Reproducibility – principles and challenges

… some reflections for our discussion on scientometrics

Jesper W. Schneider

Danish Centre for Studies in Research and Research Policy,

Department of Political Science

Aarhus University, Denmark

jws@ps.au.dk

My points

- The concepts of “reproducibility” and “replication” are more ambiguous than we seem to imply through common sense understandings
- Research within scientometrics are to some degree very different from comparable fields that seem to have “reproducibility” problems
- But we do see some of the same practices in our field that have been identified as contributing to “reproducibility” problems
- Some changes in practice would help
- But most important, more openness and incentives for such is required, and here my emphasis is on “thick” methodological descriptions, assumptions and choices

Some beliefs about replication and reproducibility

Fits with logic and common sense about science

- “Science demands replication”
- If the description of some natural phenomenon fits the facts, then of course it will always occur again under the appropriate circumstances
- If something is “true”, then it’s always “true”
- The “truth” that science came to understand over the last few centuries did indeed follow from studies of repeatable observations and replicable experiments

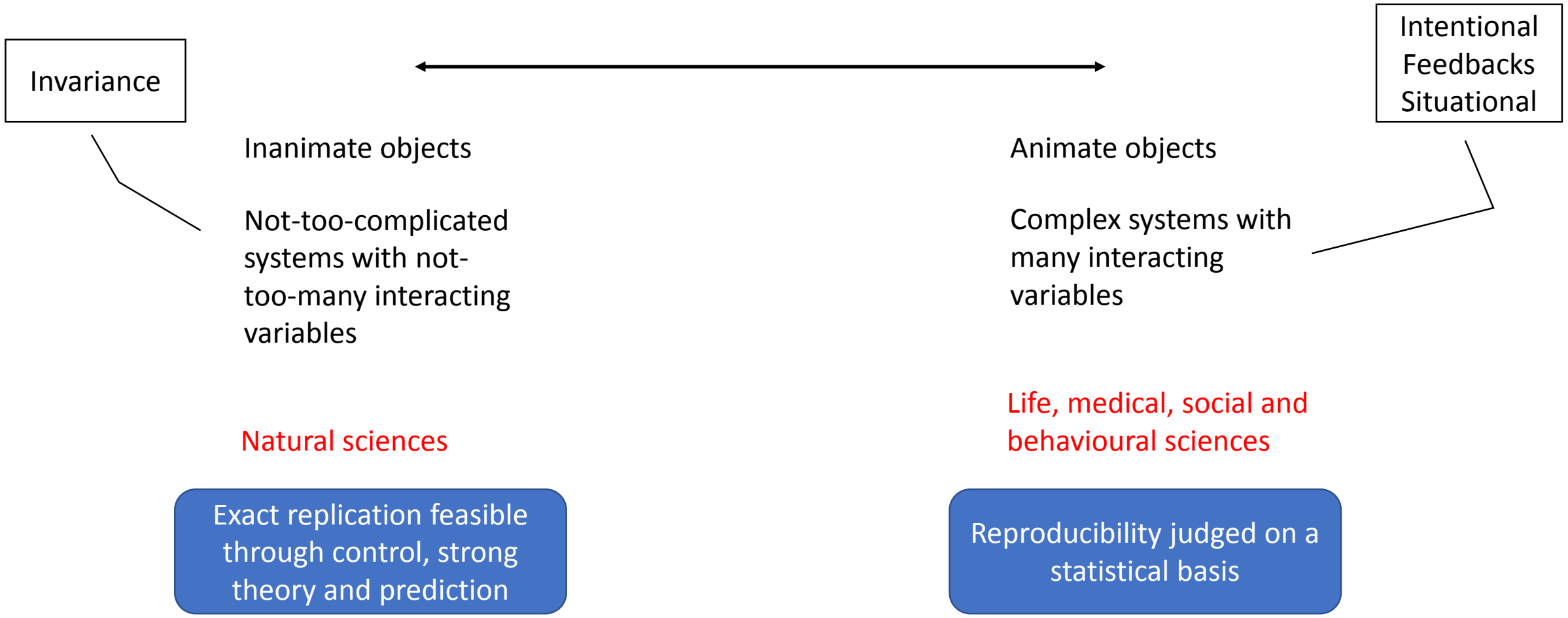
… and some beliefs about the nature of science

- That science is perpetually sceptical and doesn't form a belief until the evidence demands it
- That the scientific method ensures that science gets it right the first time because theories are accepted only after the evidence has shown them to be right
- That science is self-correcting because it changes its mind whenever the evidence demands it
  (this is incompatible with the first two points: self-correction is only needed if science doesn't get things right the first time)
- That "peer review" safeguards the objectivity and quality of science

