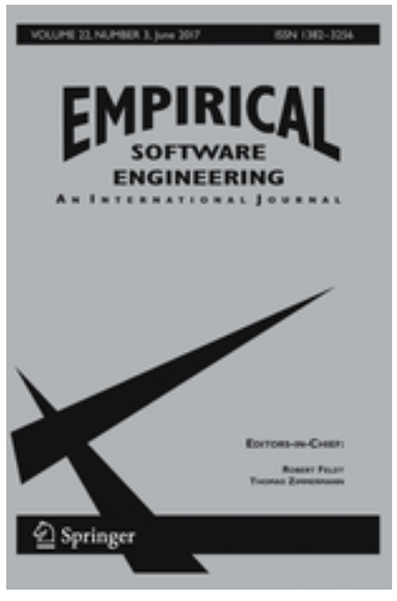
Empirical Software Engineering as a Science:
A Manifesto

METODOS track JISBD/CEDI 2021, Malaga/Online

Robert Feldt

About me and Preamble



I only know a little bit about this

You know this already

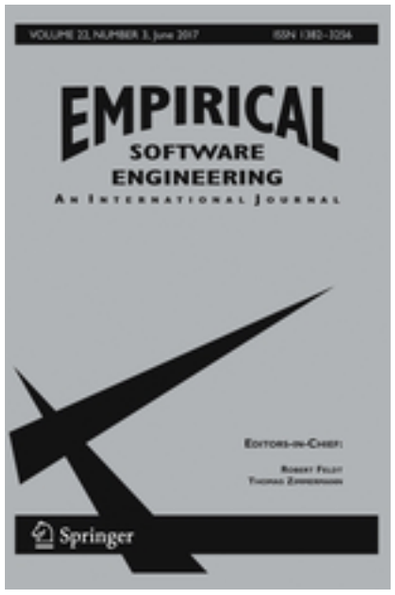
Our time here is limited

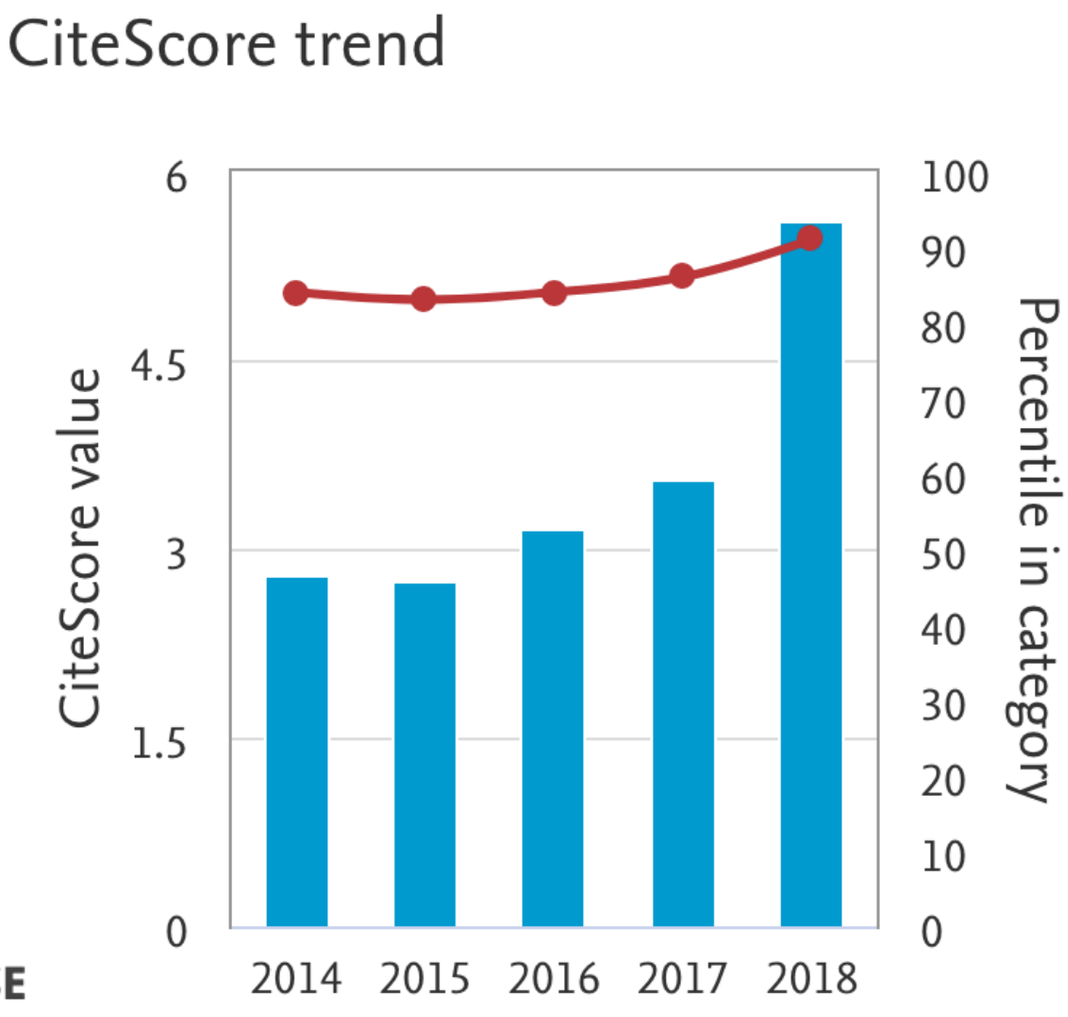
When negative, I criticise myself as much as us/you

We won!

