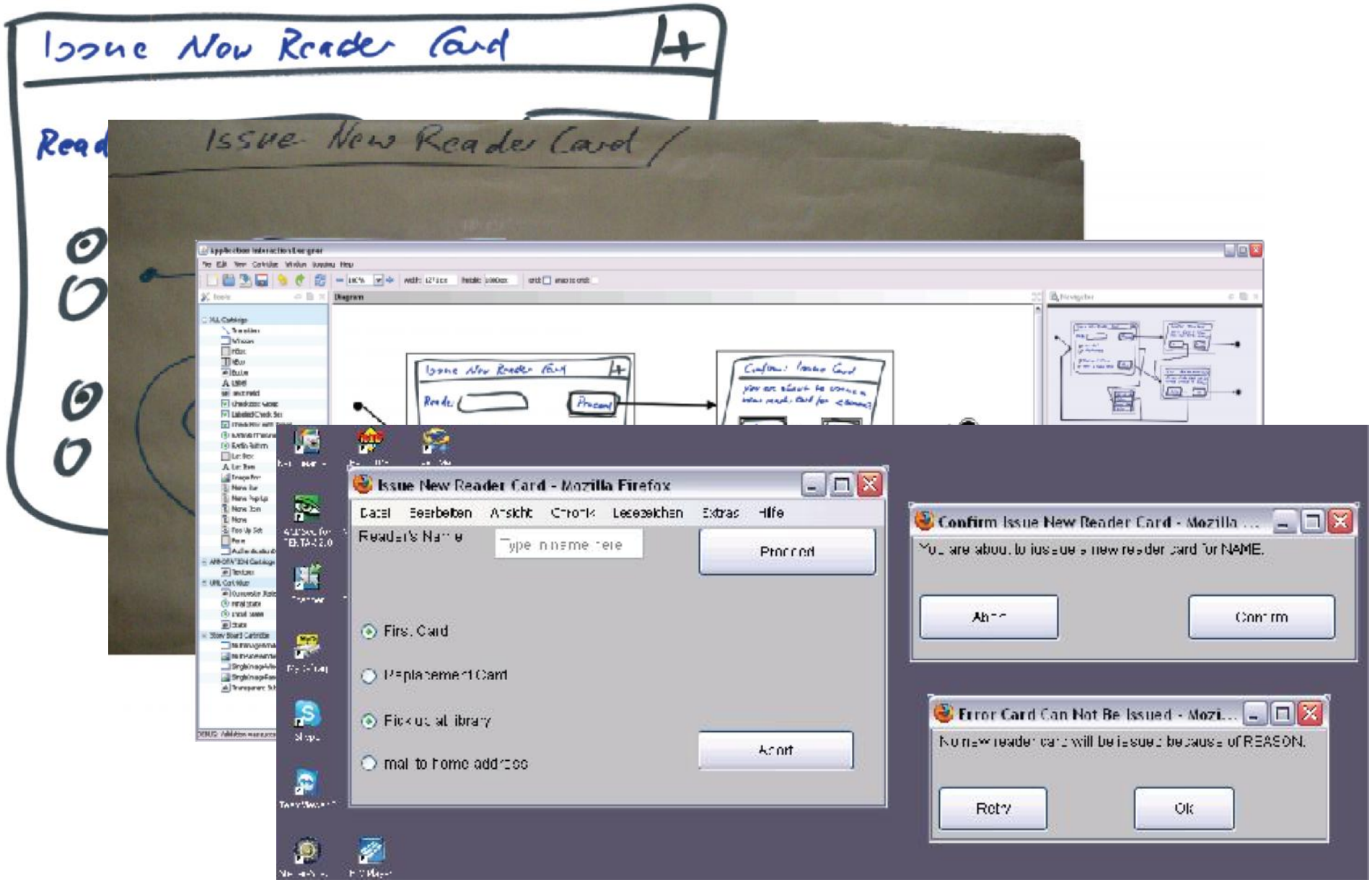
Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet (DTU)

