# Reproducibility, replication, etc.

## APAM E4990
## Modeling Social Data

Jake Hofman

Columbia University

February 22, 2019

# Questions

How should one evaluate research results?

- Was the research done and reported honestly / correctly?
- Is the result "real" or an artifact of the data / analysis?
- Will it hold up over time?
- How robust is the result to small changes?
- How important / useful is the finding?

# Honesty

Was the data accurately collected and reported?

We'll take the optimistic view that most researchers are honest, although there are exceptions

# Honesty

## List of scientific misconduct incidents

From Wikipedia, the free encyclopedia

### Social sciences [ edit ]

- Mart Bax (Netherlands), former professor of political anthropology at the Vrije Universiteit, committed multiple acts of scientific misconduct including data fabrication.[259][260][261] Bax, who has had two of his publications retracted, was found in 2013 to have never published 61 of the papers he listed on his CV.[262][263]
- Ward Churchill (US), former professor of ethnic studies at University of Colorado, was accused by a University committee of multiple counts of plagiarism, fabrication, and falsification.[264][265] After the University Chancellor recommended Churchill's dismissal to the Board of Regents, Churchill was in 2009 deemed by a jury to have been wrongly fired, although the presiding judge declined to reinstate him.[266] In 2010 the Colorado State Court of Appeals upheld the judge's decision to not reinstate Churchill, a decision that in 2012 was upheld by the Colorado Supreme Court.[267] In 2013 the Supreme Court of the United States declined to hear Churchill's appeal of the Colorado Supreme Court decision.[268]
- Jens Förster (Netherlands, Germany), a social psychologist formerly of the University of Amsterdam and the Ruhr-Universität Bochum, fabricated data reported in a number of published papers. An investigating committee in 2015 identified in Förster's work data that were "practically impossible" and displayed "strong evidence for low veracity."[269][270] Förster has had three of his publications retracted,[271][272] and four others attached to an expression of concern.[273]
- Bruno Frey (Switzerland), an economist formerly at the University of Zurich, in 2010-11 committed multiple acts of self-plagiarism in articles about the *Titanic* disaster. Frey admitted to the self-plagiarism, terming the acts "grave mistake[s]" and "deplorable."[274][275]
- Michael LaCour (US), former graduate student in political science at UCLA, was the lead author of the 2014 article When contact changes minds. Published in *Science* and making international headlines, the paper was later retracted because of numerous irregularities in the methodology and falsified data.[276][277][278][279] Following the retraction Princeton University rescinded an assistant professorship that had been offered to LaCour.[280]
- Karen M. Ruggiero (US), former Assistant Professor of Psychology at Harvard University, fabricated NIH-sponsored research data on gender and discrimination.[281][282][283] Ruggiero has had two research publications retracted.[284]
- Diederik Stapel (Netherlands), former professor of social psychology at Tilburg University, fabricated data in dozens of studies on human behaviour,[285] a deception described by the New York Times as "an audacious academic fraud."[286] Stapel has had 58 of his publications retracted.[287]
- Brian Wansink (US), former John S. Dyson Endowed Chair in the Applied Economics and Management Department at Cornell University, was found in 2018 by a University investigatory committee to have "committed academic misconduct in his research and scholarship, including misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship."[288][289][290] Wansink has had 18 of his research papers retracted (one twice), seven other papers have been attached to an expression of concern, and 15 others have been corrected.[291][292][293]

# Reproducibility

Can you independently verify the exact results using the *same data* and the *same analysis*?

Though a low bar, most research doesn't currently pass this test:

- Data or code aren't available / complete
- Code is difficult to run / understand
- Complex software dependencies

# Reproducibility

Can you independently verify the exact results using the *same data* and the *same analysis*?