Our venues increasing in size and importance

+30 to +50%/year in submissions 2018-2021





Clarivate's Journal Impact Factor

Year	JIF	Rank overall	Rank SE
2018	4.457	8th of 107	2nd of 18
2017	2.933	11th of 104	2nd of 18
2016	3.275	7th of 106	1st of 18
2015	1.393	27th of 106	6th of 18

Empirical SE concepts in ICSE

Median number of empirical “concepts” mentioned per ICSE paper
(for 20 random ICSE papers, per year)

Concept	1999	2009	2019	2022 submitted
Experiment	1.0	1.0	7.5	9.0
Empiric*	0.5	1.0	3.0	3.0
Validity	0.0	0.5	1.0	2.0

Also, the most common keyword in ICSE 2021: ”empirical study”

Increasing use of statistical analysis

Evolution of statistical analysis
in empirical software engineering research: Current state and steps forward

Francisco Gomes de Oliveira Neto[a,*], Richard Torkar[a], Robert Feldt[a,b], Lucas Gren[a], Carlo A. Furia[c], Ziwei Huang[a]

[a] *Chalmers and the University of Gothenburg, SE-412 96 Gothenburg, Sweden*
[b] *Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden*
[c] *Università della Svizzera italiana, CH-6900 Lugano, Switzerland*

Abstract

Software engineering research is evolving and papers are increasingly based on empirical data from a multitude of sources, using statistical tests to determine if and to what degree empirical evidence supports their hypotheses. This is not only crucial for research progress but also for practitioners in judging the practical significance. To investigate the practices and trends of statistical analysis in empirical software engineering (ESE), this paper presents a review of a large pool of papers from top-ranked software engineering journals. First, we manually reviewed 161 papers producing a review protocol based on a view of the recent state of art concerning statistical analysis and how researchers discuss practical significance. In a second phase of our method, we used the protocol as ground truth for a more extensive semi-automatic classification of papers spanning the years 2001–2015 targeting a total of 5,196 papers.

https://arxiv.org/abs/1706.00933 accepted for publication in JSS, July 2019

Increasing use of statistical analysis

Quantitative

Stat. Test

Parametric

Nonparametric

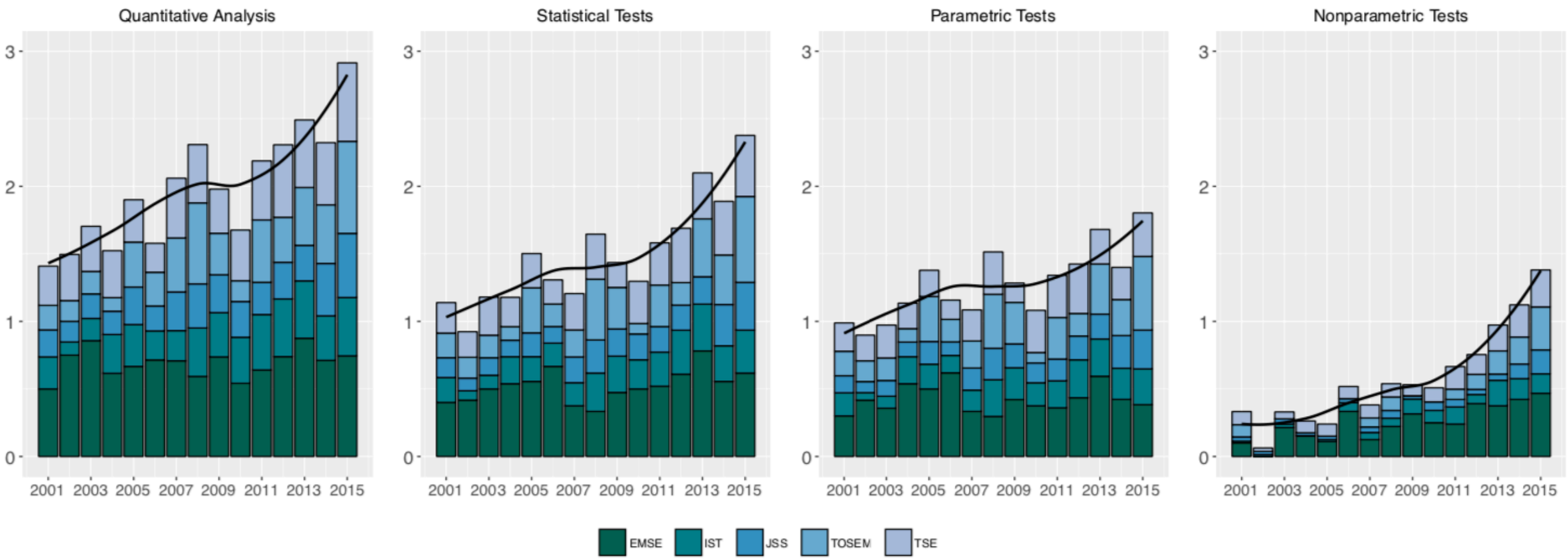
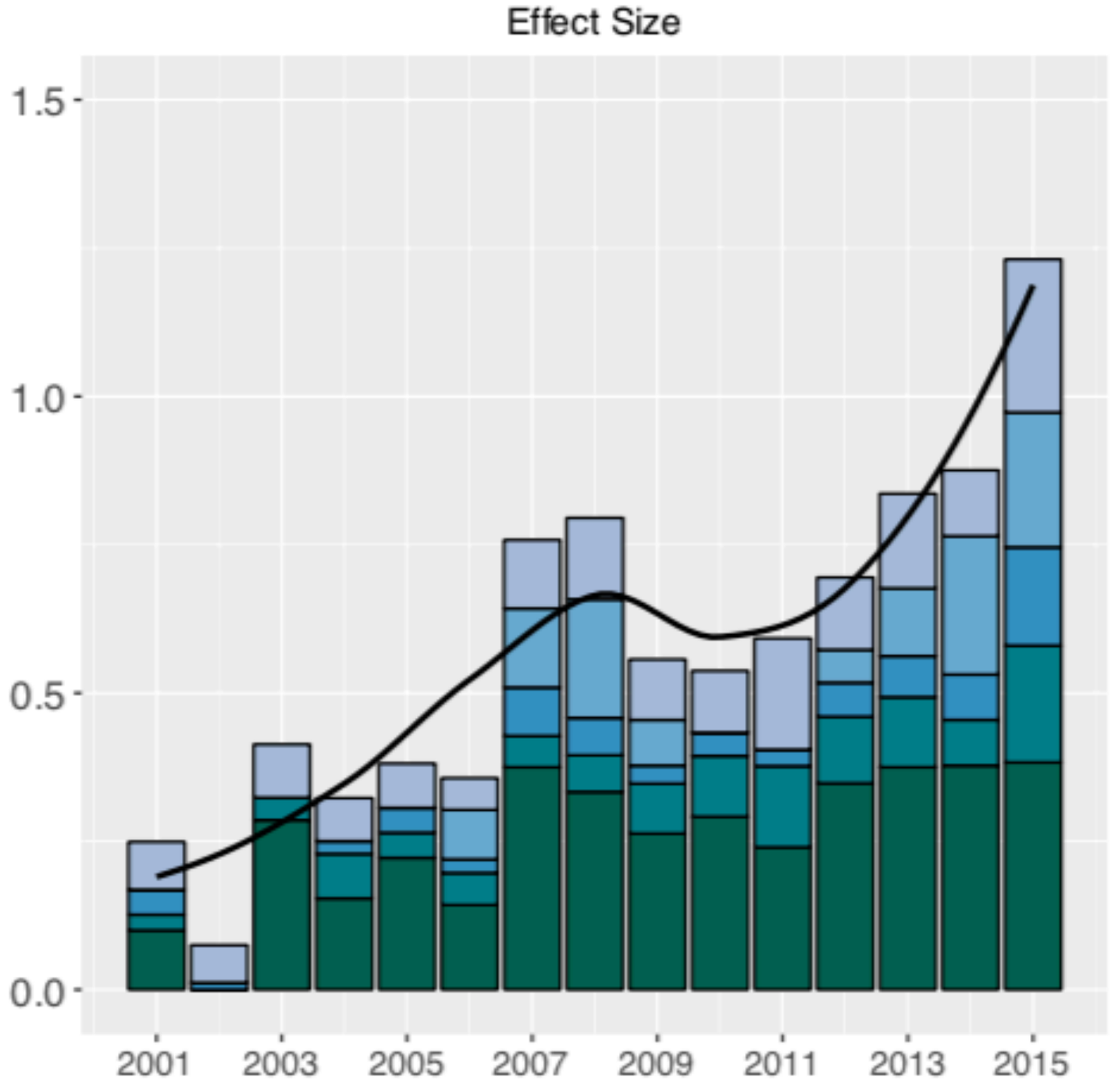


