Fixing Big Problems in Science

(as a process)

Understanding Political Numbers

April 29, 2019

Review

Humans doing science

Academia: biases toward novelty & significance for publication

Politics, policy, advocacy, industry: quest for "good enough" can sacrifice rigor or interpretability

Peer review isn't fool-proof

Publication bias (many findings are false or over-estimated)

p-hacking, "Dichotomania," Garden of Forking Paths, HARKing

Null hypothesis testing and the philosophy of science

The Statistical Crisis in Science

BY ANDREW GELMAN, ERIC LOKEN

Data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don't hold up.

How to improve?

Values

- Rigorous methods
- Replicability and reproducibility
- Transparency and honesty

How to improve?

Values

- Rigorous methods
- Replicability and reproducibility
- Transparency and honesty

Tactics

- Technology
- Institutions
- Culture

Technology

Transcription/Reporting Errors

The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nuijten[1] · Chris H. J. Hartgerink[1] · Marcel A. L. M. van Assen[1] · Sacha Epskamp[2] · Jelte M. Wicherts[1]

> "This study documents reporting errors in a sample of over 250,000 p-values reported in eight major psychology journals from 1985 until 2013, using the new R package `statcheck`. `statcheck` retrieved null-hypothesis significance testing (NHST) results from over half of the articles from this period. In line with earlier research, we found that half of all published psychology papers that use NHST contained at least one p-value that was inconsistent with its test statistic and degrees of freedom."

"Dynamic" Reporting (e.g. Rmarkdown)

Integrates code & writing; output matches analysis

Content control (similar to LaTeX)

(Get started or take a deep dive)

(Or, turn all your work into code)

````
```{r strata}
# number of respondents from each sample stratum
n_dane <- filter(res, geo_county == "Dane County") %>% nrow()
n_mke <- filter(res, geo_county == "Milwaukee County") %>% nrow()
n_rando <- filter(res, is.na(geo_county)) %>% nrow()
```

A total of `r nrow(res)` surveys were returned, with `r n_dane`
respondents from Dane County, `r n_mke` from Milwaukee County,
and `r n_rando` whose home counties could not be identified.
Because oversampling was performed within Census tracts, the `r
  n_rando` respondents with unknown geographic locations were
assigned no weights and thus excluded from the analysis,
resulting in an effective sample of `r nrow(res) - n_rando`
valid responses. This gives us a response rate of `r number((nrow
  (res) - n_rando) / 2400, scale = 100, accuracy = .1)` percent
before adjusting for deadwood in the sample..
````

"Dynamic" Reporting (e.g. Rmarkdown)

Integrates code & writing; output matches analysis

Content control (similar to LaTeX)

(Get started or take a deep dive)

(Or, turn all your work into code)

````
```{r strata}
# number of respondents from each sample stratum
n_dane <- filter(res, geo_county == "Dane County") %>% nrow()
n_mke <- filter(res, geo_county == "Milwaukee County") %>% nrow()
n_rando <- filter(res, is.na(geo_county)) %>% nrow()
```

A total of `r nrow(res)` surveys were returned, with `r n_dane`
respondents from Dane County, `r n_mke` from Milwaukee County,
and `r n_rando` whose home counties could not be identified.
Because oversampling was performed within Census tracts, the `r
  n_rando` respondents with unknown geographic locations were
assigned no weights and thus excluded from the analysis,
resulting in an effective sample of `r nrow(res) - n_rando`
valid responses. This gives us a response rate of `r number((nrow
  (res) - n_rando) / 2400, scale = 100, accuracy = .1)` percent
before adjusting for deadwood in the sample..
````

A total of 293 surveys were returned, with 75 respondents from Dane County, 213 from Milwaukee County, and 5 whose home counties could not be identified. Because oversampling was performed within Census tracts, the 5 respondents with unknown geographic locations were assigned no weights and thus excluded from the analysis, resulting in an effective sample of 288 valid responses. This gives us a response rate of 12.0 percent before adjusting for deadwood in the sample..