This is improving with better software engineering practices among researchers:

- Literate programming (Jupyter, Rmarkdown)
- Automated build scripts (Makefiles)
- Containers (Docker, Code Ocean)

# A Practical Taxonomy of Reproducibility for Machine Learning Research

**Rachael Tatman**
Kaggle
Seattle, WA 91803
rachael@kaggle.com

**Jake VanderPlas**
eScience Institute
University of Washington
Seattle, WA 98195
jakevdp@uw.edu

**Sohier Dane**
Kaggle
Seattle, WA 91803
sohier@kaggle.com

Issues of reproducibility in ML extend beyond ICML. Of the 679 papers presented at NIPS in 2017, for instance, only 259–or less than 40%–provided links to code on the NIPS website. Far fewer provided the environment needed to actually run that code. A few notable exceptions include Liu et al.'s paper on unsupervised image-to-image translation networks [14] and the papers presented at the MLTrain NIPS workshop. As a field, ML is making strides towards reproducibility, but there is still a long way to go.

# Reproducibility

# Replicability

Will the result hold up with new data but the same analysis?

- It's easy to be fooled by randomness
- Noise can dominate signal in small datasets
- Asking too many questions of the data can lead to overfitting

# Crisis

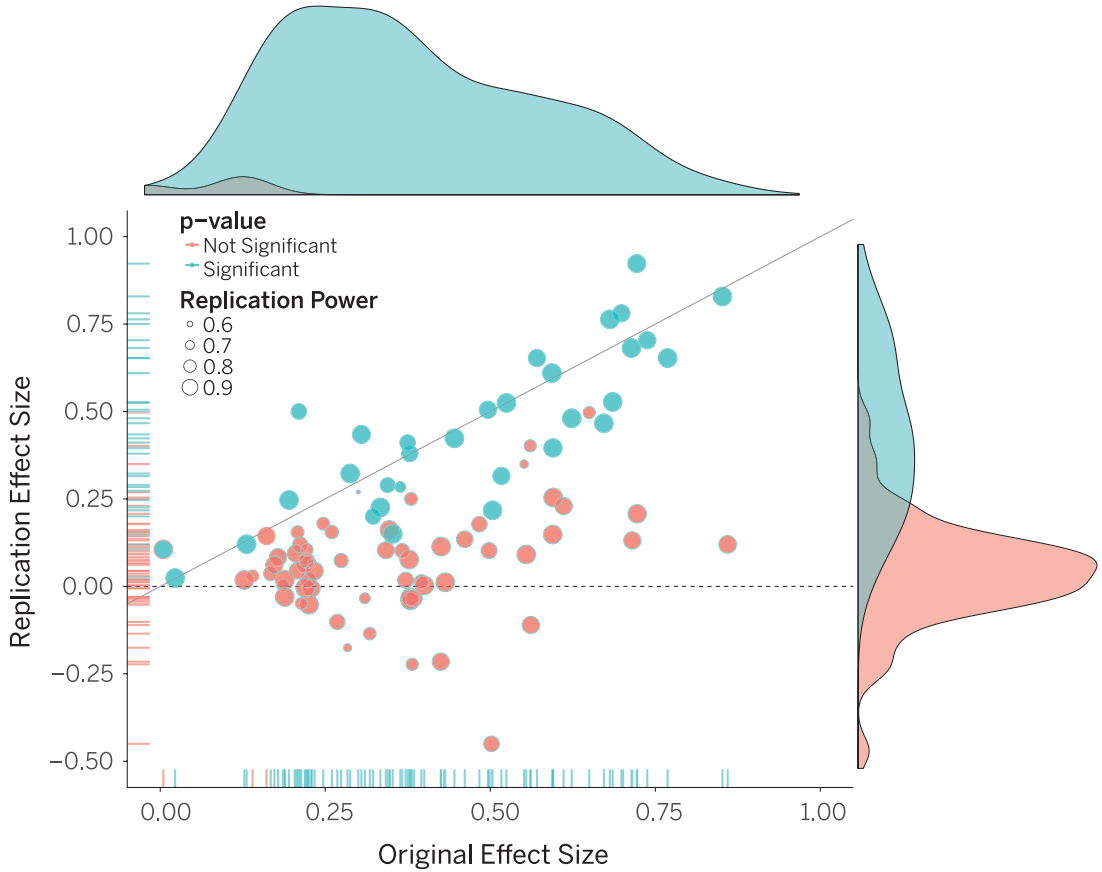
## Estimating the reproducibility of psychological science

**Open Science Collaboration***†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. 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. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Believe about half of what you read

# Crisis




**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.

## Believe about half of what you read

# Statistics!

Hypothesis testing?
P-values?
Statistical significance?
Confidence intervals??
Effect sizes???


xkcd.com/892

# Quiz #1

Which treatment would you prefer?

