

SW Engineering empirical research: Are we doing the right thing? (mental yoga)

Lutz Prechelt
2021-02-24

30 slides

The SE research community

To make our research relevant for practitioners' success

Are we doing the right thing?
(mental yoga)

Questioning what we do to maintain our mental agility



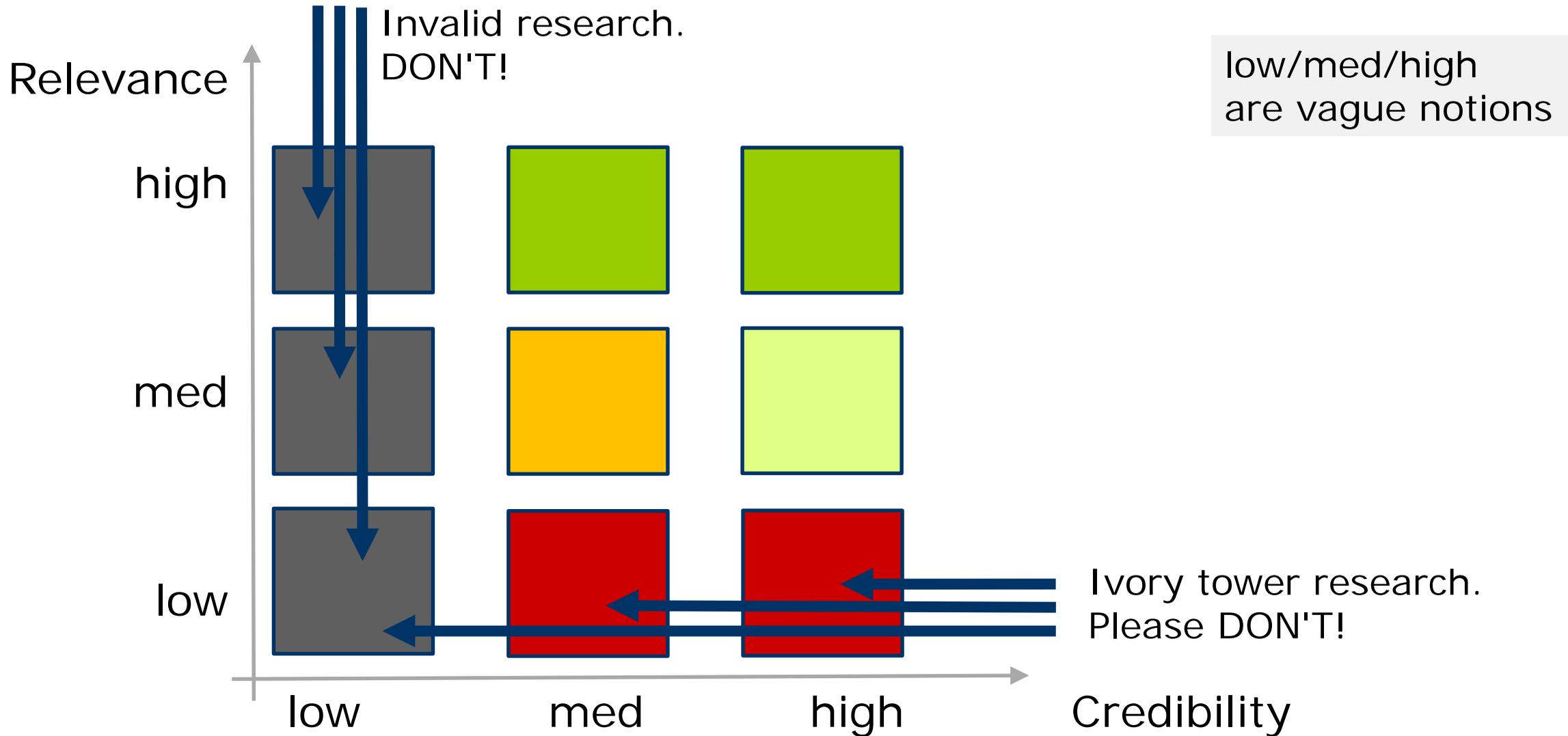
Disclaimer ("Packungsbeilage")

1. There is no single Right Thing™
2. The talk will simplify strongly
3. It is a work-in-progress report and contains provocative opinion
4. Problem 1 does not apply to tool-building work, problems 2, 3, 4 do.