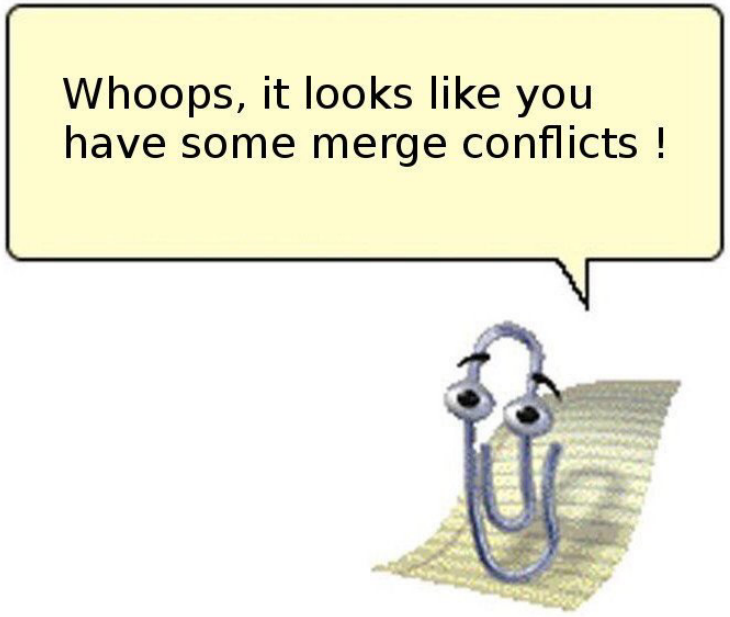
Git

Version control (like "track changes" but for code)

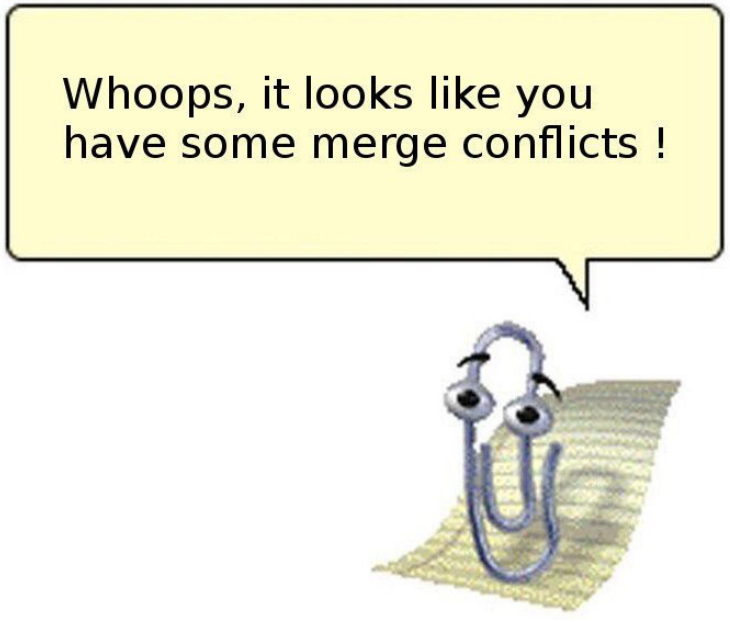
Edit code incrementally, "commit" changes

Rewind project history, branch off, merge branches

VERY important for data science (collaboration, complex projects)



Git



Version control (like "track changes" but for code)

Edit code incrementally, "commit" changes

Rewind project history, branch off, merge branches

VERY important for data science (collaboration, complex projects)

Looking at a file's history

```
224  224
225      - This text got deleted
     225 +
     226 + This text got added
226  227
```

