# Transparency and Data Management Workshop

Maarten Voors & Paul Hofman

3 December 2018

Wageningen University

# Workshop

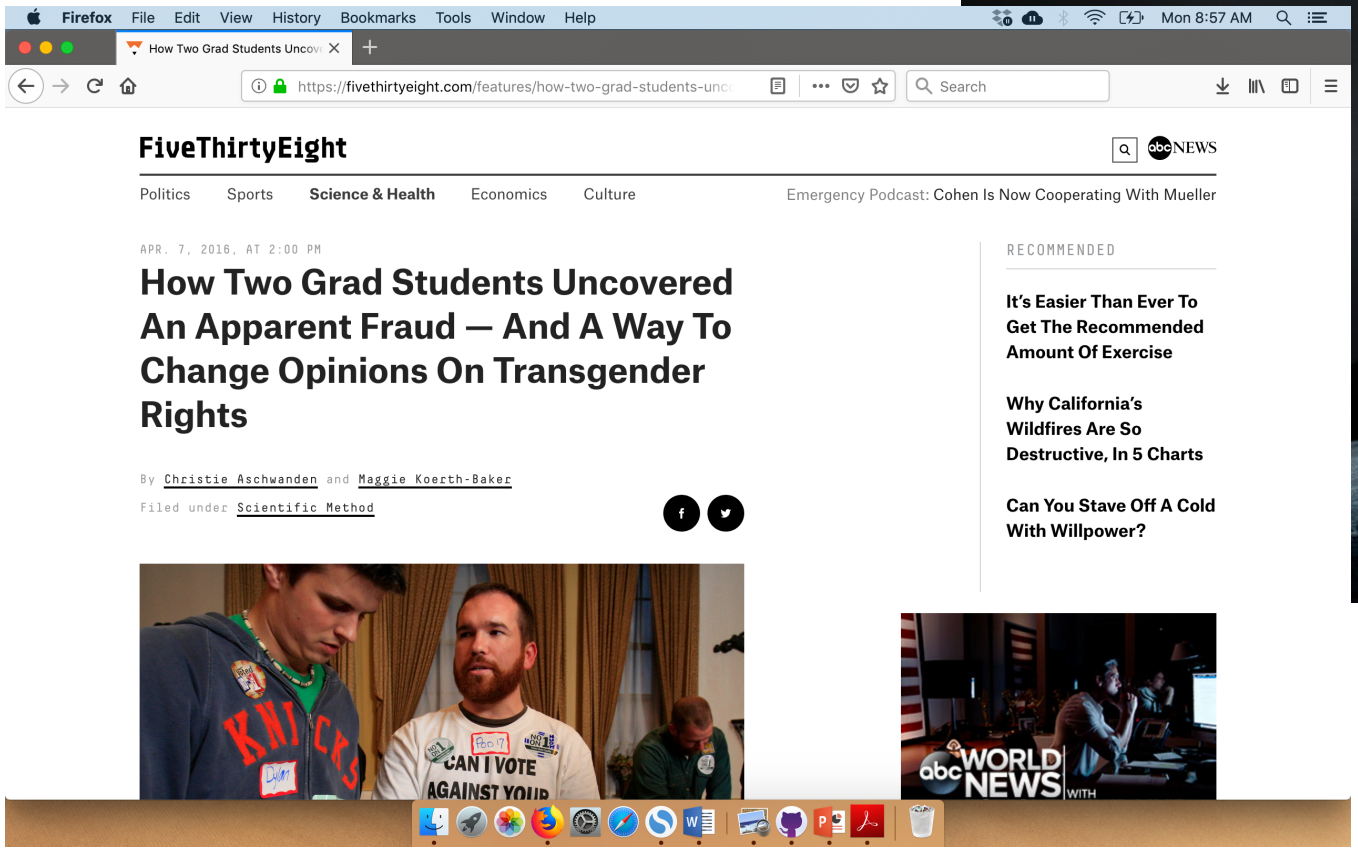
- Objective:
  - Share perspectives, experiences and best practices
  - Principles: efficiency and normative
- Agenda
  - 1230-1315 What is it? And why should we care? (Maarten)
  - 1330-1400 Workflow Manual (Maarten)
  - 1415-1530 Tools (Paul)

# Discuss

- When you read a paper what are the things you look for to check of you believe the results?

- Do you know colleagues who tried to manipulate results?

# Dishonesty Happens



# Why we worry

- There are many claims and no evidence

- Results not **replicable**
- Results not **stable**
- Results not **credible**
- Results do not **cumulate**