- Treatment A was found to improve health over a placebo by 10 points on average (with a standard error of 5 points) in a study with N = 100 participants.
- Treatment B was found to improve health over a placebo by 10 points on average (with a standard error of 5 points) in a study with N = 1,000 participants.

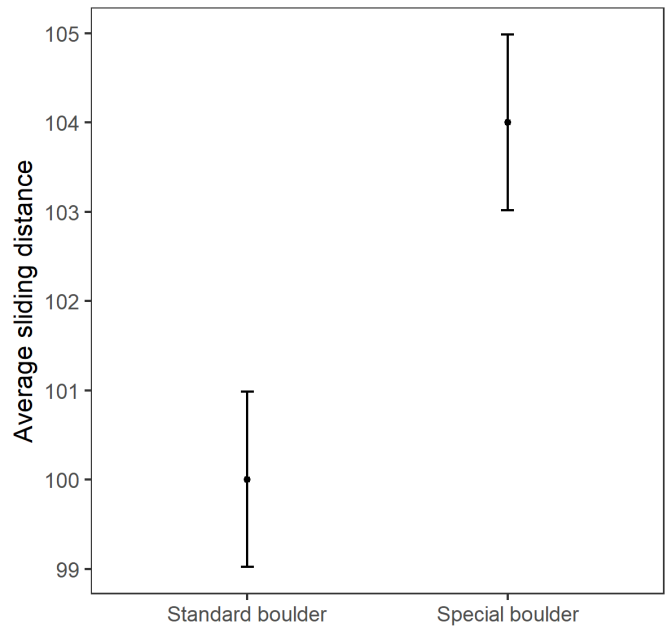
# Quiz #2 (Oakes 1986)

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant (t = 2.7, df = 18, p = .01). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).
(2) You have found the probability of the null hypothesis being true.
(3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
(4) You can deduce the probability of the experimental hypothesis being true.
(5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
(6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

# Quiz #3a (Hofman, Hullman, Goldstein 2019?)

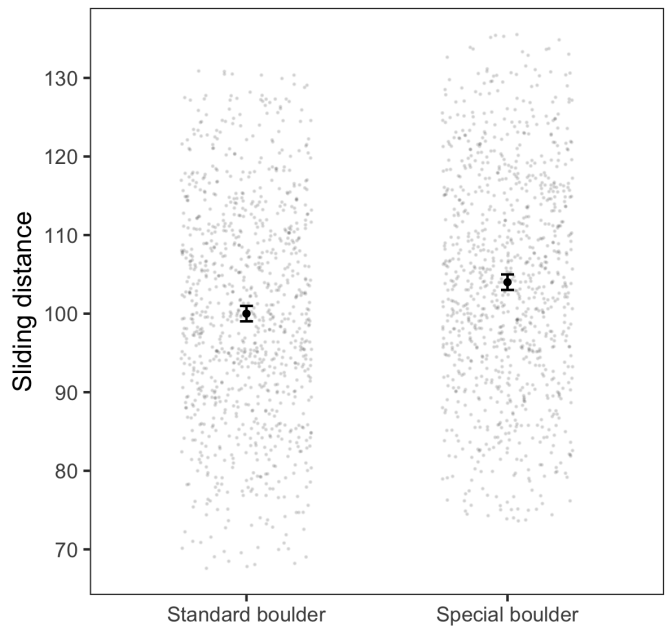
Below are results of an experiment with 1,000 slides of a standard boulder (left) and special boulder (right), with bars showing one standard error on the mean.




Estimate the probability that a slide of the special boulder goes farther than a slide of the standard boulder.

# Quiz #3b (Hofman, Hullman, Goldstein 2019?)

Below are results of an experiment with 1,000 slides of a standard boulder (left) and special boulder (right), with bars showing one standard error on the mean and points showing individual slides.




Estimate the probability that a slide of the special boulder goes farther than a slide of the standard boulder.