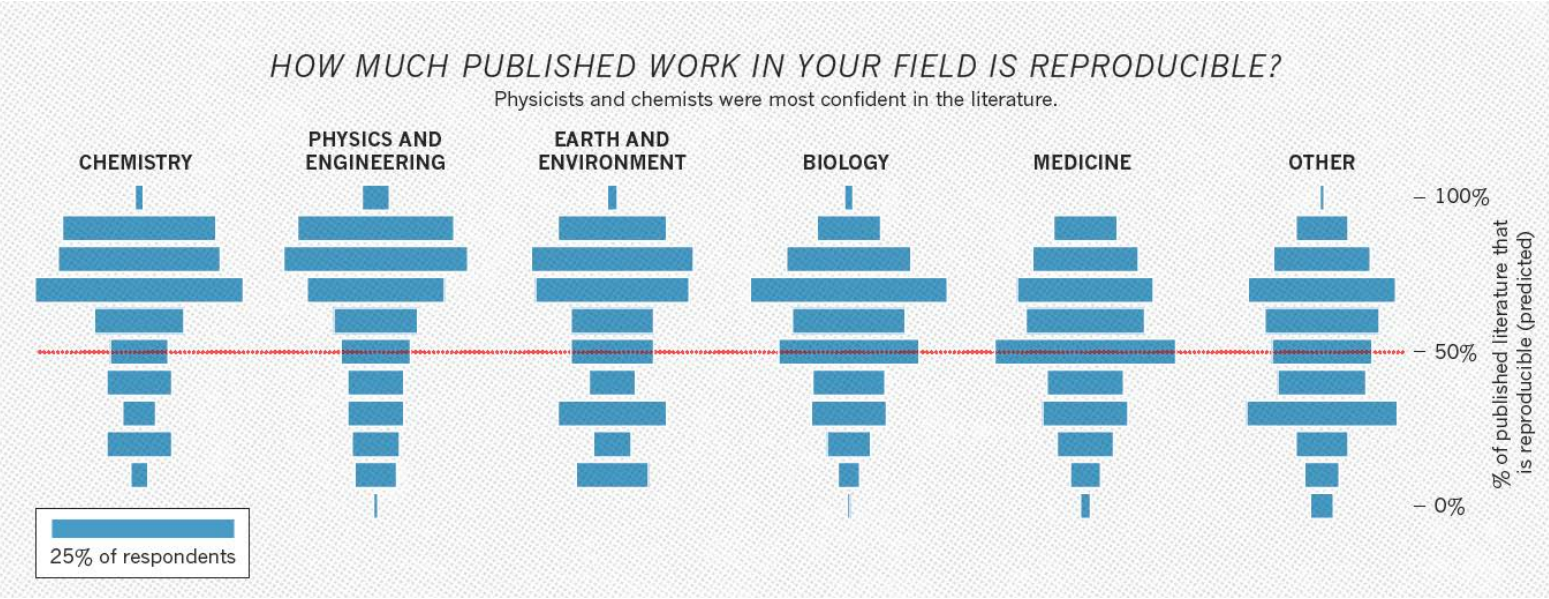
To some degree these are mistaken and they influence the way we perceive "reproducibility"

Some important nuances

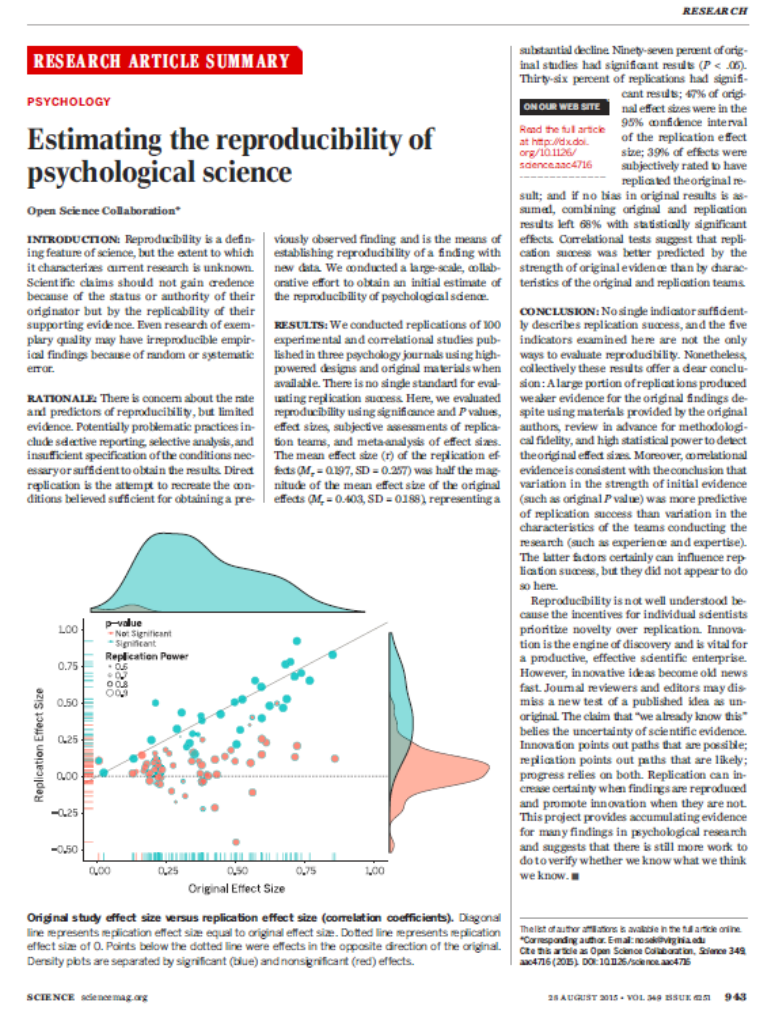
What is genuinely reproducible?