Figure 5: The y-axis in each chart is the normalization of ratings of the number of papers where we found positive evidence. Notice that the scale on the y-axis is still lower than the maximum ratio (5). The thick line is a local regression (loess) of the data.

https://arxiv.org/abs/1706.00933 accepted for publication in JSS, July 2019

Increasing use of statistical analysis



https://arxiv.org/abs/1706.00933 accepted for publication in JSS, July 2019

But....

Identity?

Real progress?

Next steps?

Manifesto for Empirical Software Engineering

Through systematic research we are
uncovering a science of software engineering
so that we can better help software practitioners.
Through this work we have come to value:

Empirical evidence over theoretical & formal arguments

Systematic & explicit methods over one-off, unique studies

Practical context & impact over clean but simplified lab studies

That is, while there is value in the items on the right,
we value the items on the left more.

Manifesto for Empirical Software Engineering 2.0

Empirical evidence over theoretical & formal arguments

Systematic & explicit methods over one-off, unique studies

Practical context & impact over clean but simplified lab studies

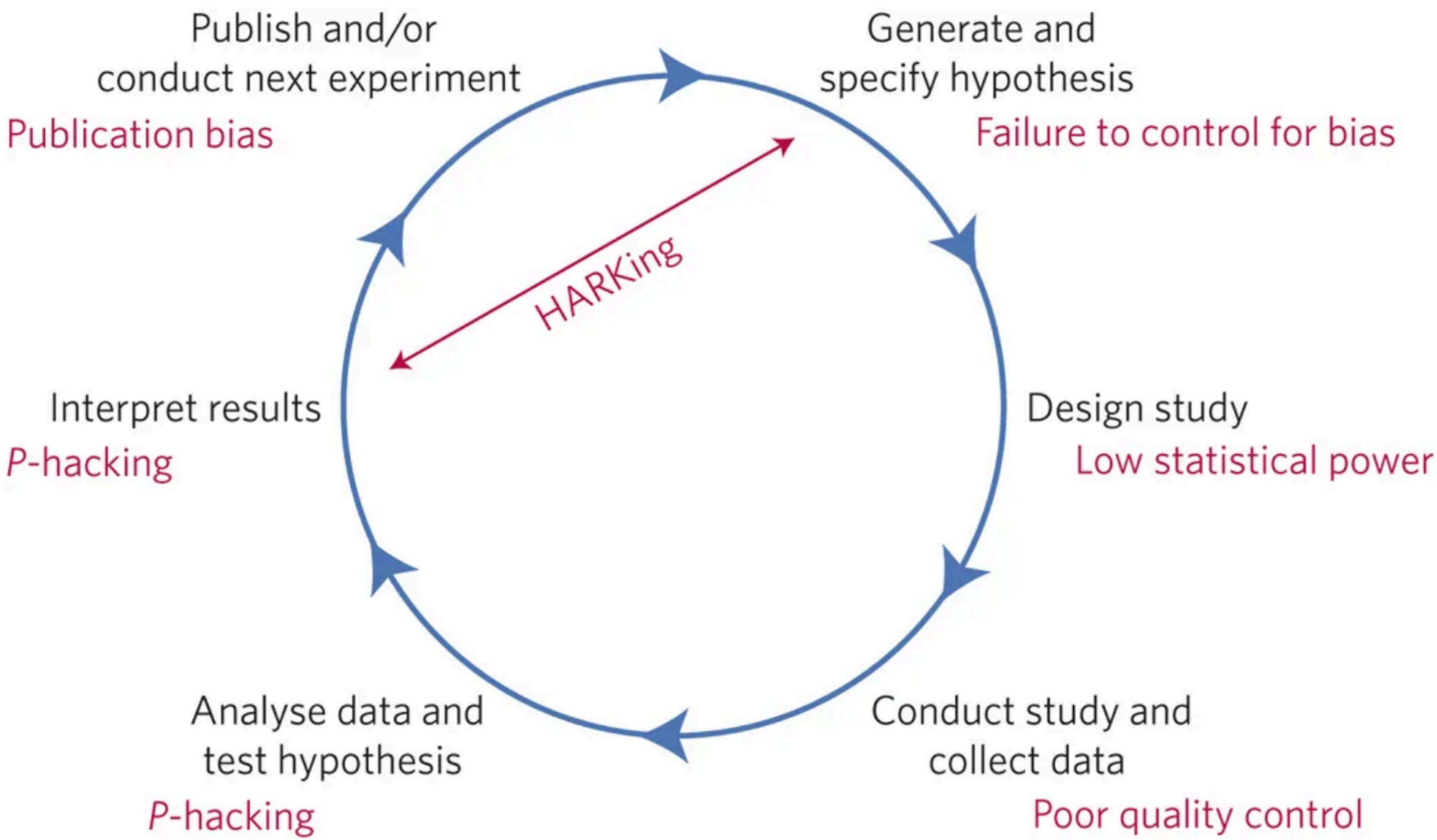
Truth over novelty, relevance and importance