FIELD WORK

Storyboards vs. UML StateMachines



Elaborating WEDs to prototypes



Medium ~ Group Work (1)

Facilitating Overview and Detail View
Allows concurrent activities

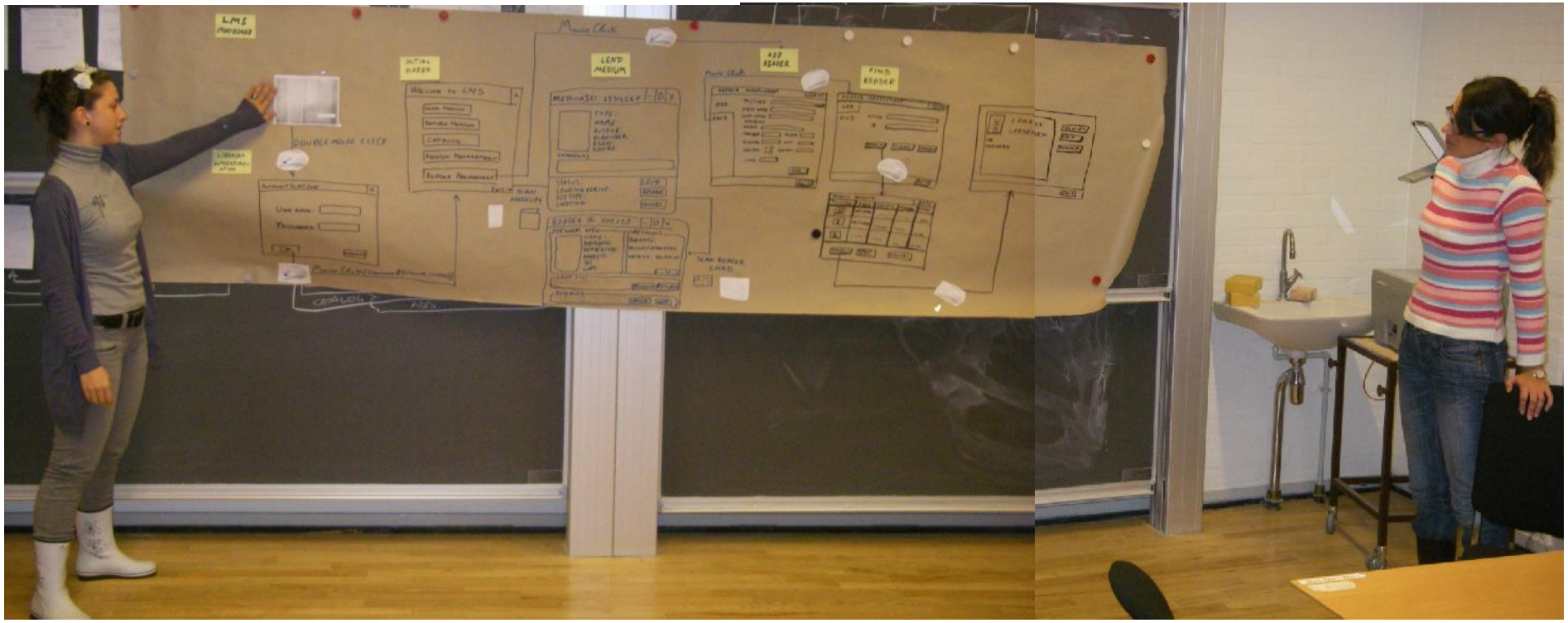
Maintain focus when changing detail level
Embodied Interaction and Manipulation

Medium ~ Group Work (2)

Division of labor
for independent tasks

Communication
Deictic Interaction

Medium ~ Group Work (3)