# Underpowered studies

## Why Most Published Research Findings Are False

**John P. A. Ioannidis**

**Summary**

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

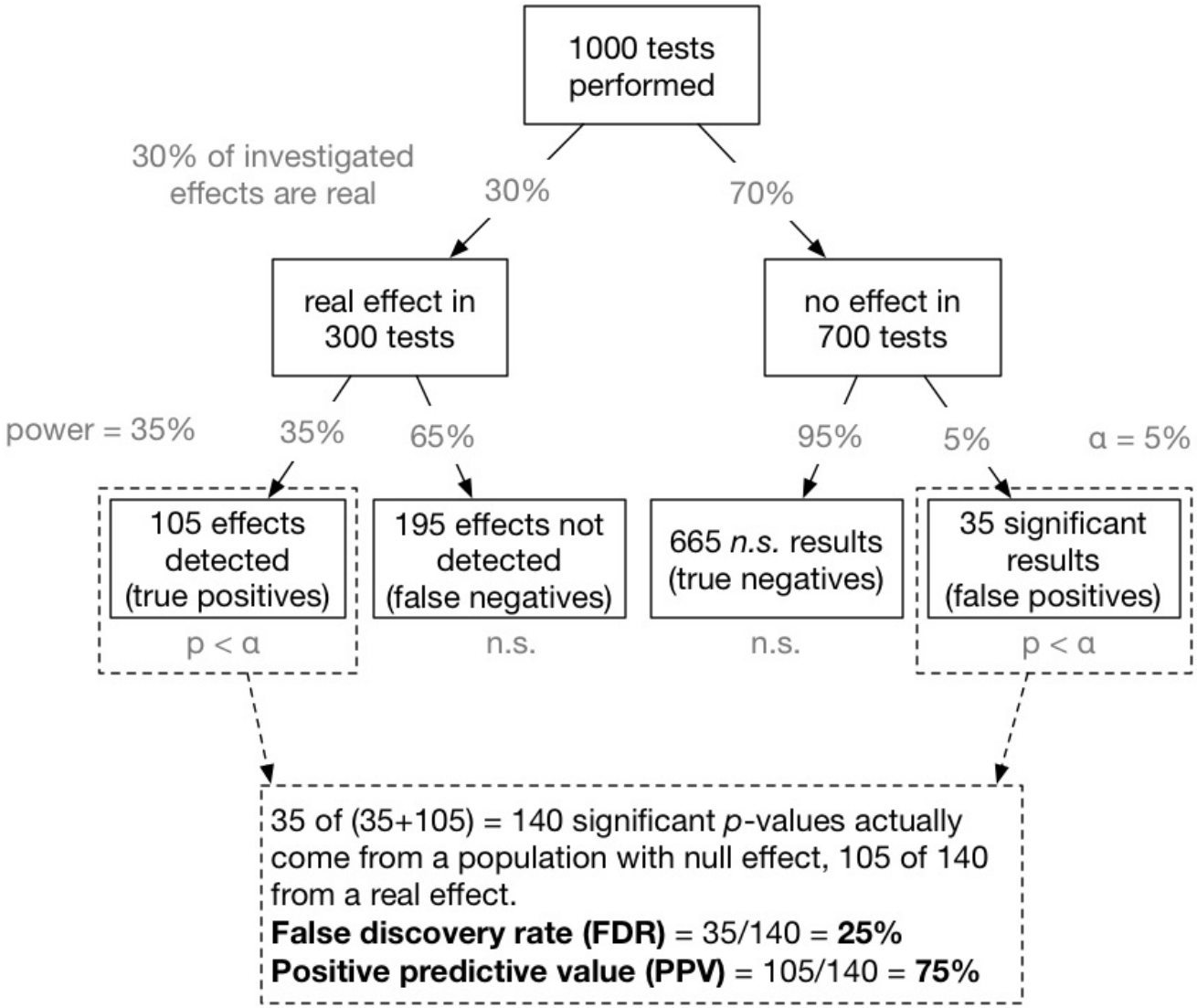
### Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a $p$-value less than 0.05. Research is not most appropriately represented and summarized by $p$-values, but, unfortunately, there is a widespread notion that medical research articles

**It can be proven that most claimed research findings are false.**

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, $\alpha$. Assuming that $c$ relationships are being probed in the field, the expected values of the $2 \times 2$ table are given in Table 1. After a research finding has been claimed based on

# Underpowered studies




bit.ly/fdrtree

# P-hacking

## False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

**Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]**
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

**Abstract**

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

# P-hacking

### *Study 1: musical contrast and subjective age*

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older. In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song by Mr. Scruff that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated survey: They answered the question "How old do you feel right now?" by choosing among five options (*very young*, *young*, *neither young nor old*, *old*, and *very old*). They also reported their father's age, allowing us to control for variation in baseline age across participants.