Plurality & nuance over simple, dichotomous claims

Human factors over algorithms & technology

Explanations & theories over descriptions of data at hand

Some threats to finding the Truth



from Munafò et al, "A Manifesto for Reproducible Science", Nature, 2017

A Truth root challenge: Neophilia

neophilia / (ˌniːəʊˈfɪlɪə) /

noun

1 a tendency to like anything new; love of novelty

Some effects of Neophilia

Publication bias / “results paradox”: We accept clear and positive results ($p<0.05$) while rejecting “negative” or inconclusive ones

Isolated paper islands: Authors must create new model, system, solution, idea rather than replicating and building on what is already there.

HARKing: changing **H**ypothesis **A**fter **R**esults are **K**nown

Truth Fix: (Pre-)Registered Reports

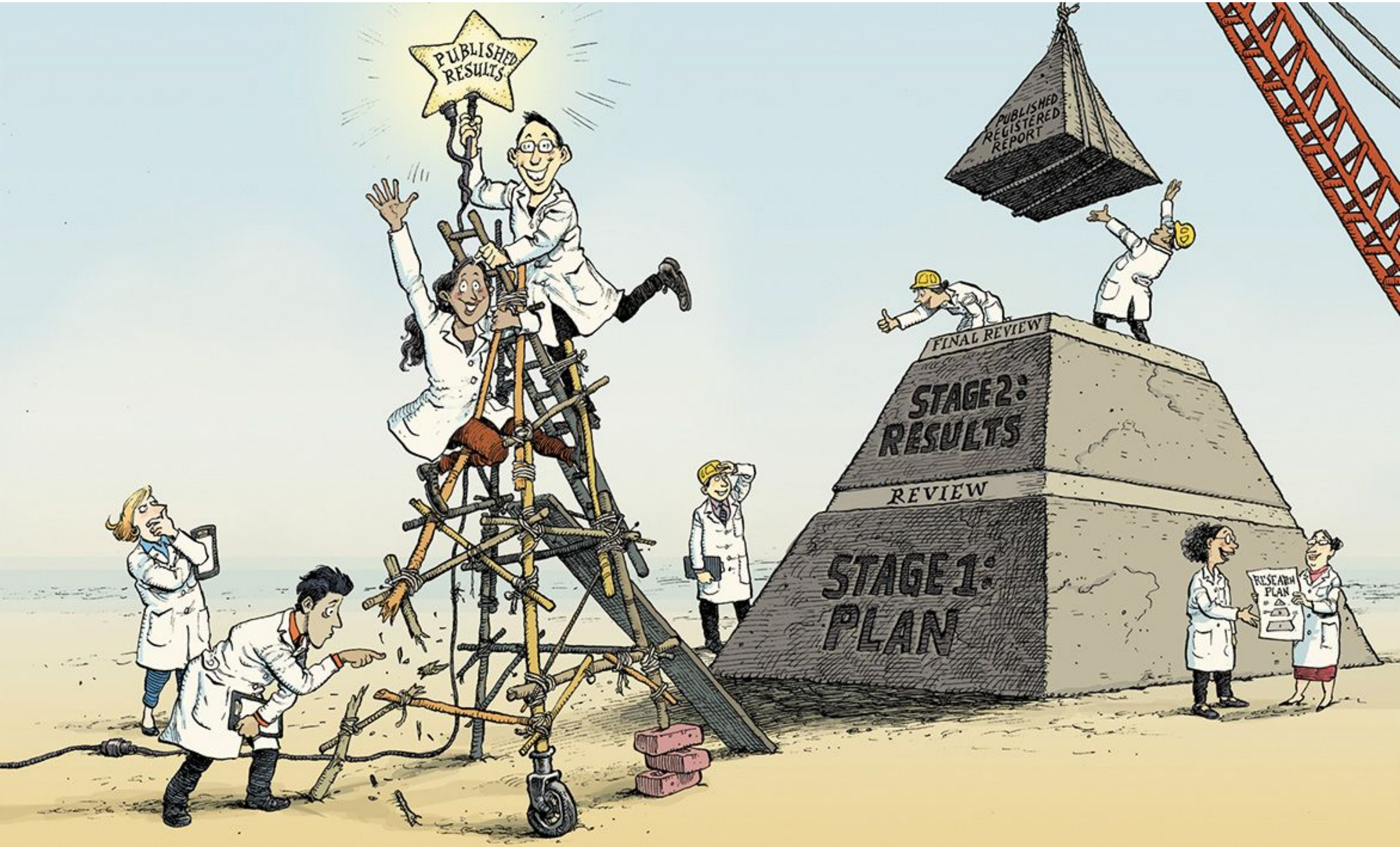
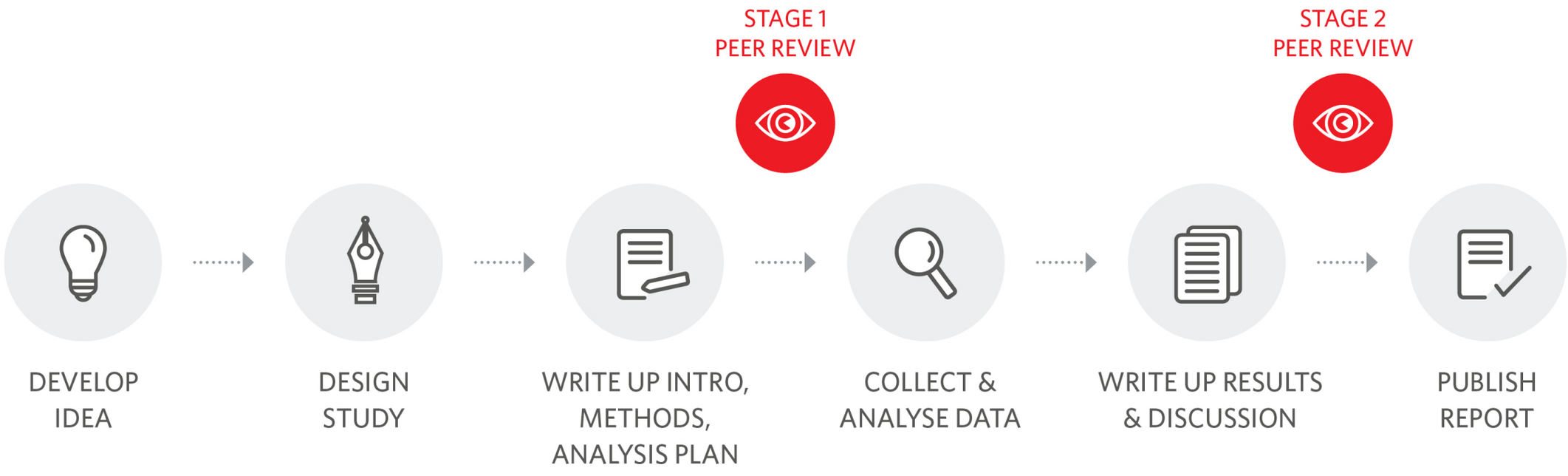