Presentation
Overview
Pride in work result

Modeling Practices in Industry

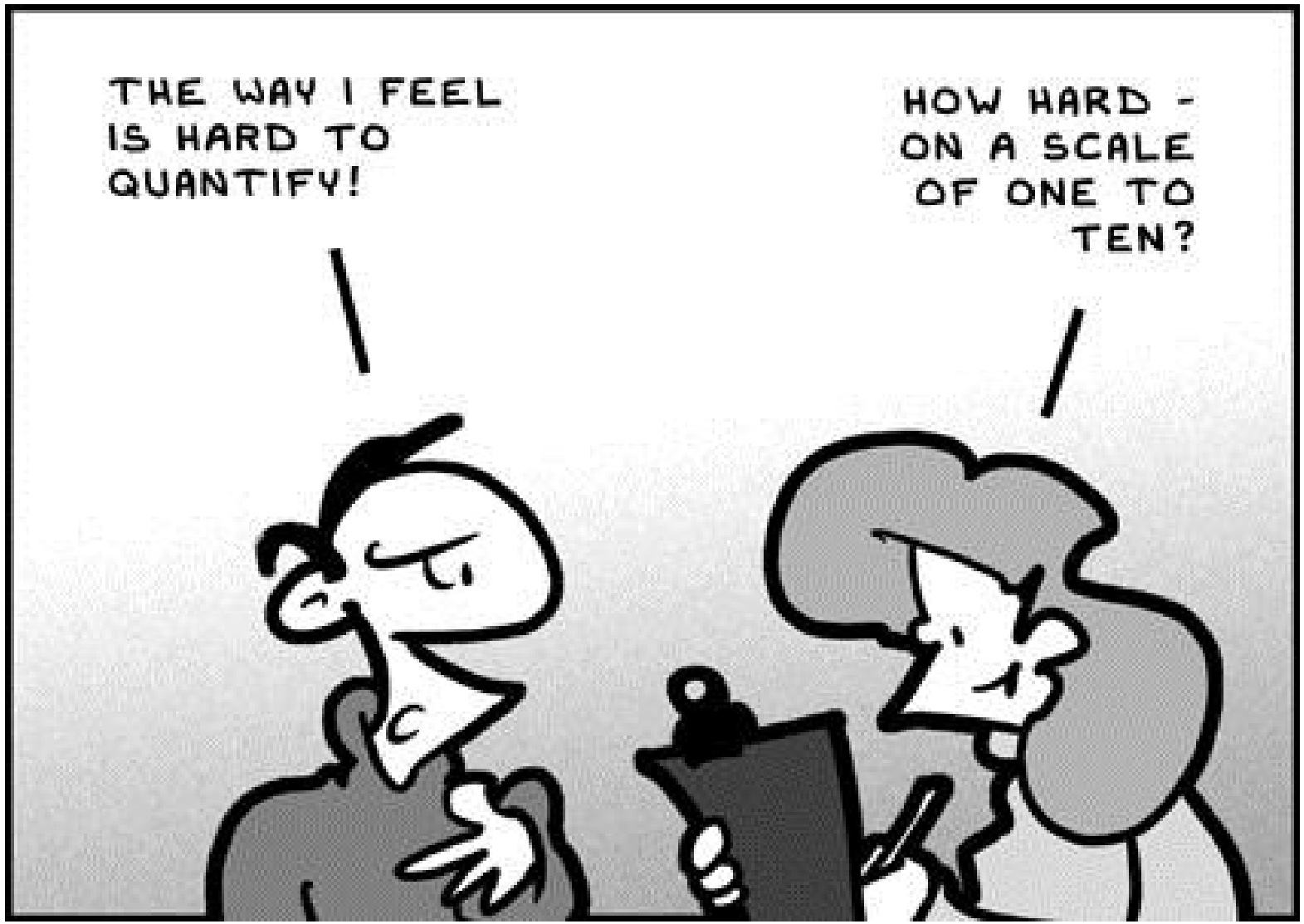
- **There are many textbooks on modeling and model based software development, but much of what they describe does not resonate with my own practical experience.**
  - Examples are always tiny and tidy.
  - In reality, you don't use always use all models.