- Possibly especially true for social interventions and studies

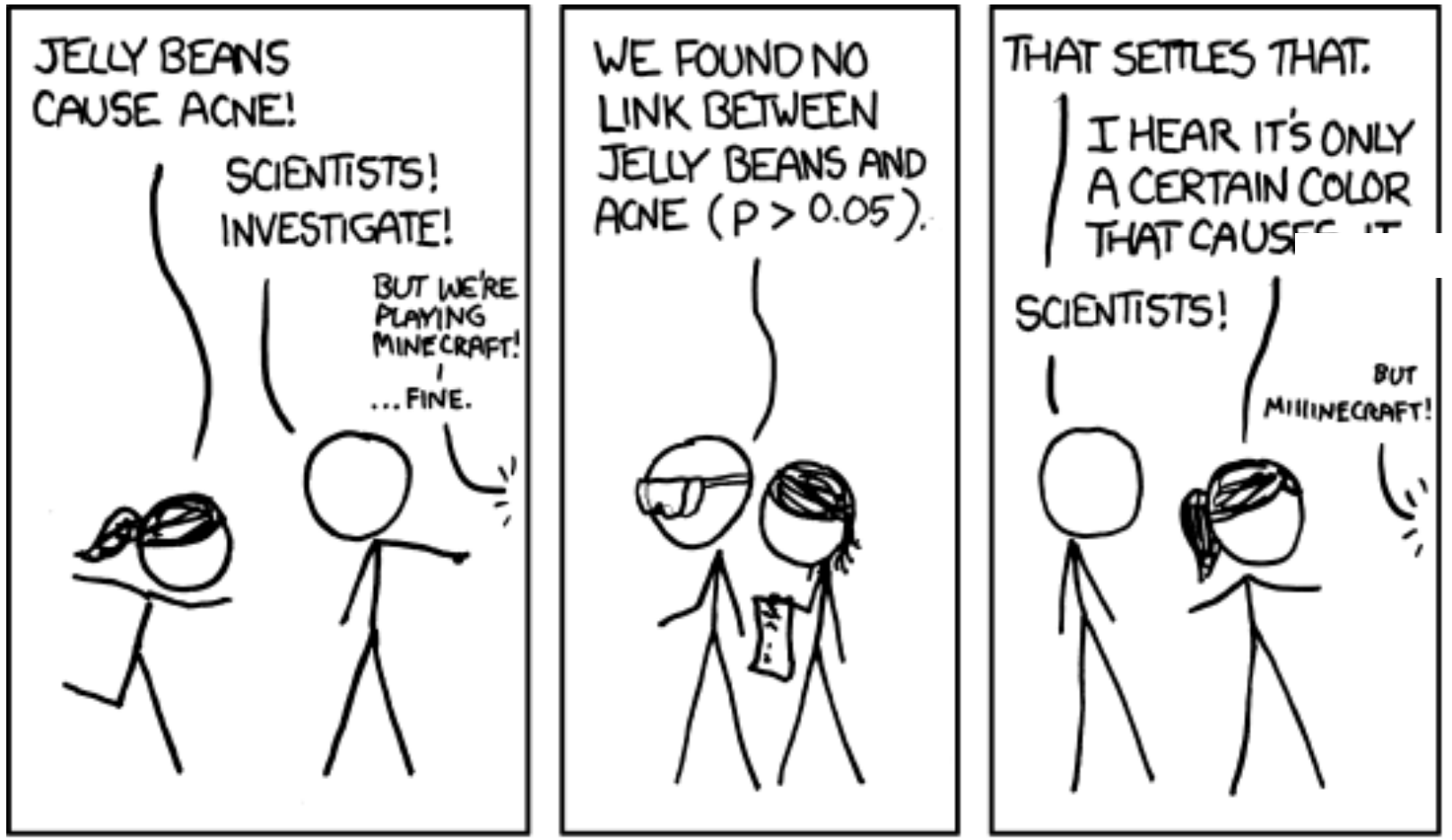
# Many Possible Reasons

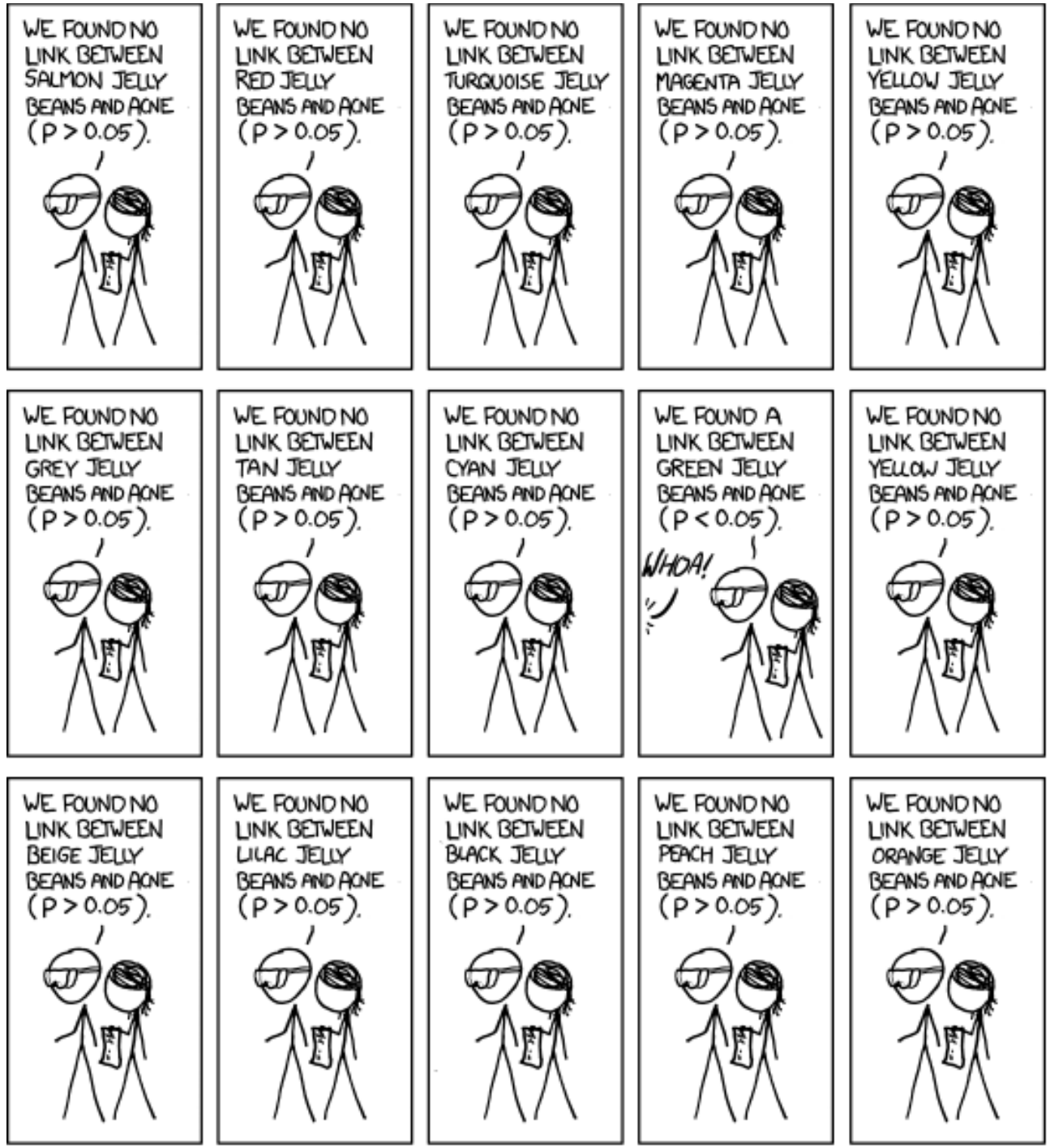
- Bad Designs
- Bad Analyses
- Bad Reporting
- Bad Promotion
- Error and Fraud

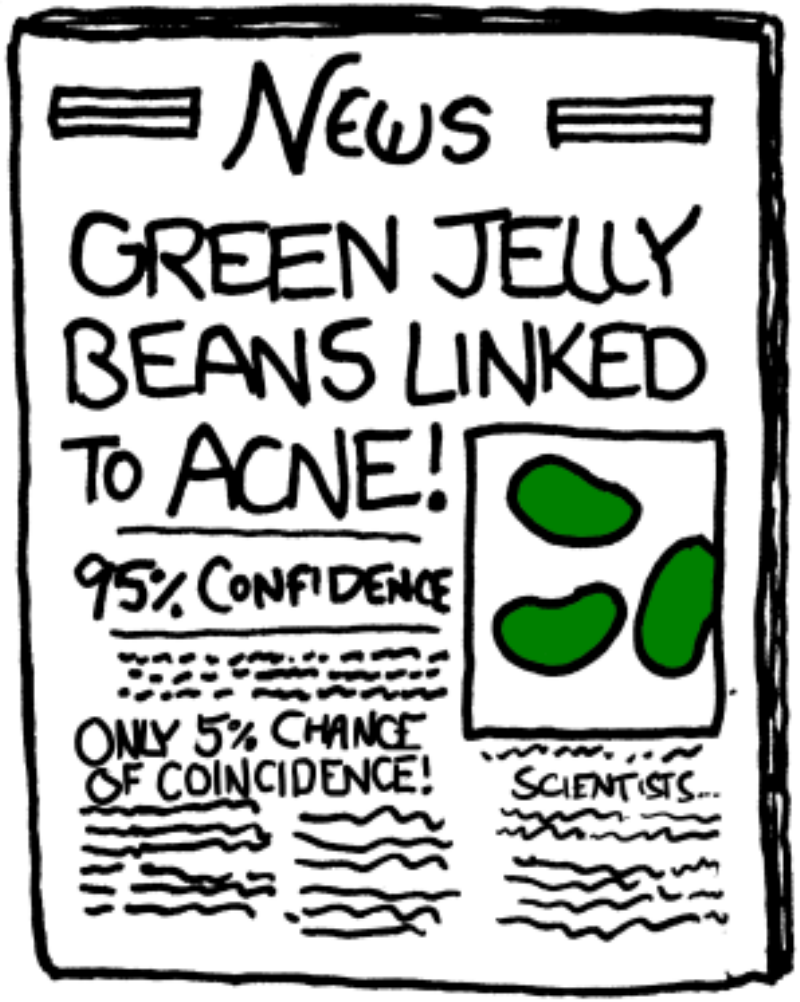
# What we find

Weak academic norms can distort the body of evidence.

- Publication bias (“file drawer” problem)
- p-hacking
- Non-disclosure
- Selective reporting
- Failure to replicate


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT
SCIENTISTS!
BUT MIIIINECRAFT!