Häufige Nebenwirkungen (die weniger als 1 von 10 Personen betreffen):

- Herzinfarkt
- **Gefäßentzündliche Nebenwirkungen (die bei weniger als 1 von 100 Personen auftreten):**
- Schilddrüsenfunktionsstörung
- Herzversagen, Herzrhythmusstörungen, schneller Puls
- **Seltene Nebenwirkungen (die bei weniger als 1 von 1000 Personen auftreten):**
- Magen-, Speiseröhren- oder Darmschmerzen (Blutung), Darmruptur (kann Magen- oder Darmblutungen verursachen), dunkle oder schwarze Stühle, Speiseröhrenentzündung (mit möglichen Schluckbeschwerden), Entzündung der Bauchspeicheldrüse (kann zu Magenschmerzen führen)
- **(die bei weniger als 1 von 10.000 Personen auftreten):**
- Blutung im Inneren des Gehirns mit Todesfolge
- Schwere allergische Reaktionen (einschließlich anaphylaktischer Schock) die zu Hautausschlag, Anschwellen von Gesicht, Lippen, Mund, Zunge oder Hals, Keuchhusten oder Atemnot, Schilddrüsenfunktionsstörung, führen können
- Leberfunktionsstörung, Leberschaden und schwere Leberentzündung (manchmal lebensbedrohlich oder eine Lebertransplantation erforderlich. Die Symptome können Übelkeit, sich krank fühlen, Durchfall, Gelbsucht (Gelbfärbung der Haut oder Augen), Durchfall, dunkler Urin, helle Stühle, Tendenz zu blauen, juckenden oder schmerzhaften Hautausschlag umfassen)
- Meningitis (Entzündung der Membran, die das Hirn und Rückenmark umgibt)
- Verengung einer Arterie oder Vene im Auge, die zu teilweisem oder vollständigem Sehverlust führen kann, Bindehäutenentzündung, Blutung im Auge

Message 2020: Balance credibility and relevance!
Message 2021: Insist on good relevance!



Are we doing the right thing?

View 1: Yes

"Yes. SW Engineering has made a lot of progress!"

Indication:

- SW projects have become more productive [SerPin15]
- Big reliability accidents are not becoming more frequent
 - despite increased complexity of the SW



Are we doing the right thing?

View 2: No

"No. SE research is hardly contributing."

Arguments:

- Most productivity improvements stem from better building blocks
- Agile processes were invented by practitioners
- SW security is bad and not getting better [MurKhrHam16]

→ The relevance of SE research tends to be low.

Mental yoga:
Please *assume*
this position



Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions
 3. Disdain for specialization
 4. Low attention to sociological effects
- Seriousness assessment
- Why does that happen?
- A silver lining

Is a problem for:

empirical research

empirical & constructive research

empirical & constructive research

empirical & constructive research

Frederick Brooks: "[The Computer Scientist as Toolsmith II](#)", CACM 1996

The scientist *builds in order to study;*
the engineer *studies in order to build.*

- Science is about knowledge
- Engineering is about usefulness
 - Cf. the IEEE's [mission statement](#):
"IEEE's core purpose is to foster technological innovation and excellence for the benefit of humanity."