- **In a series of studies, Dobing and Parsons tried to find out which UML diagram types are actually used for what purposes and by whom, in industry.**

- **I am currently conducting a series of interviews, with very senior modelers from industry and investigate their modeling practices:**
  - what exactly are they doing
  - why are they doing it
  - what effects do they observe

- **Observing student modelers, I study influence factors on model usage.**
  - Modeling medium
  - Group composition
  - Constraints

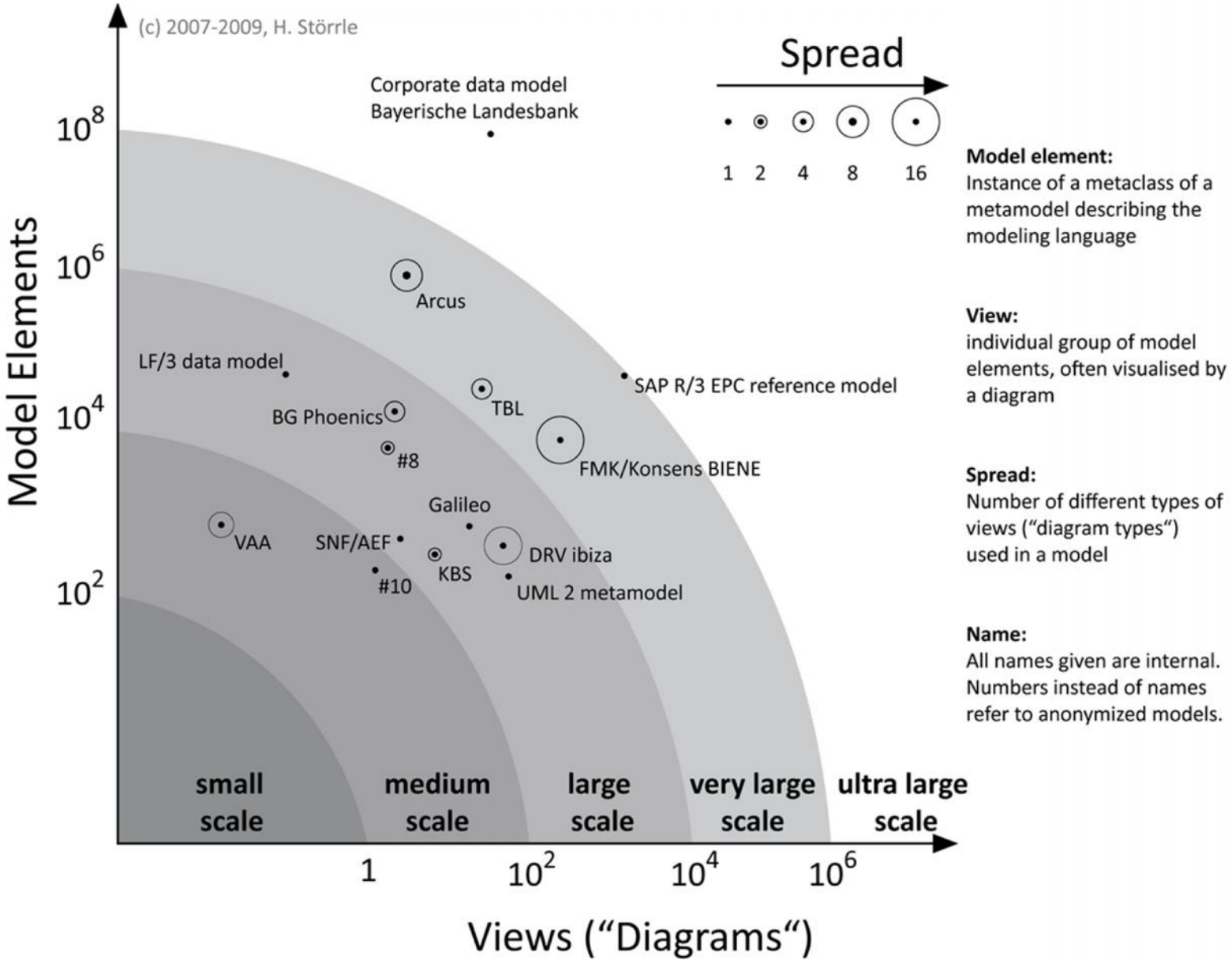
Quantification / Subjective Assessment



Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet (DTU)