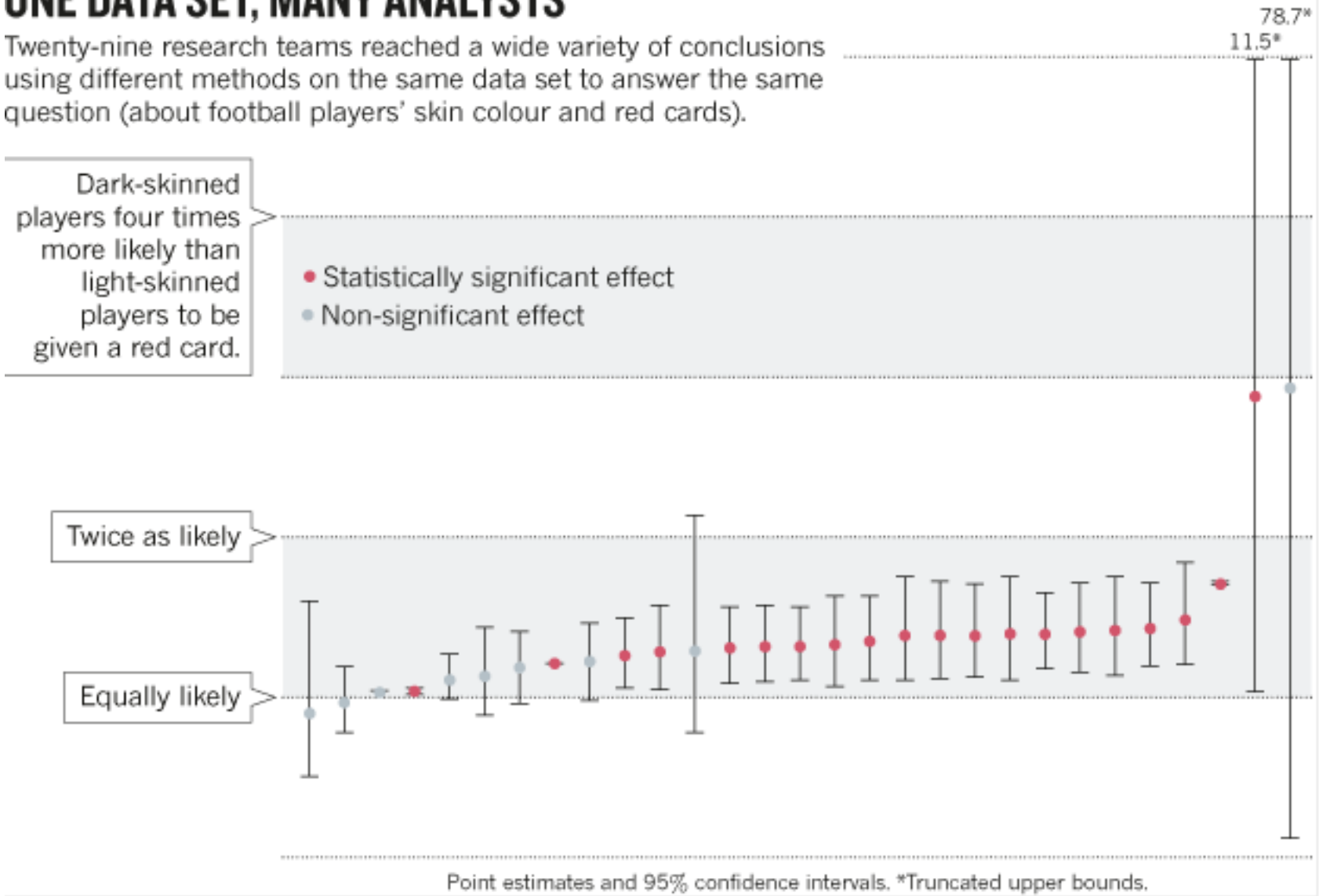

WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).




- Credit: http://www.xkcd.com/

# ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).




Point estimates and 95% confidence intervals. *Truncated upper bounds.

# What is a p-value?

# With what you know about p values…

- If I run 100 regressions or ttests
- How many should I expect to produce significant results at 5%?

# Solution: Increasing transparency

- Transparency is part of being an ethical researcher.

- Scientific values espoused by Robert Merton (Merton 1942):
  - Universalism: anyone can make a claim regardless of status.
  - Communism: open sharing of knowledge.
  - Disinterestedness: truth as motivation, not financial (or other) gains.
  - Organized skepticism: peer review, replication.

# Ecosystem for Open Science



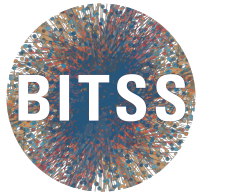











# Publication Bias

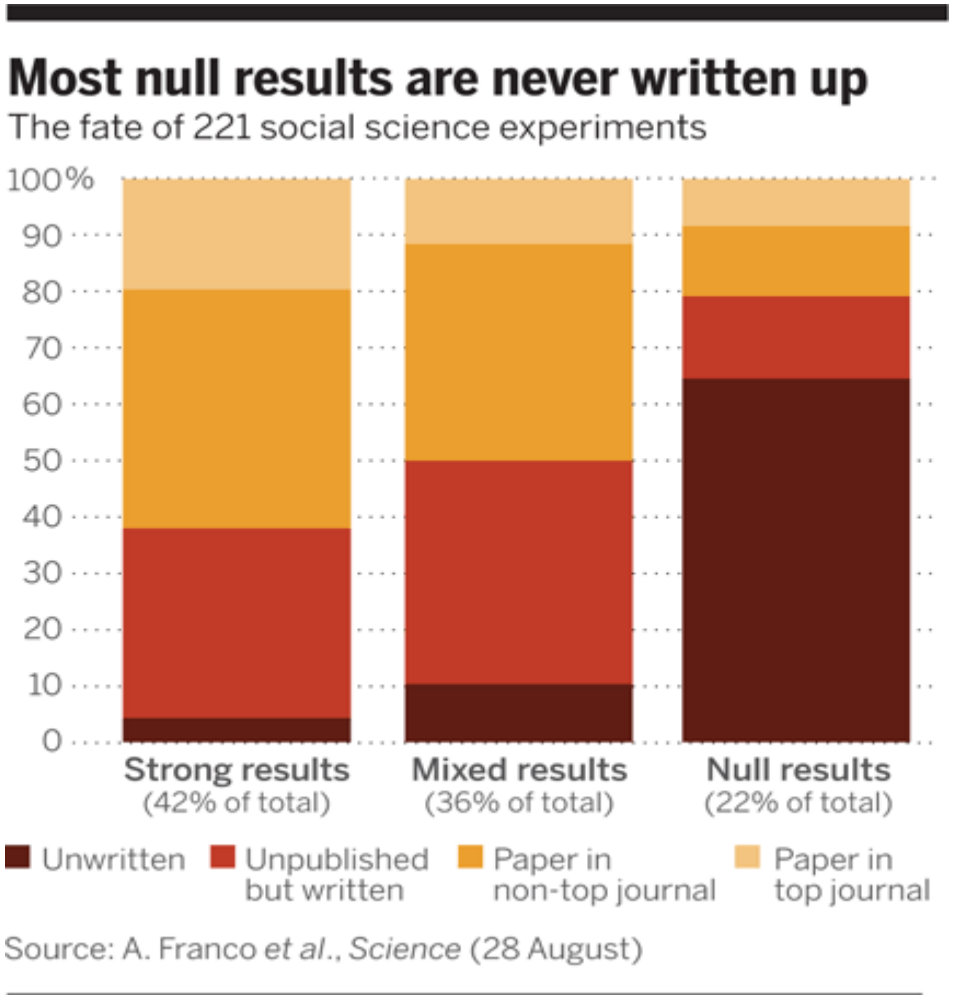
*“File drawer problem”*

# Publication Bias

- Status quo: Null results are not as “interesting.”
  - What if you find no relationship between a school intervention and test scores? (in a well-designed study…)
    - *It’s less likely to get published, so null results are hidden.*