Illustration by David Parkins in Nature, September 2019

Truth Fix: (Pre-)Registered Reports



MSR EMSE

A form of self-blinding, next step after double blind!

200+ Journals today offer pre-registration!

Acceptance rate in stage 2: 90% (Cortex journal)

Null results: 66% RR replicat., 50% RR novel, 5-20% non-RR

Counterpoint: (Pre-)Registered Reports

RRs for confirmatory, hypothesis-driven research

They are not a good fit for more exploratory work

Alternative: Explorative Reports?

Counterpoint: (Pre-)Registered Reports

ELSEVIER

Cortex

Volume 96, November 2017, Pages A1-A4

Editorial

Exploratory reports: A new article type for *Cortex*

Robert D. McIntosh

Show more

https://doi.org/10.1016/j.cortex.2017.07.014

Previous article in issue

Next article in issue

There are many ways to find things out. In science, the process of discovery can be divided conceptually into exploratory and confirmatory phases. In the exploratory phase, we observe and explore, generating theories to explain the patterns that we find. Useful theories will support predictions about what we should and should not find in the future if

MSR/EMSE 2021 RR: 2 paper types

Paper Types, Evaluation Criteria, and Acceptance Types

The RR track of MSR 2021 supports two types of papers:

Confirmatory: The researcher has a fixed hypothesis (or several fixed hypotheses) and the objective of the study is to find out whether the hypothesis is supported by the facts/data.

Exploratory: The researcher does not have a hypothesis (or has one that may change during the study). Often, the objective of such a study is to understand what is observed and answer questions such as WHY, HOW, WHAT, WHO, or WHEN. We include in this category registrations for which the researcher has an initial proposed solution for an automated approach (e.g., a new deep-learning-based defect prediction approach) that serves as a starting point for his/her exploration to reach an effective solution.

MSR/EMSE 2021 RR: 2 outcomes (well 3 ;))

The outcome of the RR report review is one of the following:

- **In-Principal Acceptance (IPA)**: The reviewers agree that the study is relevant, the outcome of the study (whether confirmation / rejection of hypothesis) is of interest to the community, the protocol for data collection is sound, and that the analysis methods are adequate. The authors can engage in the actual study for Stage 2.
  If the protocol is adhered to (or deviations are thoroughly justified), the study is published. Of course, this being a journal submission, a revision of the submitted manuscript may be necessary. Reviewers will especially evaluate how precisely the protocol of the accepted pre-registered report is followed, or whether deviations are justified.

- **Continuity Acceptance (CA)**: The reviewers agree that the study is relevant, that the (initial) methods appear to be appropriate. However, for exploratory studies, implementation details and post-experiment analyses or discussion (e.g., why the proposed automated approach does not work) may require follow-up checks. We'll try our best to get the original reviewers. All PC members will be invited on the condition that they agree to review papers in both, Stage 1 and Stage 2. Four (4) PC members will review the Stage 1 submission, and three (3) will review the Stage 2 submission.

Note: For MSR 2021, we will only offer IPA to confirmatory study. Exploratory study in software engineering often cannot be adequately assessed until after the study has been completed and the findings are elaborated and discussed in a full paper. For example, consider a study in an RR proposing defect prediction using a new deep learning architecture. This work falls under the *exploratory* category. It is difficult to offer IPA, as we do not know whether it is any better than a traditional approach based on e.g., decision trees. Negative results are welcome;

Truth & Nuance Fix: Beyond p-values

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

To link to this article: https://doi.org/10.1080/00031305.2019.1583913

Truth & Nuance Fix: Beyond p-values

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Truth & Nuance Fix: What instead of p-values?