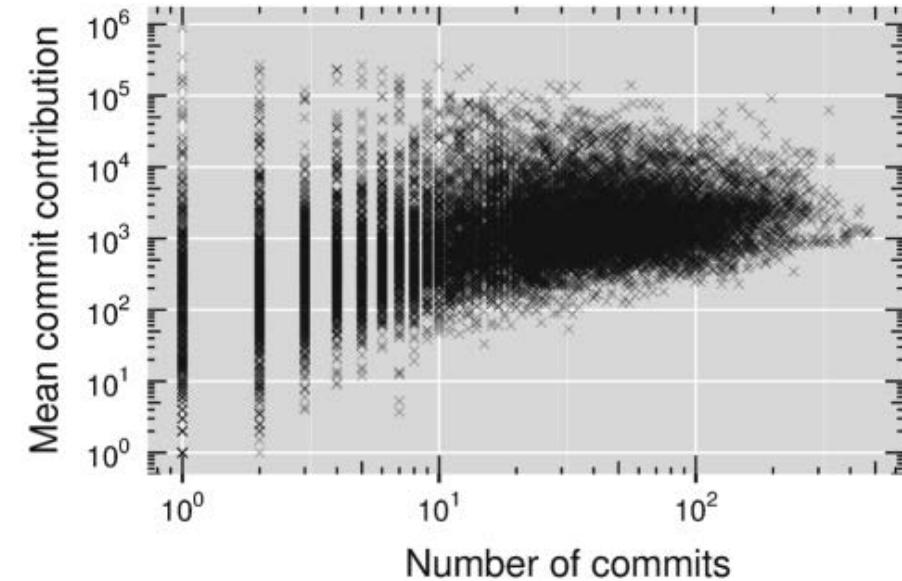
Therefore:

- **Articles that do not explain how their contribution might be useful are (presumably) not Software Engineering.**

Problem 1: Confusing engineering with science

Example

- [SchMavSch16]: "[From Aristotle to Ringelmann: a large-scale analysis of team productivity and coordination in Open Source Software projects](#)", EMSE 2016
- Looks at absolute productivity of 49 Open Source projects in relation to their changing team size over time.
 - Complicated analysis, lots of good approaches in many respects
- Finds that absolute productivity increases sublinearly.
- Does not provide any explanation how this knowledge might be relevant
 - (it might be, but that needs to be explained)



Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. **Making unwarranted assumptions**
 3. Disdain for specialization
 4. Low attention to sociological effects
- Seriousness assessment
- Why does that happen?
- A silver lining

Problem 2: Making unwarranted assumptions

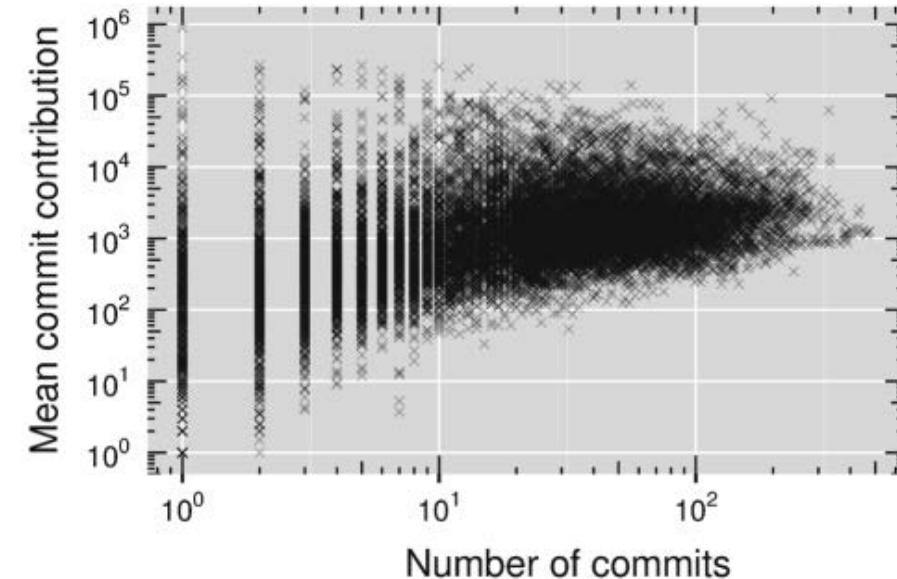
- All empirical work interprets data
 - and hence needs to make assumptions
- Some frequent assumption types:
 1. Controlled experiment: The results from our toy tasks generalize
 2. Quant. questionnaire survey: Respondents assume the same meaning we do
 - e.g. "Do you use Pair Programming?"
 3. Qualitative interview: Respondents can provide accurate accounts and do not rationalize post-hoc
 4. Tool-building: The constraints of my tool can often be met
- Insofar as these assumptions are invalid, the research findings will often become invalid.