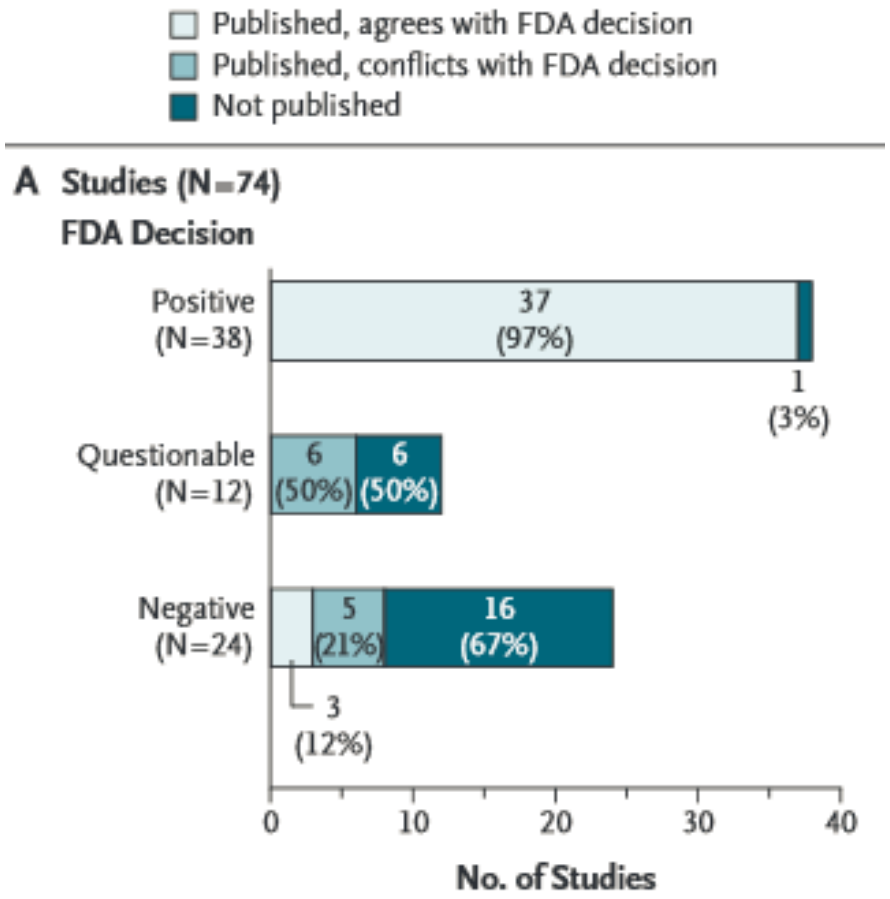
- How do we know? Rosenthal 1979:
  - Published: 3 published studies, all showing a positive effect…
  - Hidden: A few unpublished studies showing null effect
    - *The significance of positive findings is now in question!*

# In social sciences…




[Franco, Malhotra, Simonovits (2014)]

# In medicine…




ClinicalTrials.gov

# P-curves

- Scientists want to test hypotheses
  - i.e. look for relationships among variables (schooling, test scores)
  - Observed relationships should be statistically significant
    - Minimize the likelihood that an observed relationship is actually a false discovery
    - Common norm: p-value < 0.05

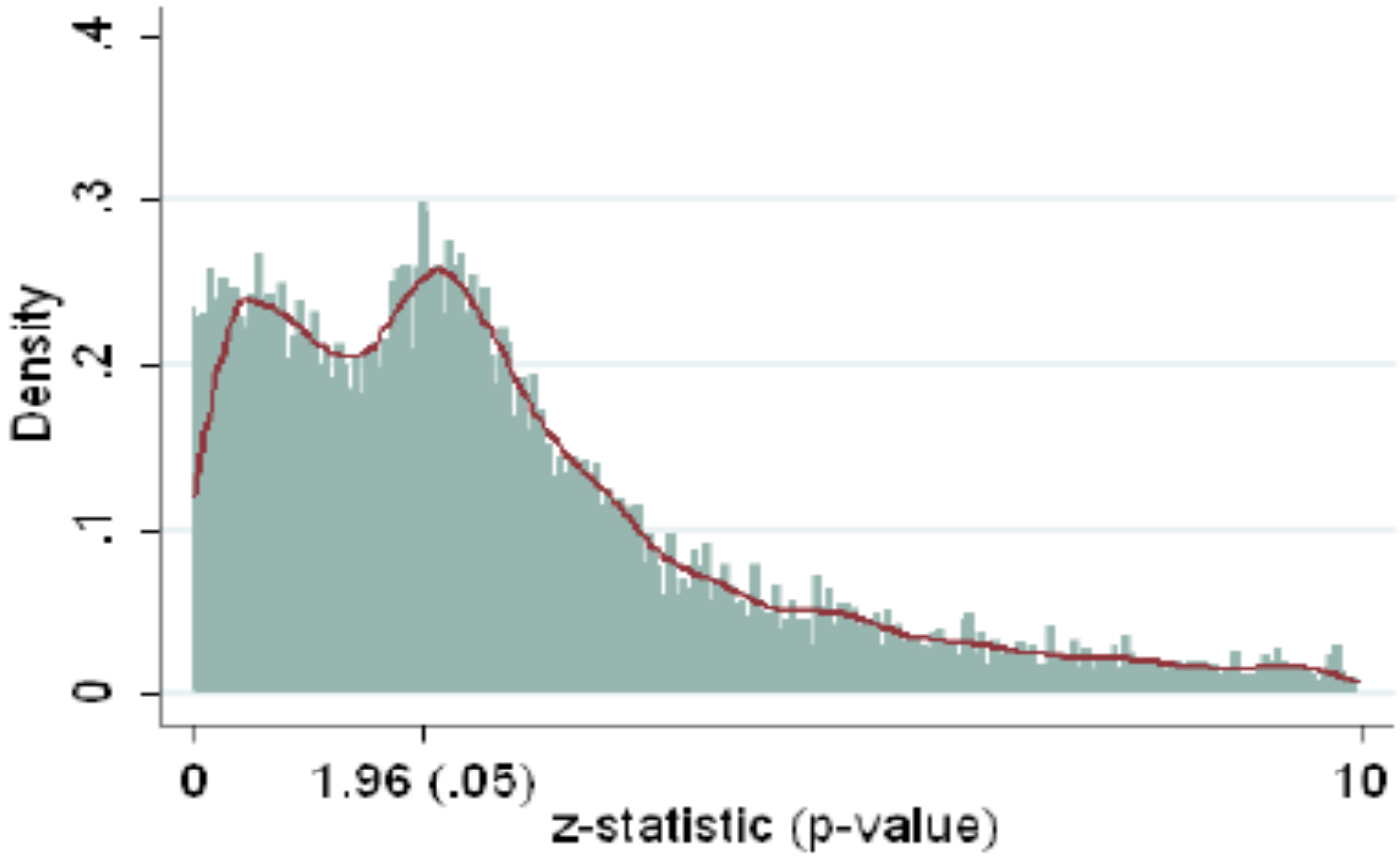
But null results not “interesting” ...