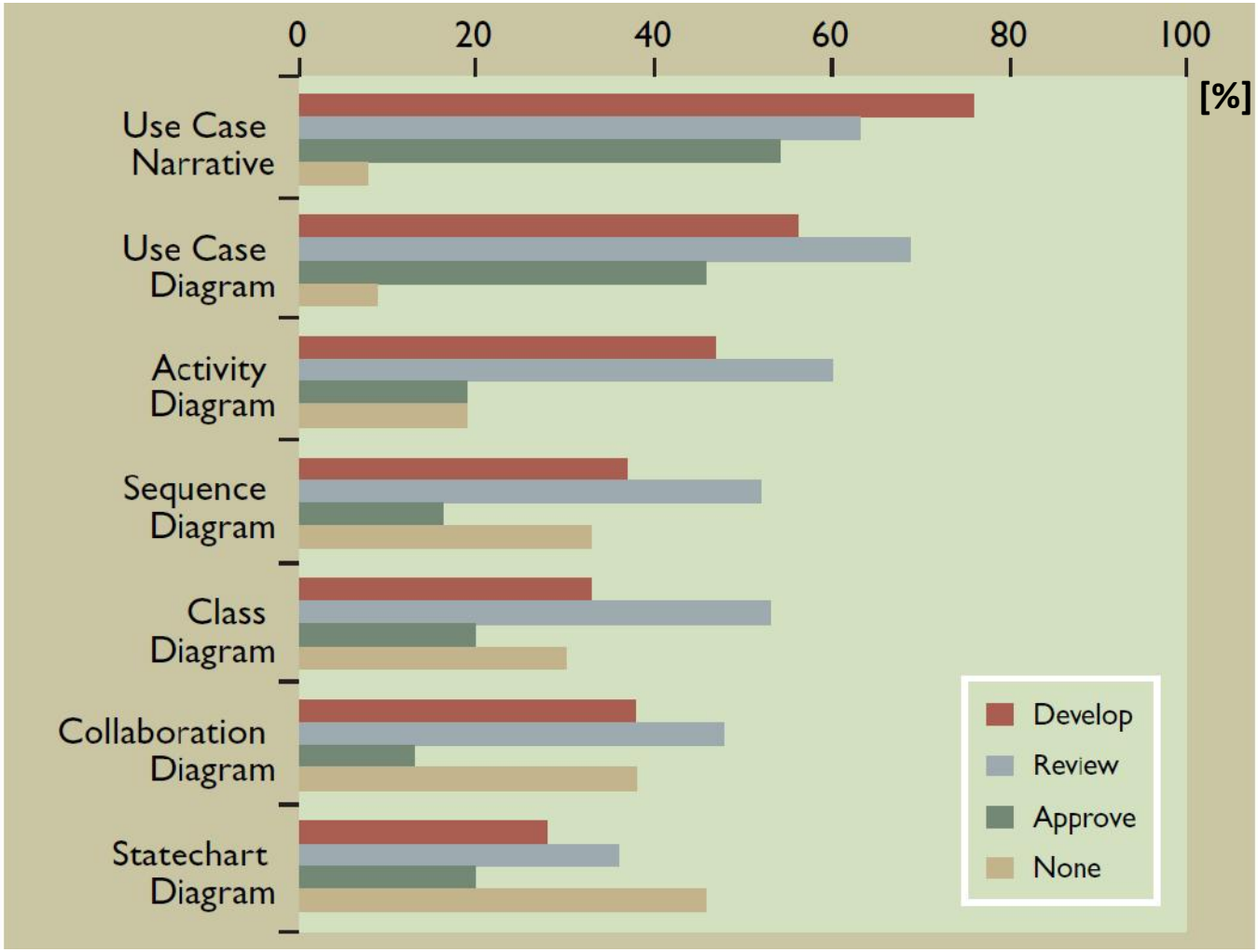
SECONDARY STUDIES

Recruiting Bias


(c) 2007-2009, H. Störrle
Spread
1 2 4 8 16
Corporate data model
Bayerische Landesbank
Arcus
LF/3 data model
SAP R/3 EPC reference model
TBL
BG Phoenics
#8
FMK/Konsens BIENE
Galileo
VAA
SNF/AEF
DRV ibiza
KBS
#10
UML 2 metamodel
Model Elements
10^8
10^6
10^4
10^2
small scale
medium scale
large scale
very large scale
ultra large scale
1
10^2
10^4
10^6
Views ("Diagrams")
Model element:
Instance of a metaclass of a metamodel describing the modeling language
View:
individual group of model elements, often visualised by a diagram