Ioannidis:
alpha = 0.005!

Greenland & 800 signatories:
Stop dichotomising!
Compatibility Intervals!

Wagenmakers:
Bayes factors!

Gelman:
No tests, just full
Bayesian analysis!

Truth & Nuance Fix: What instead of p-values?

Now: Lower alpha, acknowledge problem,
study compatibility interval and how to report on them!

Medium-term: Educate yourself about Bayesian analysis

Longer-term: Start using flexible Bayesian models.
When Causal analysis matures, learn it.

Truth & Nuance Fix: What instead of p-values?

Bayesian Data Analysis in Empirical Software Engineering Research

Carlo A. Furia, Robert Feldt, and Richard Torkar

Abstract—Statistics comes in two main flavors: frequentist and Bayesian. For historical and technical reasons, frequentist statistics have traditionally dominated empirical data analysis, and certainly remain prevalent in empirical software engineering. This situation is unfortunate because frequentist statistics suffer from a number of shortcomings—such as lack of flexibility and results that are unintuitive and hard to interpret—that curtail their effectiveness when dealing with the heterogeneous data that is increasingly available for empirical analysis of software engineering practice.

In this paper, we pinpoint these shortcomings, and present Bayesian data analysis techniques that provide tangible benefits—as they can provide clearer results that are simultaneously robust and nuanced. suffer from a number of shortcomings, which limit their scope of applicability and usefulness in practice, and may even lead to drawing flat-out unsound conclusions in certain contexts. In particular, the widely popular techniques for null hypothesis statistical testing—based on computing the infamous *p*-values—have been *de facto* deprecated [7], [41], but are still routinely used by researchers who simply lack practical alternatives: techniques that are rigorous yet do not require a wide statistical know-how, and are fully supported by easy-to-use flexible analysis tools.

Bayesian statistics has the potential to replace frequentist

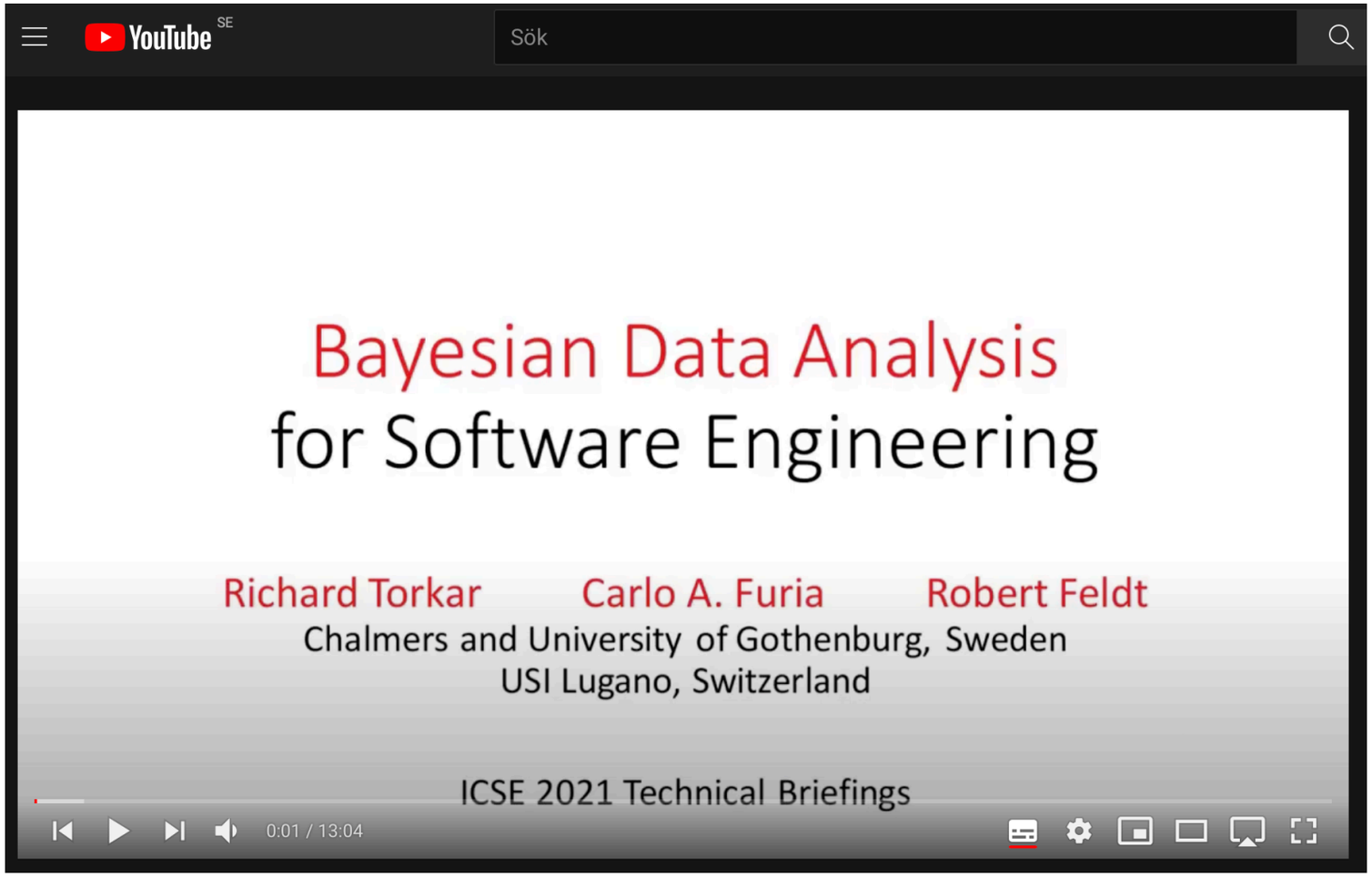
https://arxiv.org/abs/1811.05422 accepted for publication in TSE, July 2019

Truth & Nuance Fix: Bayes in SE ICSE Tutorial

We held an ICSE 2021 Tutorial/Tech briefing on the use of Bayesian methods in Software Engineering