*So incentive is to look for (or report) the positive effects,*

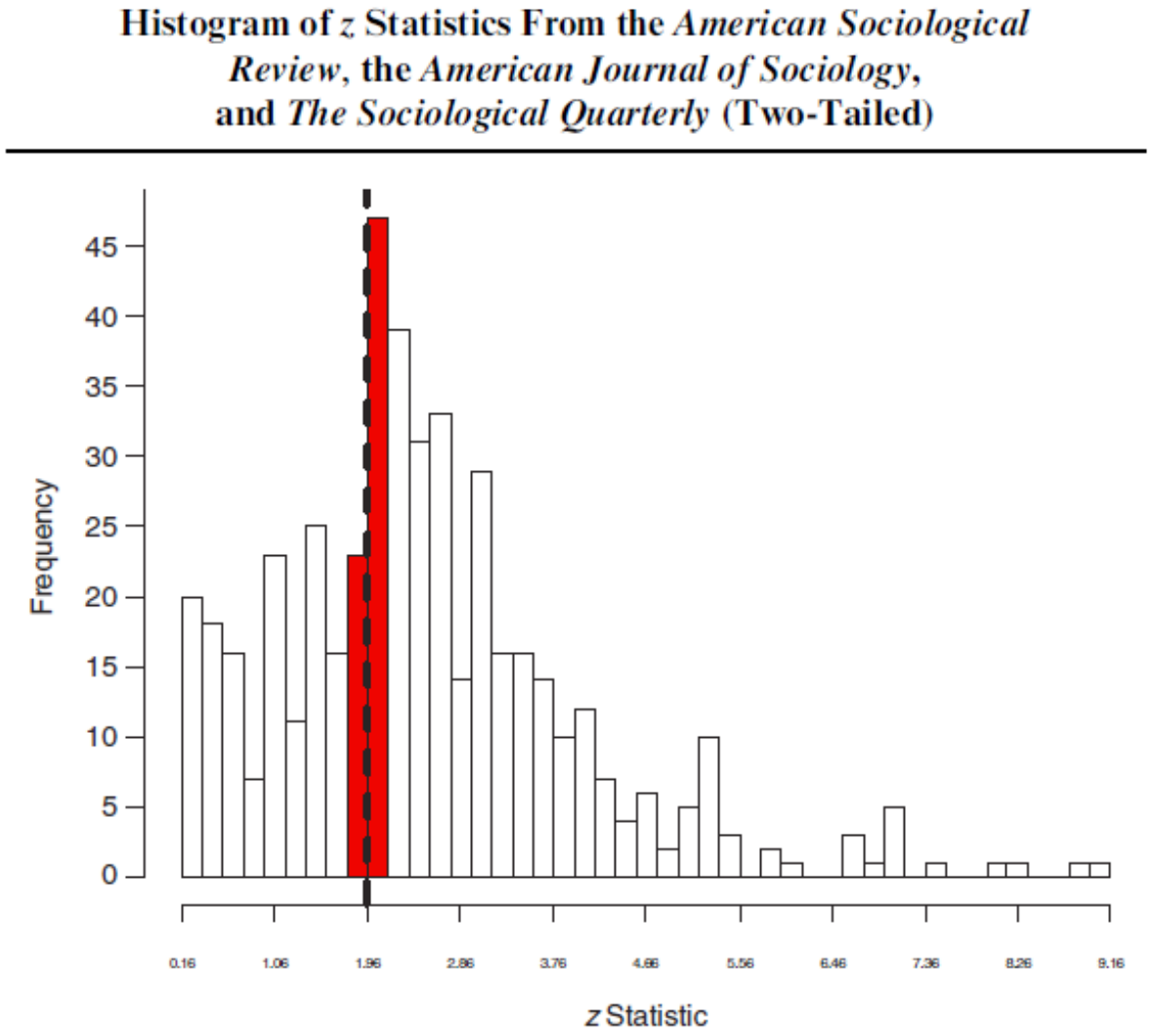
*even if they’re false discoveries*

# In economics

(b) Unrounded distribution of z-statistics.



# In sociology



# In political science…

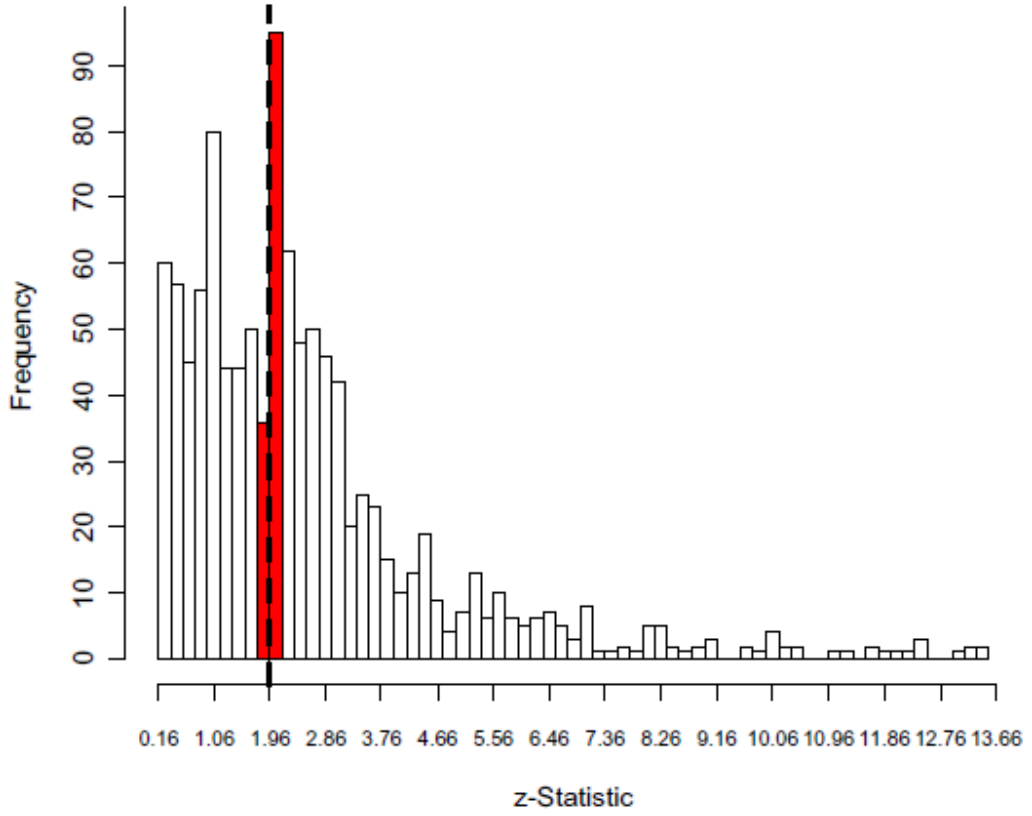



Figure 1(a). Histogram of $z$-statistics, *APSR & AJPS* (Two-Tailed). Width of bars (0.20) approximately represents 10% caliper. Dotted line represents critical $z$-statistic (1.96) associated with $p = 0.05$ significance level for one-tailed tests.

"What would you prefer to see first, Minister, the actual figures,
or the gloss we've put on the figures?"

- If we only write up/publish significant results, and we have no record of all the insignificant results, we have no way to tell if our ‘significant’ results are real, or if they’re the 5% we should expect due to noise.