Problem 2: Making unwarranted assumptions

Example

- Again: "[From Aristotle to Ringelmann: a large-scale analysis of team productivity and coordination in Open Source Software projects](#)"
- Finding: OSS absolute productivity increases sublinearly with team size
 - "Productivity": size of code changes per time unit
- Key assumptions:
 1. size of code changes is a reasonable measure of output [explicit, defensible]
 2. expected commit size does not depend on developer [implicit, ridiculous]
- But: OSS participants' engagement varies by several orders of magnitude: "core/periphery" [CroWeiHow12].
The periphery mostly sends small bug fixes.



Problem 2: Making unwarranted assumptions

General case

**This is
very bad!**

- SE thinking is dominated by a rationalist world view, which claims:
 1. Teams work **methodically**: They strive to follow an idealized process, involving reasonable planning and reasonable decision-making
 2. They are approximately successful in their striving; any deviance is a defect
- In reality, however, *very many* process steps are ad-hoc, planning and reality often differ *drastically*, and some decisions *hardly* reflect the available information:
 - "**amethodical** development" [TruBasTra00], plenty of evidence that it exists
 - It appears to be the norm, not the exception [Ralph10]
- **The point:** The rationalist view's dominance creates misleading assumptions
 - e.g.: "A given system structure is the result of design decisions"
 - when in fact most of those decisions were not based on design considerations

Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions
 - 3. Disdain for specialization**
 4. Low attention to sociological effects
- Seriousness assessment
- Why does that happen?
- A silver lining

Problem 3: Disdain for specialization

- What is the better tool:
 - The general Monkey Wrench?
 - The specialized set of fixed wrenches?
- Hint: General methods are called "weak methods" in the problem solving literature: lower cognitive fit
 - I. Vessey, R. Glass: "[Strong vs. Weak Approaches to Systems Development](#)", CACM 1998
- SE still operates like professional mechanics never would:
We strive for the most general tools, not the most appropriate ones
 - and likewise for empirical studies
- Example: You don't need one, do you?



Problem 3: Disdain for specialization [continued]

Conjecture:

- To understand and improve amethodical development, we will need empirical studies (and tools as well) to become intentionally more specialized



Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions
 3. Disdain for specialization
 4. **Low attention to sociological effects**
- Seriousness assessment
- Why does that happen?
- A silver lining

Problem 4: Low attention to sociological effects

- We now agree that SE is a socio-technical activity [[ICSE21CfP](#)]
- Yet few works *focus* on socio (e.g. culture) effects [DybDin08]
 - (this is currently getting better)

Conjecture:

- To understand and improve amethodical development, we will need empirical studies (and tools as well) to pay more attention to non-technical phenomena.

- Rashina Hoda researched roles in self-organization of agile teams
 - she has a Computer Science background
 - [HodNobMar13] Hoda et al.: "[Self-Organizing Roles on Agile Software Development Teams](#)", TSE 2013
 - **Findings:** 6 Roles: Mentor, Coordinator, Translator, Champion, Promoter, Terminator [*fits with methodical development*]



- Helena Barke did the same again later
 - she has a background in CS and in Gender and Diversity
 - [BarPre19] Barke, Prechelt: "[Role clarity deficiencies can wreck agile teams](#)", PeerJ CS 2019
 - **Findings:** Roles are unique, crafted to fit person+team+project. If person and team cannot agree on a role, people often leave the team [*more socio-intensive, fits with amethodical development*]



Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions
 3. Disdain for specialization
 4. Low attention to sociological effects
- **Seriousness assessment**
- Why does that happen?
- A silver lining

How serious
is this?



1. Confusing engineering with science

- Three sciency articles that survived ICSE 2020 reviewing indicate we are not well sensitized to the science/engineering distinction:
- a246: how programmers use 'unsafe' in Rust (no usefulness claim)
- a691: strange study about interference when learning 2nd prog language
- a752: defect taxonomy (no usefulness claim)

Ongoing work with
Franz Zieris

2. Making unwarranted assumptions

- Assumptions can be reliable, reasonable, risky, or ridiculous.
Three 'ridiculous' assumptions in ICSE 2020 indicate a community problem:
- a542: Assuming usefulness of numbers 98%/49%/55k involves assumptions:
"all compiler updates are important", "all vulnerabilities are fixed at once",
"all smart contracts are security-critical"
- a752: Defects belong into neatly disjoint classes
- b328: The effort for finding a bug is worthwhile for any bug [a popular one!]