Github: online Git repositories

Collaboration and distribution of open-source projects

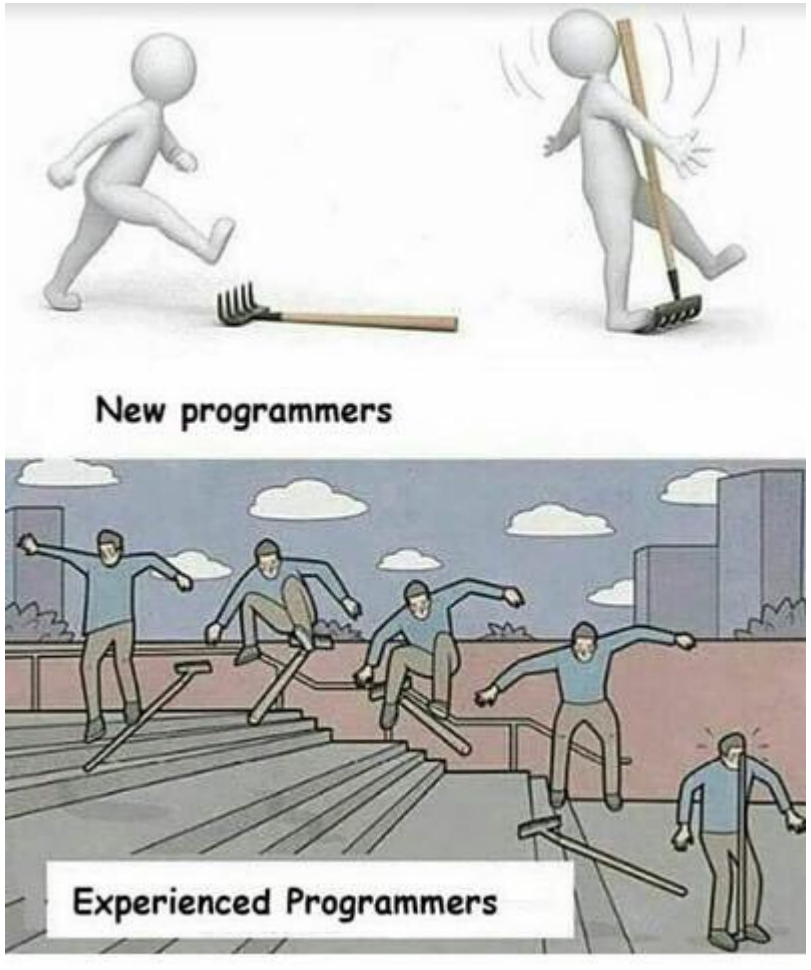
Github repository for ps-270

More Git resources:

- Excuse me, do you have a moment to talk about version control?
- Happy Git and GitHub for the useR

Technology, not a magic bullet

It can't make you *want* to do good science



Statistical(?) Technology

Larger samples (statistical power)

Causal inference with potential outcomes

- Minimal requisite assumptions to identify "treatment effect"
- Unit i receives treatment ($z = 1$) or control ($z = 0$)
- $y(z)$: outcome value given z
- Treatment effect is $y(z = 1)_i - y(z = 0)_i$

Bayesian statistical analysis

Big $ in your future

USING FULL PROBABILITY MODELS TO COMPUTE PROBABILITIES OF ACTUAL INTEREST TO DECISION MAKERS

Frank E. Harrell, Jr.
University of Virginia School of Medicine

Ya-Chen Tina Shih
MEDTAP International Inc.

Abstract

The objective of this paper is to illustrate the advantages of the Bayesian approach in quantifying, presenting, and reporting scientific evidence and in assisting decision making. Three basic components in the Bayesian framework are the prior distribution, likelihood function, and posterior distribution. The prior distribution describes analysts' belief *a priori*; the likelihood function captures how data modify the prior knowledge; and the posterior distribution synthesizes both prior and likelihood information. The Bayesian approach treats the parameters of interest as random variables, uses the entire posterior distribution to quantify the evidence, and reports evidence in a "probabilistic" manner. Two clinical examples are used to demonstrate the value of the Bayesian approach to decision makers. Using either an uninformative or a skeptical prior distribution, these examples show that the Bayesian methods allow calculations of probabilities that are usually of more interest to decision makers, e.g., the probability that treatment A is similar to treatment B, the probability that treatment A is at least 5% better than treatment B, and the probability that treatment A is not within the "similarity region" of treatment B, etc. In addition, the Bayesian approach can deal with multiple endpoints more easily than the classic approach. For example, if decision makers wish to examine mortality and cost jointly, the Bayesian method can report the probability that a treatment achieves at least 2% mortality reduction and less than $20,000 increase in costs. In conclusion, probabilities computed from the Bayesian approach provide more relevant information to decision makers and are easier to interpret.

Institutional Reform

What are institutions?