# Solution: Design Peer Review

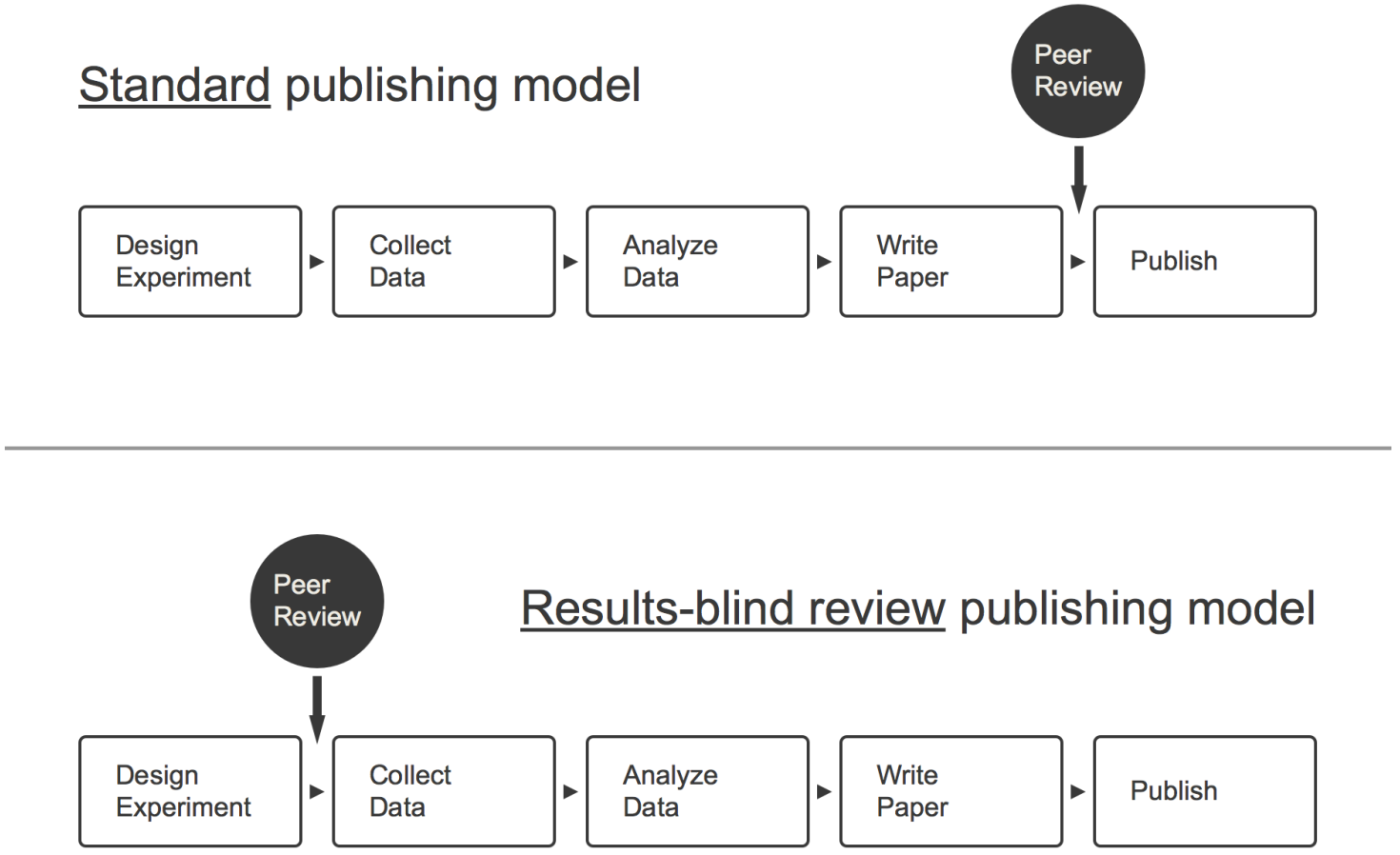
Prospectively register hypotheses in a public database.

“Paper trail” to solve the “File Drawer” problem.

Differentiate *confirmatory* hypothesis testing from *exploratory.*

- Medicine & Public Health: clinicaltrials.gov
- **Economics: AEA registry: socialscienceregistry.org**
- **Political Science: EGAP Registry: http://egap.org/content/registration**
- Development: 3IE Registry: ridie.3ieimpact.org
- Open Science Framework: http://osf.io

# Solution: Design Peer Review



# Solution: Design Peer Review

- Journal of Development Economics (JDE),
- Pilot of a “pre-results review”: submit work for peer review before there are any empirical results.
  - Stage 1, authors submit a proposal that includes an introduction, methods, and data analysis plan for a prospective study.
  - High-quality studies are accepted based on pre-results review, which constitutes a commitment by the JDE to publish the upcoming paper, regardless of the results.
  - Authors can then collect and analyze the data and submit a full paper for final review and publication
  - Stage 2 review makes sure that the final paper is aligned with the research design accepted in Stage 1.

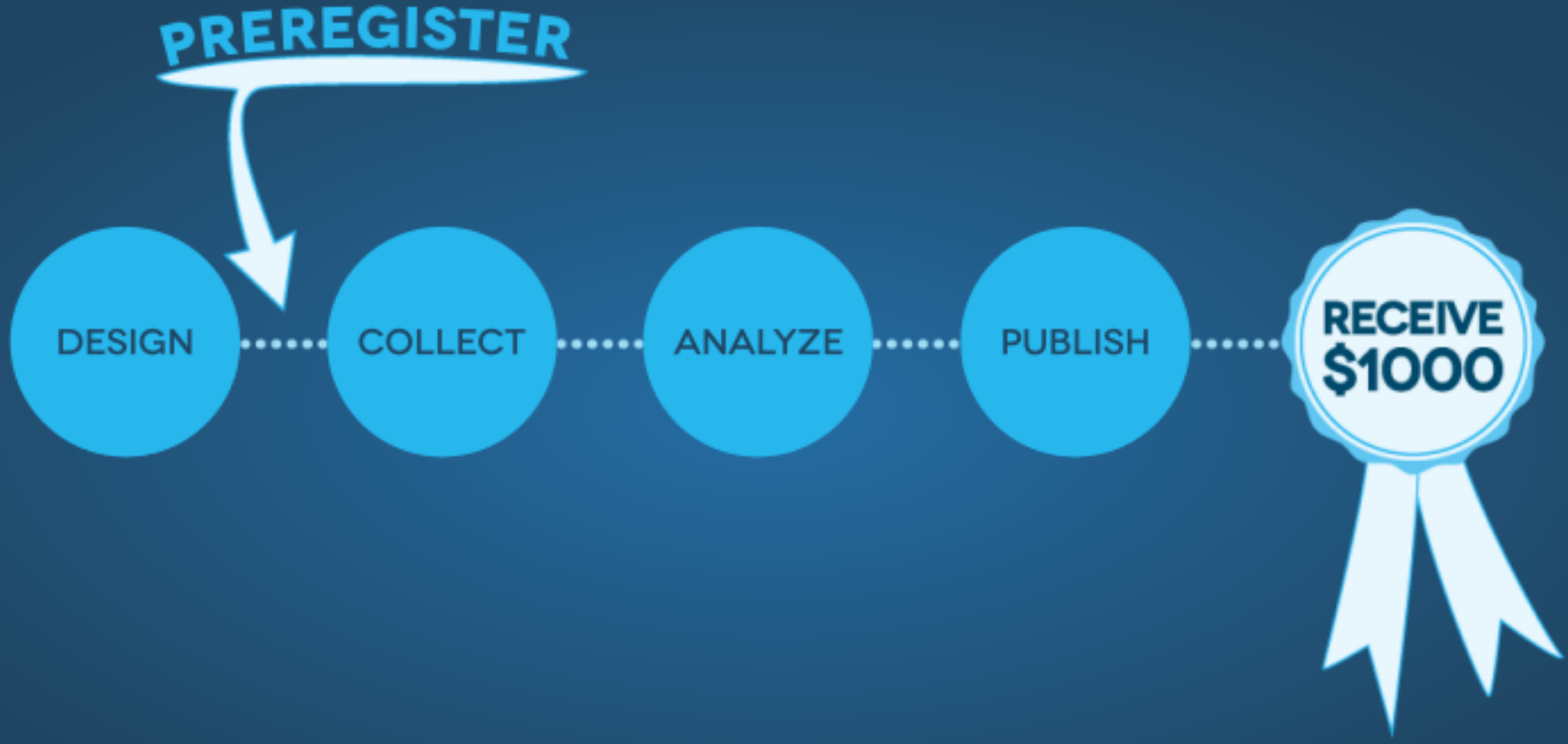
- https://www.elsevier.com/__data/promis_misc/JDE_RR_Author_Guidelines.pdf

# Pre-registration

- Have a look:
  - http://egap.org/content/registration
  - https://www.socialscienceregistry.org/