Perceptions of the reproducibility crisis



Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature, 533*(7604), 4 52-454.



RESEARCH

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (*r*) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac4716

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know. ■

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: nosek@virginia.edu
Cite this article as Open Science Collaboration, Science 349, aac4716 (2015). DOI: 10.1126/science.aac4716

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251)

Diagnosis: Mainly a problem in life, medical social and behavioural science!?

- Lack of strong theories to predict outcomes
- Lack of “truth”
- Lack of strong designs for control

“Intrinsic”

- Reliance on inferential statistics (mainly frequentist)
- Publication demand for “new” findings
- …

“Extrinsic”

Not much we can do about the “intrinsic” factors,
but we can do a lot about the “extrinsic” causes

All kinds of quantitative research?

Replicate or reproduce means that some kind of “truth” is needed

Exploratory

- Descriptive
- Dimensionality reduction
- Ordering

→ No prescription

No “ground truth”, not beyond data

Explanatory

- Confirmatory: “weak theory testing” [significance tests]
- Falsificationist: “strong theory testing” [specific predictions]

→ “truth” is parametrized and beyond data

- → Disprove H0 = 0
- → Corroborate H1 by numerically predict outcome to be between *a* and *b*, the shorter the interval the stronger the prediction

Explanatory/confirmatory research is what is causing the fuzz

- To much research is framed as “explanatory/confirmatory” when it is in fact exploratory
- The business of “confirming” weak, but also often rather obscure theories
- Negligence of “noise” coming from measurement issues
- An overreliance on $p < 0.05$ for “confirmation”, too many “positive” findings
- The $p < 0.05$ rule has been considered to be a safeguard against noise chasing and thus a guarantor of replicability

So an artificial “truth” is established and assumed and subsequently disproved statically

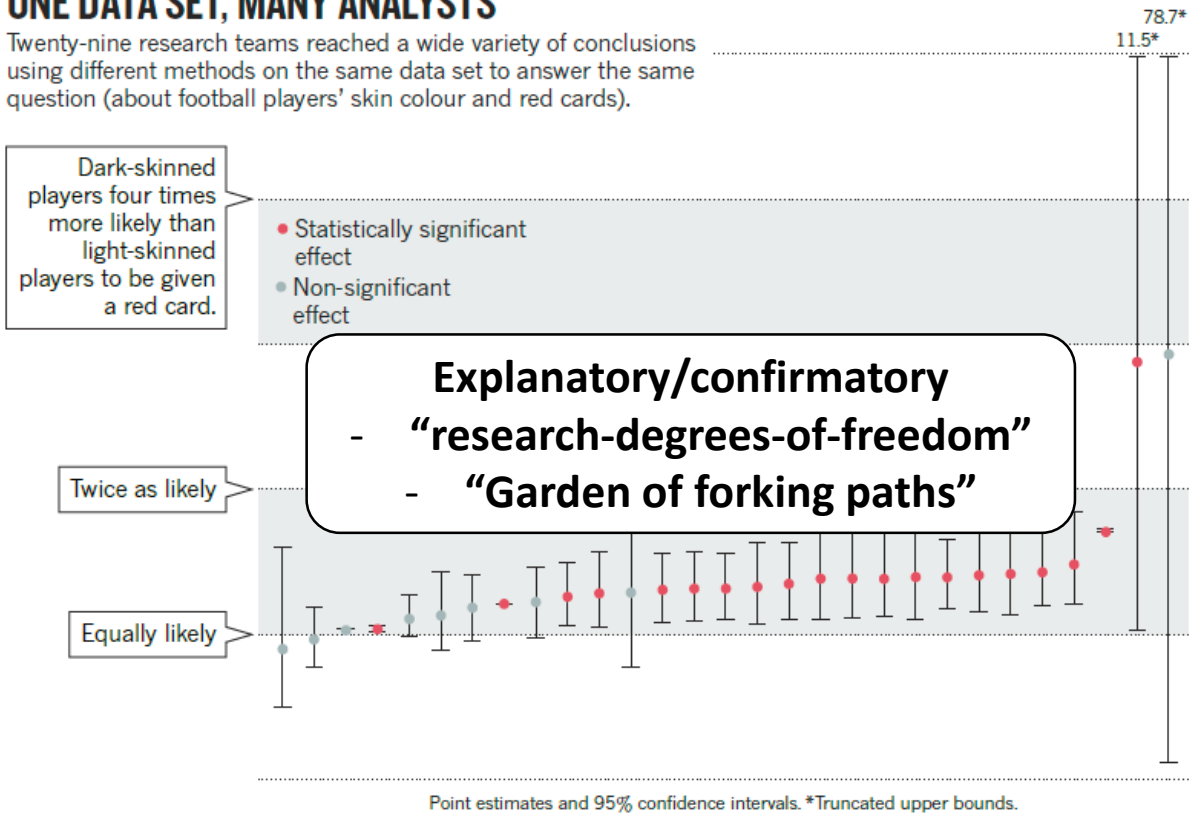
But an important issue is often overlooked