Formal and informal structures for behaviors, customs, rules, processes

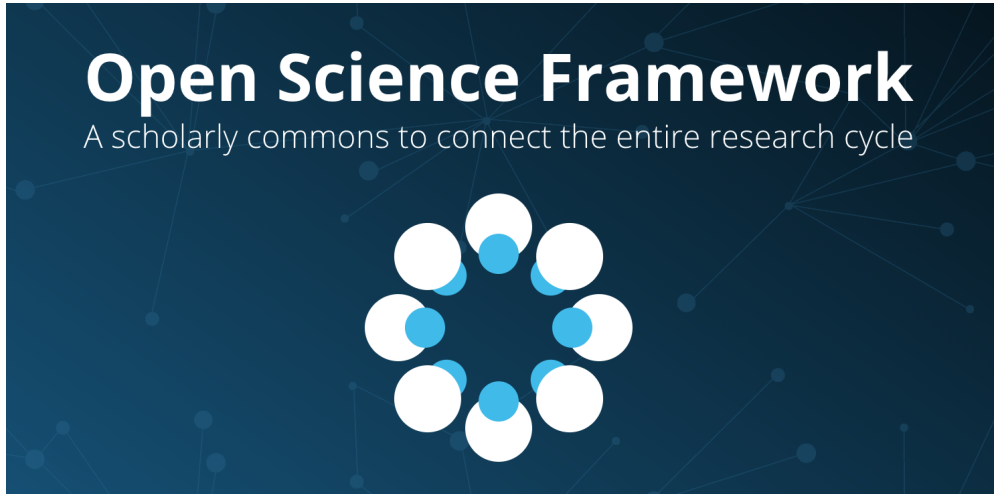
What is a collective action problem?

Scientific work is like a *market*. How do you change market behavior?

"Open Science"

Pre-analysis planning (hypotheses and analysis)

Results-blind peer review (Pre-acceptance)