3. Disdain for specialization

- Unclear how to judge that any given article should have been more specialized
- Seriousness is unclear (I have some doubts)

4. Low attention to sociological effects

- Unclear how to judge that any given article should have paid more attention to some socio-issue X. **But:**
- de Marco et al.: "[Adrenaline Junkies and Template Zombies](#)" (book, 2008/2013) sketches 86 patterns (positive or negative) of SW project behaviors. About two thirds of the negative ones suggest cultural reasons.
 - Curious example: "Shipping on time, every time"
- Yet I am not aware of much work in these directions.

Overall: Quite serious

Why does SE research tend to have low relevance?

Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions ← **conjecture: our biggest mistake**
 3. Disdain for specialization
 4. Low attention to sociological effects
- Seriousness assessment
- **Why does that happen?**
- A silver lining

Why do the 4 problems occur?

I suspect the following main reasons:

1. Positivist bias

- "detailed numbers? → good study"
- Positivists prefer the rationalist view and dislike the idea of amethodical development

2. Lack of willingness to pick high-hanging fruit

- A culture effect!

3. Too few people recognize the assumptions they make

- *Threats to Validity* sections are often depressing

Speculation.
My evidence is
anecdotal only.

Conjecture:

We need to accept:

**There are no
simple truths
in SW Eng.**

Why does SE research tend to have low relevance?

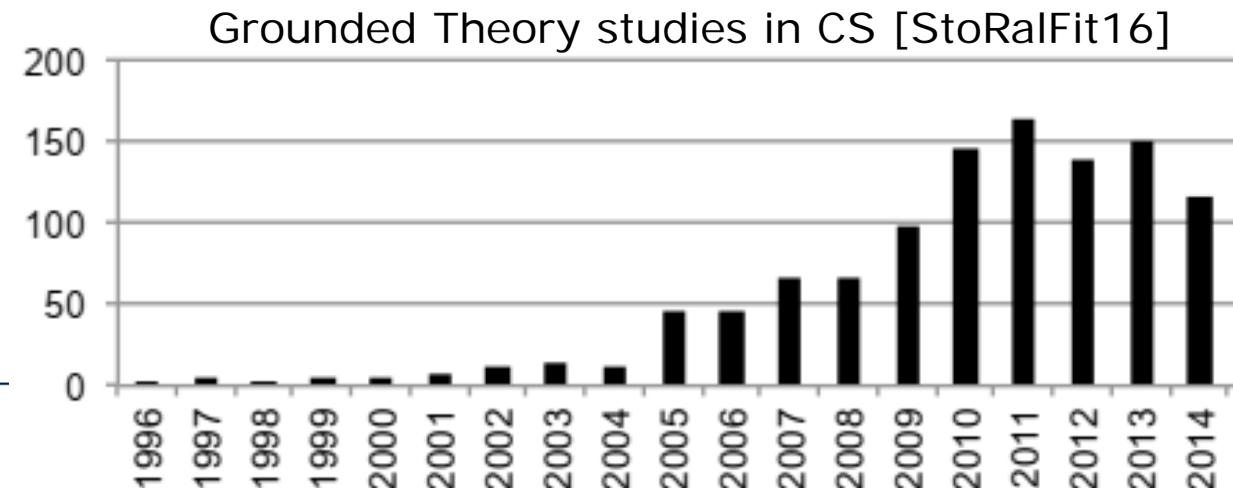
Structure of the talk:

- 4 problems that contribute
 1. Confusing engineering with science
 2. Making unwarranted assumptions
 3. Disdain for specialization
 4. Low attention to sociological effects
- Seriousness assessment
- Why does that happen?
- **A silver lining**

Any silver lining?

Yes.