Choices in data processing and analysis that are contingent on data

- Statistical significance is a lot less meaningful than is traditionally assumed for many reasons, but two very important ones, the former also has consequence for exploratory studies are

- Abundant researcher “degrees of freedom” and “forking paths”which assure researchers a high probability of finding impressive p-values, even if all effects were zero and data were pure noise
- And if not documented makes any “reproducible” attempt hopeless

Same data, different results



Explanatory/confirmatory
- **"research-degrees-of-freedom"**
- **"Garden of forking paths"**

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature, 526*, 189-191.

Scientometrics (2017) 111:1223–1224
DOI 10.1007/s11192-017-2307-0

Topic identification challenge

Kevin Boyack[1] · Wolfgang Glänzel[2] · Jochen Gläser[3] · Frank Havemann[4] · Andrea Scharnhorst[5] · Bart Thijs[2] · Nees Jan van Eck[6] · Theresa Velden[3,7] · Ludo Waltmann[6]

Received: 6 June 2016 / Published online: 15 March 2017

Over the last two years, a group of researchers used a shared dataset in order to compare their approaches to the identification of thematic structures in a set of 111,616 papers on astronomy and astrophysics published in 59 journals between 2003 and 2010. The outcomes of this comparative exercise are published in a special issue of *Scientometrics*

✉ Jochen Gläser
Jochen.Glaser@ztg.tu-berlin.de

Nees Jan van Eck
ecknjpvan@cwts.leidenuniv.nl

Theresa Velden
velden@ztg.tu-berlin.de

Ludo Waltmann
waltmanlr@cwts.leidenuniv.nl

1 SciTech Strategies, Inc., Albuquerque, NM 87122, USA
2 ECOOM and Department of MSI, KU Leuven, Louvain, Belgium
3 ZTG, TU Berlin, HBS1, Hardenbergstr. 16-18, 10623 Berlin, Germany
4 Institut für Bibliotheks-und Informationswissenschaft, Humboldt-Universität zu Berlin, Dorotheenstr. 26, 10099 Berlin, Germany
5 DANS-KNAW, Anna van Saksenlaan 51, The Hague, The Netherlands
6 Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands
7 University of Michigan, School of Information, Ann Arbor, MI 48109, USA

Springer

Exploratory/ordering
- **"researcher-degrees-of-freedom"**

Boyack, K., Glänzel, W., Gläser, J., Havemann, F., Scharnhorst, A., Thijs, B., . . . Waltmann, L. (2017). Topic identification challenge. Scientometrics, 111(2), 1223-1224

So what about scientometrics?

Some characteristics

- Definitely complex systems
- Much more “exploratory”, but still with some explicit or implicit claims of describing “reality” (naïve realism)
- “Explanatory” scientometric research is not theory-driven, its instrumental or descriptive
  - obviously, reliance on statistical significance is an issue here
- The challenges – as I see it – is the “anything goes” approach
- Too many unwarranted choices and researcher-degrees-of freedom
- Problem if we: extrapolate our findings, 2) do not make our findings contingent, 3) if we do not provide sufficient information so that others can “reproduce”, 4) if we do not argue for our choices, 5) if we do not make assumptions clear, 6) if we neglect robustness and comparisons …

Thank you for your attention