- Very useful repository of possible designs

CENTER FOR OPEN SCIENCE




# THE PREREGISTRATION CHALLENGE

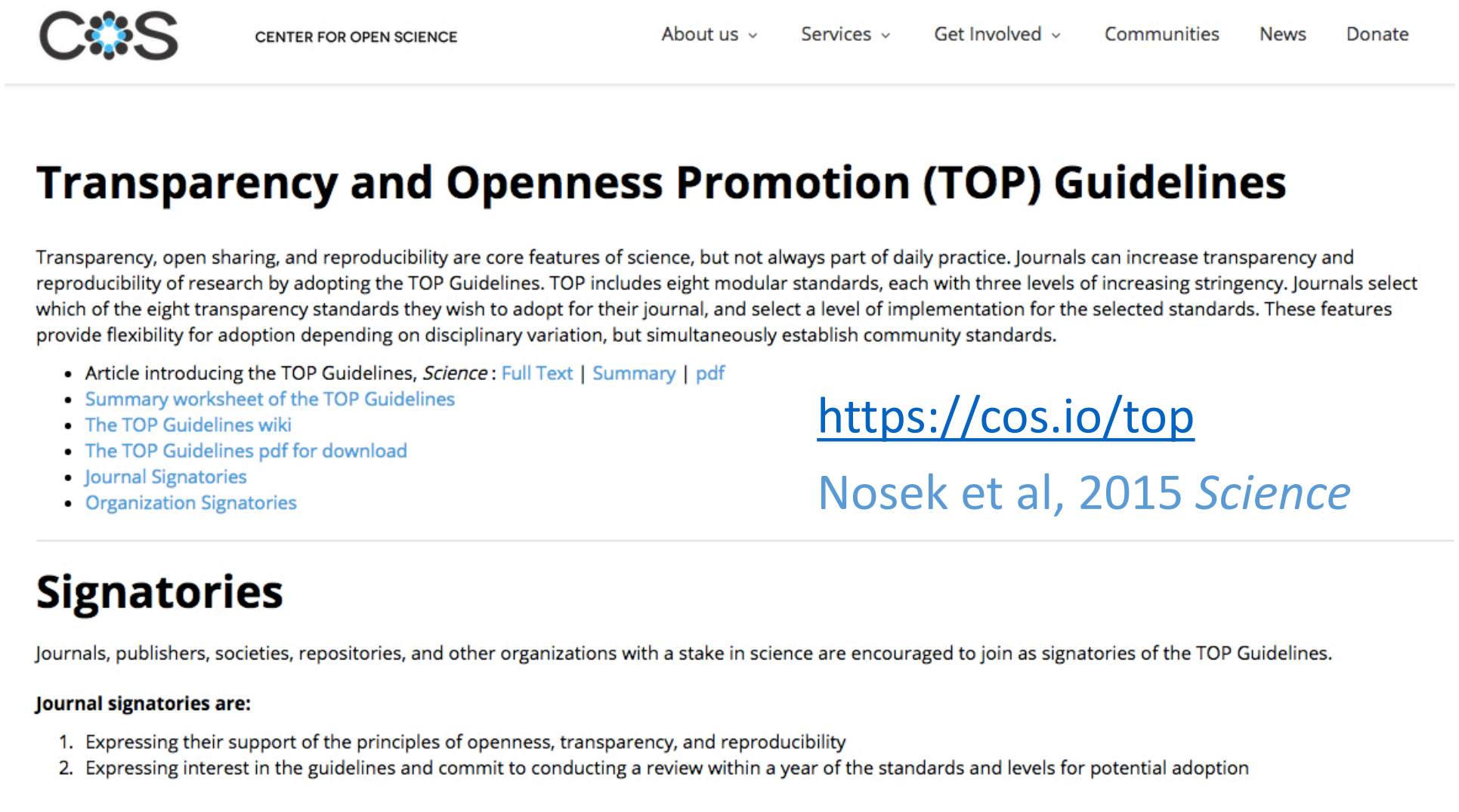
Learn more at cos.io/prereg

# Non-disclosure

- To evaluate the evidentiary quality of research, we need complete reporting of methods and results….
  - *Challenge: limited real estate in journals*
  - *Challenge: heterogeneous reporting*
  - *Challenge: perverse incentives*

- It’s impossible to replicate or validate findings if methods are not disclosed.

# Solution: Standards




https://cos.io/top

Nosek et al, 2015 *Science*

# Solution: Standards, Grass Roots Efforts

- **DA-RT Guidelines**: http://dartstatement.org
  - DATA ACCESS & RESEARCH TRANSPARENCY

- ***Psych Science* Guidelines**
  - Checklists for reporting excluded data, manipulations, outcome measures, sample size. Inspired by grass-roots psychdisclosure.org.

- **21 word solution** in Nelson, Simmons and Simonsohn (2012):
- "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study."

# Selective Reporting:

- Problem: Cherry-picking & fishing for results
- Can result from vested interests, perverse incentives…

You can tell many stories with any data set…

Example: Casey, Glennerster and Miguel (2012, *QJE*)

GoBifo Treatment Effects by Research Hypothesis

| Hypotheses by family | (1) GoBifo mean treatment effect endex | (2) Naive $p$-value | (3) FWER-adjusted $p$-value for all 12 hypos | (4) FWER-adjusted $p$-value for 11 hypos in 2009 PAP |
|---|---|---|---|---|
| Family B: Institutional and social change or "software" effects | | | | |
| Mean effect for family B (Hypotheses 4–12; 155 unique outcomes) | 0.028<br>**(0.020)** | 0.155 | | |
| H4: Participation in GoBifo increases collective action and contributions to local public goods (15 outcomes) | 0.012<br>(0.037) | 0.738 | 0.980 | 0.981 |
| H5: GoBifo increases inclusion and participation in community planning and implementation, especially for poor and vulnerable groups; GoBifo norms spill over into other types of community decisions, making them more inclusive, transparent, and accountable (47 outcomes) | 0.002<br>(0.032) | 0.944 | 0.980 | 0.981 |
| H6: GoBifo changes local systems of authority, including the roles and public perception of traditional leaders (chiefs) versus elected local government (25 outcomes) | 0.056<br>(0.037) | 0.134 | 0.664 | 0.667 |
| H7: Participation in GoBifo increases trust (12 outcomes) | 0.042<br>(0.046) | 0.360 | 0.913 | 0.914 |
| H8: Participation in GoBifo builds and strengthens community groups and networks (15 outcomes) | 0.028<br>(0.037) | 0.450 | 0.913 | 0.914 |
| H9: Participation in GoBifo increases access to information about local governance (17 outcomes) | 0.038<br>(0.037) | 0.301 | 0.913 | 0.913 |
| H10: GoBifo increases public participation in local governance (18 outcomes) | 0.090*<br>(0.045) | 0.045 | 0.315 | 0.322 |
| H11: By increasing trust, GoBifo reduces crime and conflict in the community (8 outcomes) | 0.010<br>(0.043) | 0.816 | 0.980 | 0.981 |
| H12: GoBifo changes political and social attitudes, making individuals more liberal toward women, more accepting of other ethnic groups and "strangers," and less tolerant of corruption and violence (9 outcomes) | 0.041<br>(0.043) | 0.348 | 0.913 | 0.914 |

# Solution: Pre-specify

1. Define hypotheses

2. Identify all outcomes to be measured

3. Specify statistical models, techniques, tests (# obs, sub-group analyses, control variables, inclusion/exclusion rules, corrections, etc)

# Reduce Cherry-Picking

| Outcome variable | (1) Mean for controls | (2) Treatment effect |
|---|---|---|
| Panel A: GoBifo "weakened" institutions | | |
| Attended meeting to decide what to do with the tarp | 0.81 | –0.04+ |
| Everybody had equal say in deciding how to use the tarp | 0.51 | –0.11+ |
| Community used the tarp (verified by physical assessment) | 0.90 | –0.08+ |
| Community can show research team the tarp | 0.84 | –0.12* |
| Respondent would like to be a member of the VDC | 0.36 | –0.04* |
| Respondent voted in the local government election (2008) | 0.85 | –0.04* |
| Panel B: GoBifo "strengthened" institutions | | |
| Community teachers have been trained | 0.47 | 0.12+ |
| Respondent is a member of a women's group | 0.24 | 0.06** |
| Someone took minutes at the most recent community meeting | 0.30 | 0.14* |
| Building materials stored in a public place when not in use | 0.13 | 0.25* |
| Chiefdom official did not have the most influence over tarp use | 0.54 | 0.06* |
| Respondent agrees with "Responsible young people can be good leaders" and not "Only older people are mature enough to be leaders" | 0.76 | 0.04* |
| Correctly able to name the year of the next general elections | 0.19 | 0.04* |