Spread:
Number of different types of views ("diagram types") used in a model
Name:
All names given are internal. Numbers instead of names refer to anonymized models.

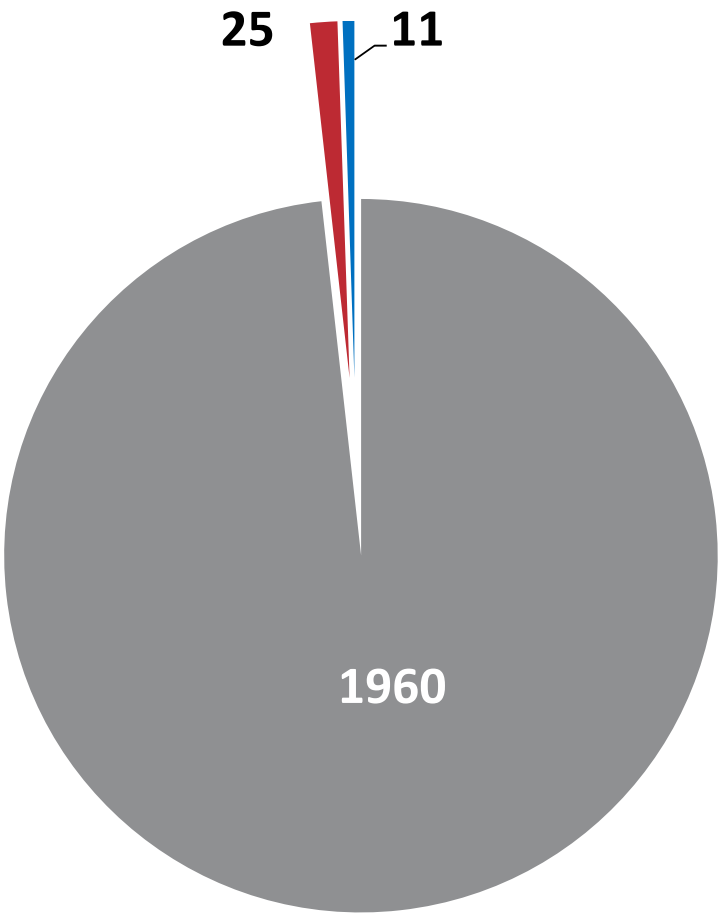
UML Diagrams used in the Analysis Phase

- **Dobing and Parsons showed which UML diagram types are used at the analysis level.**



[Dobing, Parsons: How UML is used. CACM, May 2006 (49) 5, pp. 109—113]

What do we know about Agile SE?

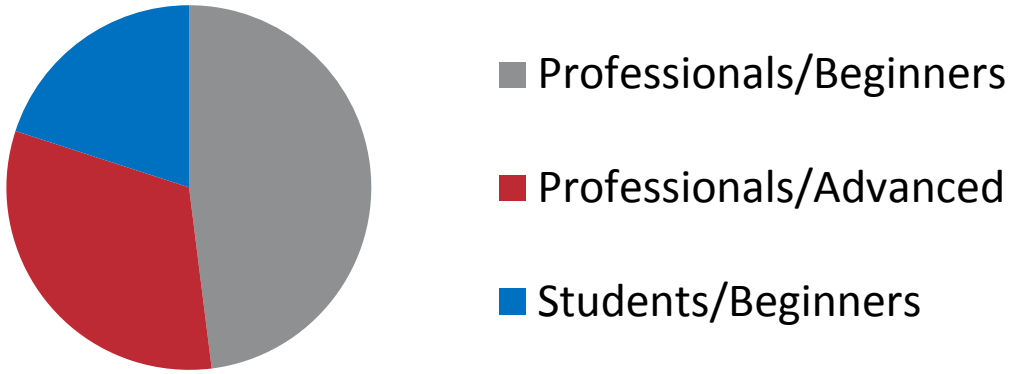


Meta-review (SLR) of all published studies relating to agile practices up to 2005.