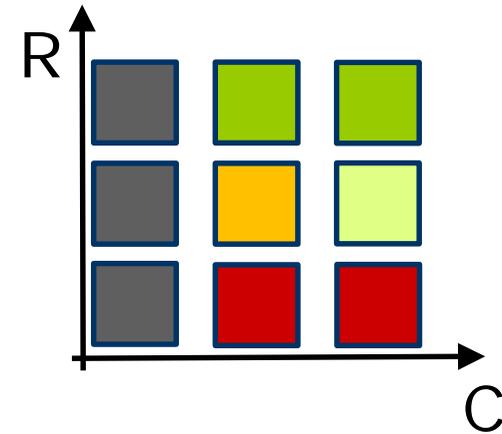
- Controlled experiments have gone out of fashion
- Increase of qualitative studies (qS) promises improvement:
 - although so far they are often very broken [StoRaIFit16]
 - Lack of specialization? qS are usually specialized
 - Attention to sociology? qS methods are usually closer to people
- Not sure about:
 - Unwarranted assumptions? Are qS assumptions more reliable?
 - Science vs. engineering:
Near-irrelevant questions can still be asked in qS
- Pessimistic about:
 - Our attitude regarding acceptable effort



- [CroWeiHow12] Crowston et al.: "Free/Libre Open-Source Software Development: What We Know and What We Do Not Know", ACM Comp.Surveys 2012
- [DybDin08] Dybå, Dingsøyr: "Empirical studies of agile software development: A systematic review", IST 2008
- [MurKhrHam16] Murtaza et al.: "Mining trends and patterns of software vulnerabilities", JSS 2016
- [Ralph10]: "Comparing Two Software Design Process Theories", DESIRIST 2010
- [SerPin15] Serrador, Pinto: "Does Agile work? — A quantitative analysis of agile project success", Int'l. J. of Project Mgmt. 2015
- [TruBasTra00] Truex et al.: "Amethodical systems development: the deferred meaning of systems development methods", AMIT 2000.

Thank you!

Discussion, please!

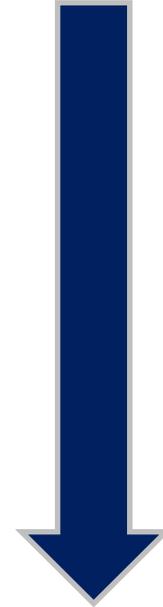


1. As reviewers, will you insist on plausible relevance?
2. Are you going to change your own research questions?
Your methods? Do longer studies? Go into the field?

- Yoga man: <https://www.theluxuryspot.com/wp-content/uploads/2010/01/yoga1228623392.jpg>
- Packungsbeilage: <https://www.offenesblog.de/wp-content/uploads/2011/08/packungsbeilage.jpg>
- Trump rally: <https://spacecoastdaily.com/wp-content/uploads/2020/06/Trump-rally-600-2.jpg>
- Rescue: <https://www.sn.at/panorama/oesterreich/retter-werden-immer-oeffter-bei-der-arbeit-behindert-40214887>, Robert Ratzner
- Tipp Kick: [https://www.tipp-kick.com/out/pictures/generated/product/1/540_340_75/cup-spielszene\(4\).jpg](https://www.tipp-kick.com/out/pictures/generated/product/1/540_340_75/cup-spielszene(4).jpg)
- Wrenches: <https://www.obl.de>
- Brushes: <https://www.pinterest.com/pin/416864509232053678/>
- Band-aid legs: <https://babyccinokids.com/blog/2017/04/11/tuesday-tips-a-pretty-plaster-for-every-pain/>
- Surgery: <https://www.india.com/lifestyle/what-is-open-heart-surgery-things-you-need-to-known-about-open-heart-surgery-1709556/>

Given a broad research interest, e.g.

- How should we use X?
 - e.g. models or modeling or pair programming or ...
- How does X compare to Y?
 - e.g. maintainability of Java code versus Python code, or ...



- A sensible progression of research could be:
 - Understand relevant factors
 - identify, describe
 - Formulate a theory of their relationships (mechanisms)
 - talks about the development process
 - Validate the theory
 - Measure the size of certain effects in the theory
 - Quantification, based on the qualitative theory

Given a broad research interest, e.g.

- **How can we best solve X?**
 - **by any kind of tool support**



- A sensible progression of research could be:
 - Understand relevant **problems**
 - identify, describe
 - Formulate a theory of their relationships (mechanisms)
 - talks about the development process
 - Validate the theory
 - **Find one or more points of attack**
 - **where improvements will be most useful**
 - **Devise and build helpful tools**

Premature tool-building is much like premature quantification