# Failure to replicate

*“Reproducibility is just collaboration with people you don’t know, including yourself next week”—Philip Stark, UC Berkeley*

*“Economists treat replication the way teenagers treat chastity - as an ideal to be professed but not to be practised.”—Daniel Hamermesh, UT Austin*

http://www.psychologicalscience.org/index.php/replication

# Why we care

- Identifies fraud, human error
- Confirms earlier findings (bolsters evidence base)

# Solution: workflow

1. Data analysis is computer programming: Code everything that can be coded.
2. Research ought to be credible communication.
3. Map inputs to outputs
   - We should know where the data came from and what operations were performed on which set of data.
4. Version control prevents clobbering, reconciles history, and helps organize work.
   - No: ”inequalityMV4_APSRsubmit_reallyfinal2”, but “20181203_inequality”
   - Version control: Github, OSF
5. Testing minimizes error.
6. Work should be reproducible.
   - Integrate Coding, Rmarkdown,
   - Organise thoughts: simplenote

- See Bowers and Voors 2016 How to improve your relationship with your future self

# Replication resources

- Replication Wiki:
- http://replication.uni-goettingen.de/

- Large-scale Replication Efforts
  - Reproducibility Project: Psychology
  - Many Labs


# Replication resources

- Data/Code Repositories:
  - Harvard Dataverse
  - Inter-university Consortium for Political and Social Research (ICPSR)
  - Open Science Framework
  - GitHub

# Replication Standards

- Replications need to be subject to rigorous peer review (no “second-tier” standards)

- Could they be pre-registered as well?

# Replication Standards

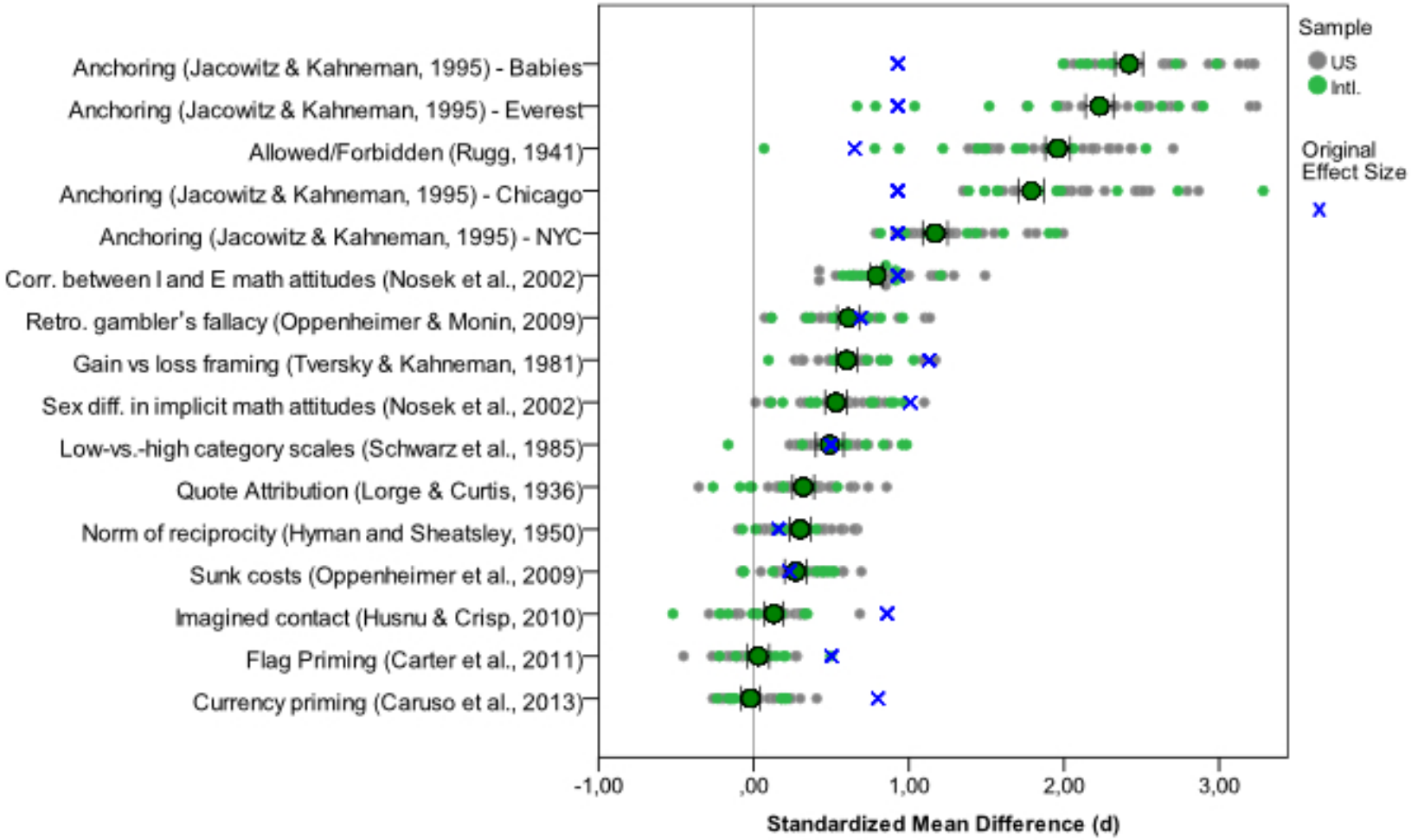
**The Reproducibility Project: Psychology** was a crowdsourced effort to estimate the reproducibility of a sample of 100 studies from the literature.

*Science (Aug 28, 2015): "*Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically 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."

https://osf.io/ezcuj/

# Many Labs

https://osf.io/wx7ck/



# Some Solutions…

- Publication bias → Pre-registration
- Non-disclosure → Reporting standards
- P-hacking→ Pre-specification
- Failure to replicate → Open data/materials, Many Labs

# Resources at WUR

- Scientific Symposium: FAIR Data Science for Green Life Sciences
- Organised by: Wageningen Data Competence Center
- Date: Wed 12 December 2018
- Venue: Orion

- https://www.wur.nl/en/activity/Scientific-Symposium-FAIR-Data-Science-for-Green-Life-Sciences.htm

- **Wageningen Data Competence Center**
- *The WDCC is established to support developments in the field of (big) data and data science at Wageningen University & Research.*
- https://www.wur.nl/en/Value-Creation-Cooperation/WDCC.htm

# Resources at WUR

- Course: Research Data Management
  - Part 1 focuses on how to set up your data collection system and how to keep it organised and ensure it is understandable.
  - Part 2 deals with creating a data management plan (DMP), data storage options and the support and services available at WUR.
  - Part 3 goes into the long-term storage and the publication of research data.

- https://wgs.crs.wur.nl/Courses/Details/84

# Disclaimer

- Many slides from
  - Garret Christensen, (Research Fellow BITSS and Berkeley Institute for Data Science)
  - Some Macartan Humphreys (EGAP Learning Days)

- See also
  - Christensen and Miguel (2018) Transparency, Reproducibility, and the Credibility of Economics Research, Journal of Economic Literature vol. 56, no. 3, September 2018(pp. 920-80)
    - https://github.com/garretchristensen/BestPracticesManual
  - Miguel et al (2014) Promoting Transparency in Social Science Research. Science 03 Jan 2014:Vol. 343, Issue 6166, pp. 30-31DOI: 10.1126/science.1245317
  - Matthew Gentzkow Jesse M. Shapiro, Code and Data for the Social Sciences: A Practitioner's Guide