Videos, slides and additional information can be found:

https://robertfeldt.github.io/research/bayesian_se/



Nuance Challenge: Pseudo-profound bullshit

Judgment and Decision Making, Vol. 10, No. 6, November 2015, pp. 549–563

On the reception and detection of pseudo-profound bullshit

Gordon Pennycook* James Allan Cheyne† Nathaniel Barr‡ Derek J. Koehler†

Jonathan A. Fugelsang†

Abstract

Although bullshit is common in everyday life and has attracted attention from philosophers, its reception (critical or ingenuous) has not, to our knowledge, been subject to empirical investigation. Here we focus on pseudo-profound bullshit, which consists of seemingly impressive assertions that are presented as true and meaningful but are actually vacuous. We presented participants with bullshit statements consisting of buzzwords randomly organized into statements with syntactic structure but no discernible meaning (e.g., "Wholeness quiets infinite phenomena"). Across multiple studies, the propensity to judge bullshit statements as profound was associated with a variety of conceptually relevant variables (e.g., intuitive cognitive style, supernatural belief). Parallel associations were less evident among profundity judgments for more conventionally profound (e.g., "A wet person does not fear the rain") or mundane (e.g., "Newborn babies require constant attention") statements. These results support the idea that some people are more receptive to this type of bullshit and that detecting it is not merely a matter of indiscriminate skepticism but rather a discernment of deceptive vagueness in otherwise impressive sounding claims. Our results also suggest that a bias toward accepting statements as true may be an important component of pseudo-profound bullshit receptivity.

Keywords: bullshit, bullshit detection, dual-process theories, analytic thinking, supernatural beliefs, religiosity, conspiratorial ideation, complementary and alternative medicine.

Humans & Plurality Fix: Lifting Qualitative Methods

1. Use **broader set of Qual methods** from Social Science!

2. **Emphasize Reflexivity**!
Researcher is part of social world she studies and the relationship to participants is explicit & transparent.

3. Adapt & **employ existing Qual checklists**!

Humans & Plurality Fix: Standards & Checklists

equator network

Enhancing the QUAlity and Transparency Of health Research

EQUATOR resources in German | Portuguese | Spanish

EQUATOR Network: what we do and how we are organised

The EQUATOR Network is an “umbrella” organisation that brings together researchers, medical journal editors, peer reviewers, developers of reporting guidelines, research funding bodies and other collaborators with mutual interest in improving the quality of research publications and of research itself.

We are developing into a global initiative covering all areas of health research and all nations, and actively involving all key stakeholders. We have launched the first four national centres that will substantially contribute to expanding EQUATOR activities: the UK EQUATOR Centre (also the EQUATOR Network’s head office), French EQUATOR Centre, Canadian EQUATOR Centre and Australasian EQUATOR Centre. The new centres will focus on national activities aimed at raising awareness and supporting adoption of good research reporting practices. They will work with partner organisations and initiatives and will also contribute to the work of the EQUATOR Network as a whole.

EQUATOR’s mission and goals

The EQUATOR mission is to achieve accurate, complete, and transparent reporting of all health research studies to support research reproducibility and usefulness. Our work increases the value of health research and helps to minimise avoidable waste of financial and human investments in health research projects.

To achieve its mission the EQUATOR Network has the following major goals:

- Maintain and further develop a comprehensive collection of online resources providing up-to-date information, tools and other materials related to health research reporting (Library for health research reporting)

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	
Economic evaluations	CHEERS	

Humans & Plurality Fix: Lifting Qualitative Methods

Behavioral software engineering - guidelines for qualitative studies

Per Lenberg[a,*], Robert Feldt[a], Lars Göran Wallgren Tengberg[b], Inga Tidefors[b], Daniel Graziotin[c]

[a] *Chalmers University of Technology*
[b] *Psychology Institution at Gothenburg University*
[c] *University of Stuttgart*

Abstract

Researchers are increasingly recognizing the importance of human aspects in software development and since qualitative methods are used to, in-depth, explore human behavior, we believe that studies using such techniques will become more common.

Existing qualitative software engineering guidelines do not cover the full breadth of qualitative methods and knowledge on using them found in the social sciences. The aim of this study was thus to extend the software engineering research community's current body of knowledge regarding available qualitative methods and provide recommendations and guidelines for their use.

https://arxiv.org/abs/1712.08341 rejected and in revision since mid 2018… ;)

Shameless plugs: Replications & Open Science

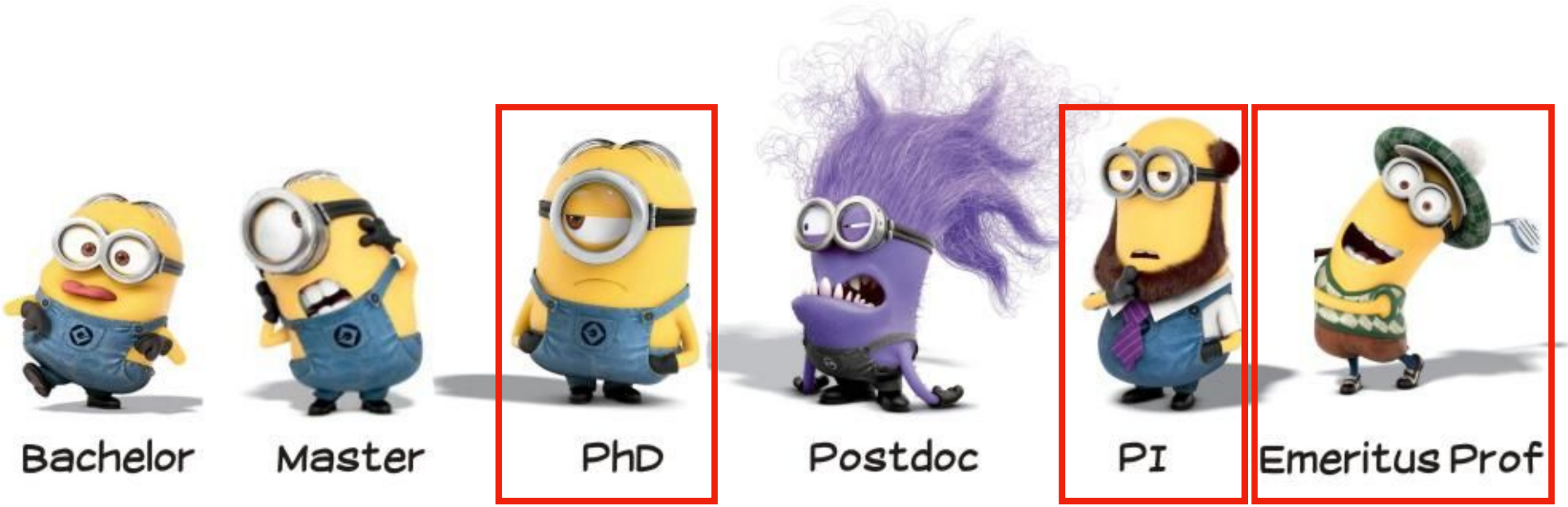
Empirical Software Engineering
https://doi.org/10.1007/s10664-019-09712-x