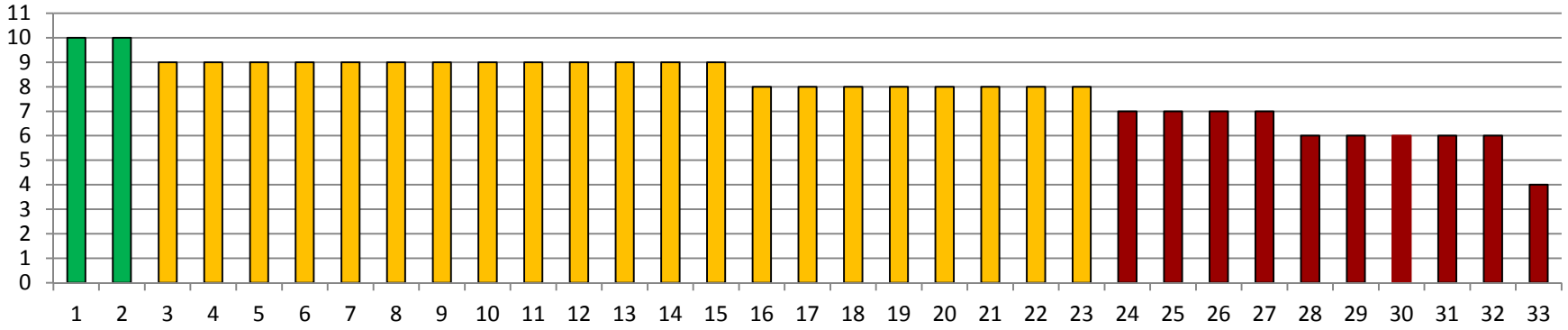
- All studies on agile approaches
- Empirical Studies (XP)
- Empirical Studies (Other)

How reliable is our knowledge?