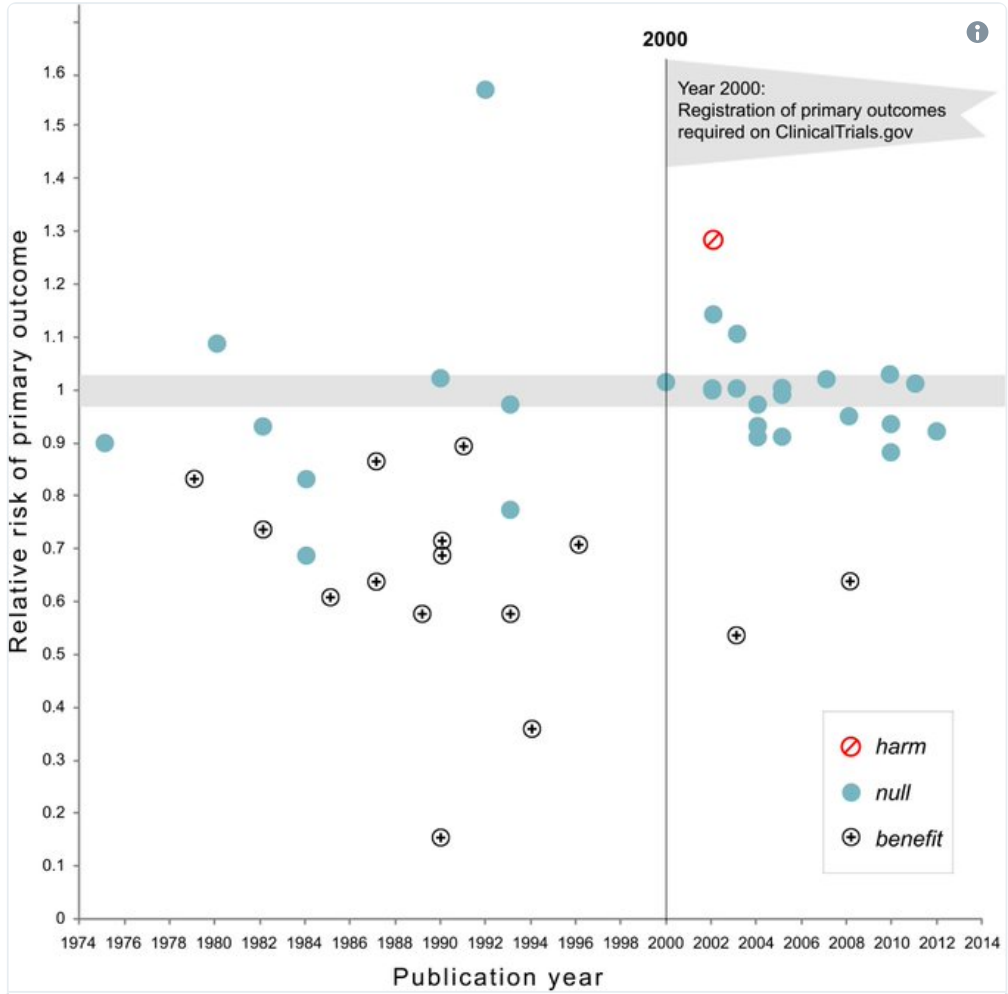
Journal reforms

Open Access

Short papers

Replication data and code ("Data Access and Research Transparency" Initiative)

But...exploratory vs confirmatory research?



John B. Holbein
@JohnHolbein1

Figure of the Day: results of clinical trials before and after preregistration of study design became a requirement.

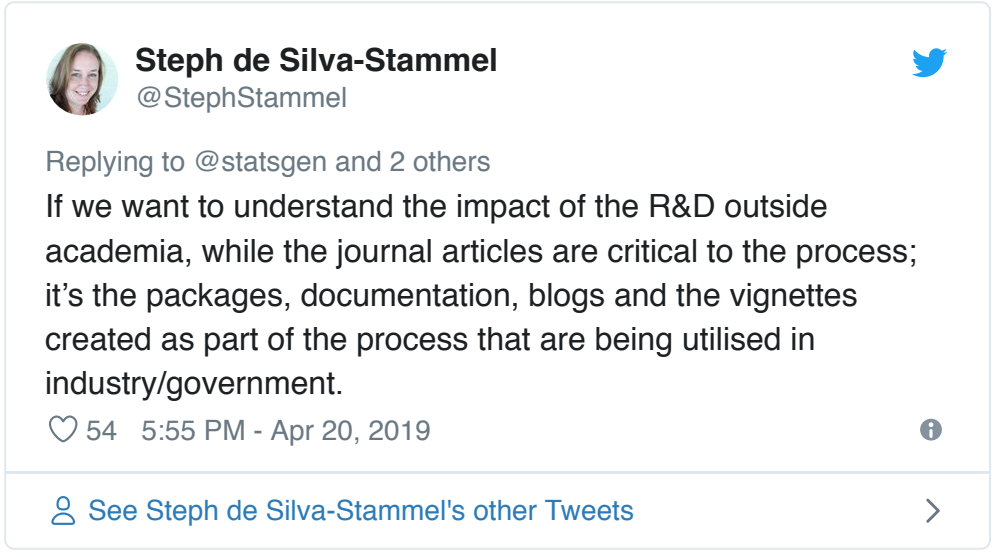
Source: journals.plos.org/plosone/articl…

Tenure reform

Address the *career incentives*: academic freedom and job security

What "output" counts toward tenure?

Journal publications vs software/data contributions, public scholarship, disciplinary service



Cultural Change

Can't just ask nicely

Replication, still largely a thankless task

Give people means to change (technology)

Give people motivation to change (institutions)

It's on you to do the right thing

No, it’s not The Incentives—it’s you

There’s a narrative I find kind of troubling, but that unfortunately seems to be growing more common in science. The core idea is that the mere existence of perverse incentives is a valid and sufficient reason to knowingly behave in an antisocial way, just as long as one first acknowledges the existence of those perverse incentives. The way this dynamic usually unfolds is that someone points out some fairly serious problem with the way many scientists behave—say, our collective propensity to p-hack as if it’s going out of style, or the fact that we insist on submitting our manuscripts to publishers that are actively trying to undermine our interests—and then someone else will say, “I know, right—but what are you going to do, those are the *incentives*.”

Yarkoni, "It's not the Incentives'"

In conclusion, data analysis is hard

In conclusion, data analysis is hard

Science, math, statistics, ethics

In conclusion, data analysis is hard

Science, math, statistics, ethics

Not to mention...subject-matter expertise

In conclusion, data analysis is hard

Science, math, statistics, ethics

Not to mention...subject-matter expertise

Luckily, it's fun to learn & challenge yourself