The open science initiative of the Empirical Software Engineering journal

Daniel Méndez Fernández[1] · Martin Monperrus[2] · Robert Feldt[3,4] · Thomas Zimmermann[5]

Published online: 02 May 2019

I'll throw in some Calls-for-action!



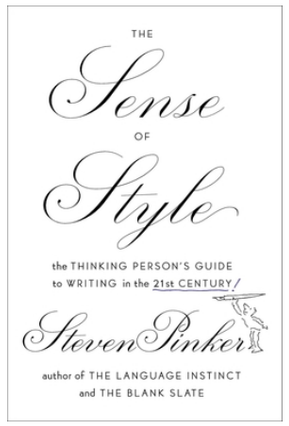
Call-for-action!

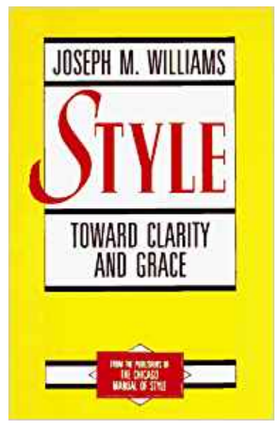
Remember why you went into science in 1st place

Seek truth & improve society. Don't fall for competition, politics, & the "numbers game".

Learn to write succinctly

Don't spread pseudo-profound bullshit.





Use diverse research methods

Broader knowledge base and equipped for pluralism & nuance.

Think deeply about actual threats to validity

Don't use as a "recipe" and "copy-n-paste".

Call-for-action!

Avoid “lamppost science”

Just because we have repositories, logs, and DBs doesn’t mean they have the information we truly need or should analyse.

Practice Open Science & try Pre-Registration

Don’t wait for venues; arXiv, GitHub, & zenodo are your friends.

Don’t preach “One paper, one message!” too strongly

Find balance between simplicity and shallow thinking / over-simplification.
Consider and discuss alternative explanations.

Raise the bar on statistical analysis

NHST is so 20th century. Causal analysis & Bayesian is the future.

Call-for-action!

Help create shared visions for the community

Multiple schools of thought ok, if clear & explicit and actively discussed.

Standardise quality checklists and guidelines

Help authors and peer reviewers. Build on what is there and adapt to ESE.

Stop the “numbers game”!

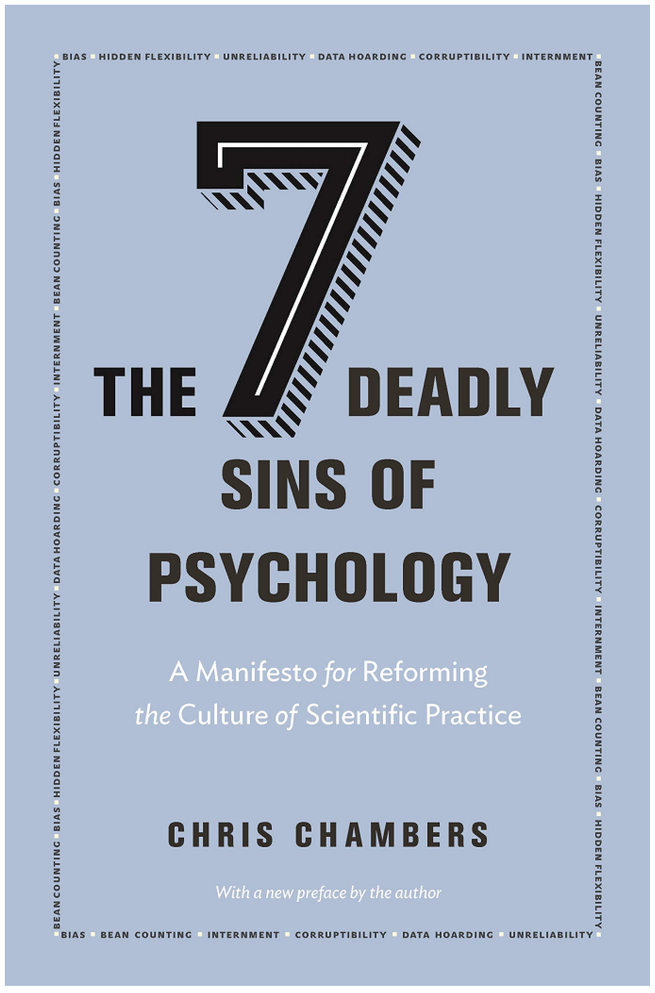
“Publish or Perish” can introduce bias that hinders truth.
Take responsibility in evaluations/promotions & discussion.

Continuous learning also from other fields

They know stuff. You’ll learn. Keep on learning & sharing.

Credits

“Replication is the immune system of science”
/ Prof. Chris Chambers:



Prof. Brian Nosek, Centre for Open Science & OSF

All my co-authors, colleagues and mentors!

The End

robert.feldt@chalmers.se