Study Type



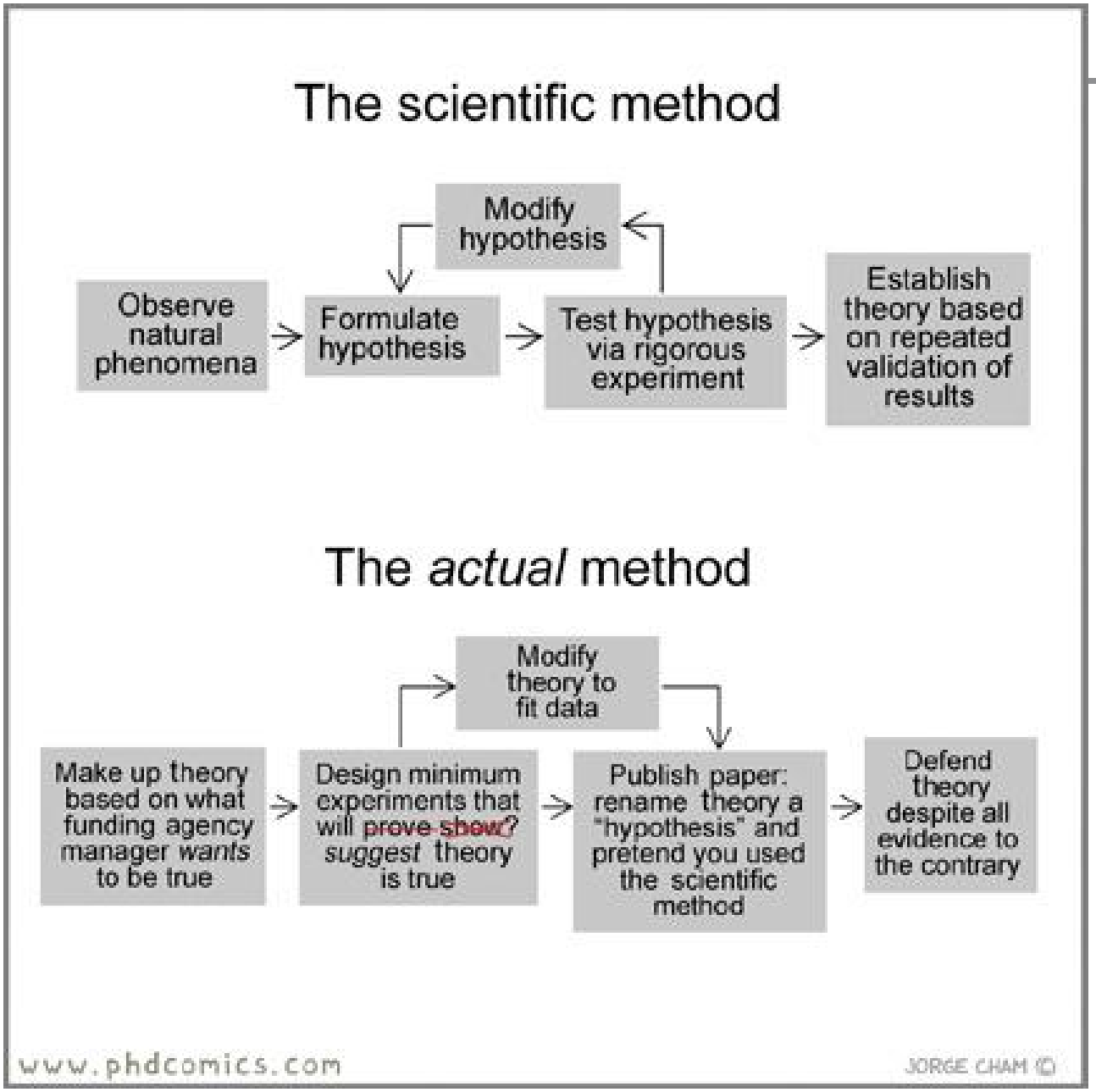
Study Quality (according to CASP)



The Critical Appraisals Skill Program (CASP) is a checklist to asses rigor, credibility, and relevance of empirical research, in particular those using qualitative methods.

Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet (DTU)

CONCLUSIONS


The scientific method
Modify hypothesis
Observe natural phenomena
Formulate hypothesis
Test hypothesis via rigorous experiment
Establish theory based on repeated validation of results
The actual method
Modify theory to fit data
Make up theory based on what funding agency manager wants to be true
Design minimum experiments that will prove show? suggest theory is true
Publish paper: rename theory a "hypothesis" and pretend you used the scientific method
Defend theory despite all evidence to the contrary
www.phdcomics.com
JORGE CHAM ©

Summary

- **So far, Informatics has focused on theory building, but validation becomes more and more important rapidly.**
  - A large portion of research ought to be validated empirically.
  - Publishing without empirical validation is increasingly difficult in SE.

- **Studying software development is effectively social science, which calls for using appropriate research methods.**
  - These research methods look soft and wooly – which makes them very hard to use, in some way, to some people.
  - There is also a certain prejudice against the non-finality of empirical results.

Empirical Methods are a powerful and indispensable complement to existing research tools in Informatics.

Prof. Dr. Harald Störrle

Software Engineering Section
Applied Mathematics and Computer Science
Technical University of Denmark

Matemtiktorvet
Building 303b, Room 056
DK-2800 Kgs. Lyngby

Tel **0045 4525 3757**
EMail **hsto@dtu.dk**
Web **www.compute.dtu.dk/~hsto**

2nd International Summer School on Empirical Research Methods in SE

August, 19th – 23rd, 2013, Kongens Lyngby, DTU
more on www.imm.dtu.dk/~hsto/ERMSE

Sign up now!

DTU