An analysis of covariance (ANCOVA) revealed the predicted effect: People felt older after listening to "Hot Potato" (adjusted $M = 2.54$ years) than after listening to the control song (adjusted $M = 2.06$ years), $F(1, 27) = 5.06$, $p = .033$.

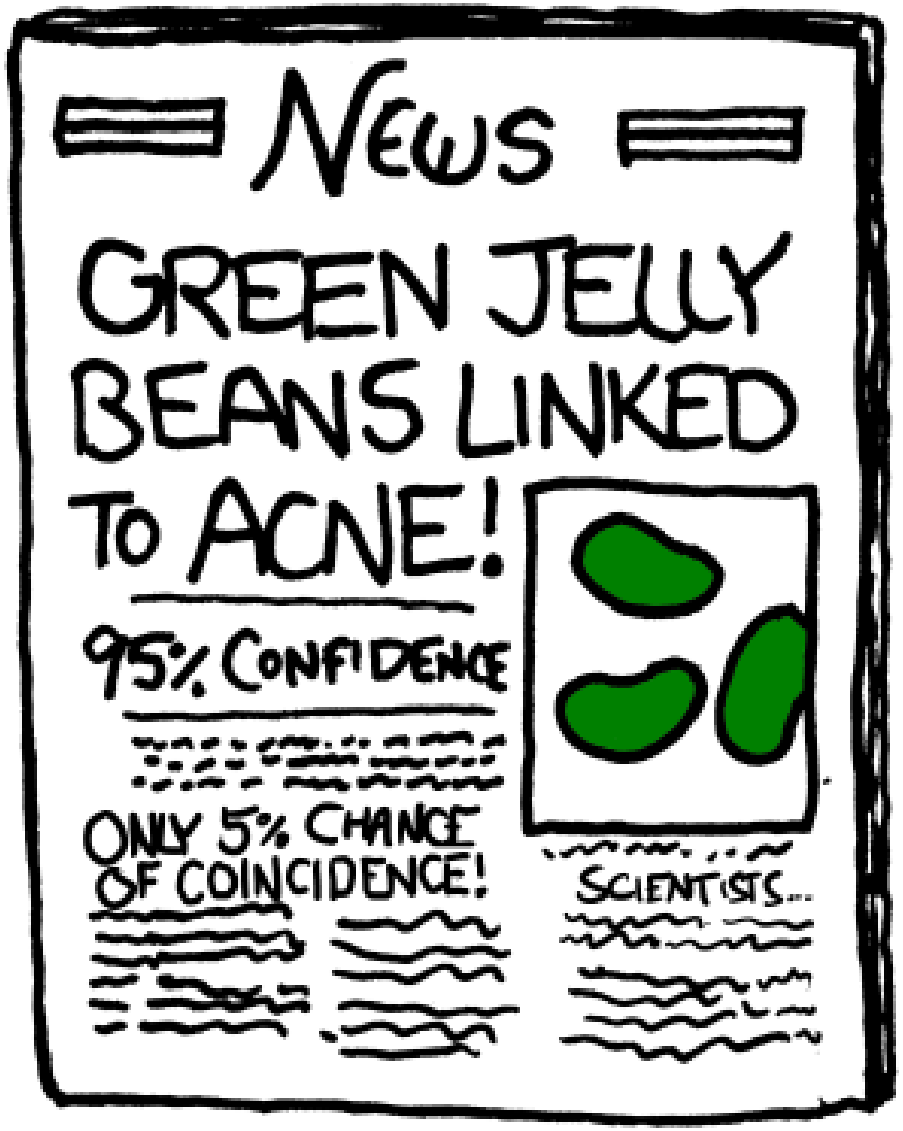
In Study 2, we sought to conceptually replicate and extend Study 1. Having demonstrated that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people *actually* younger.

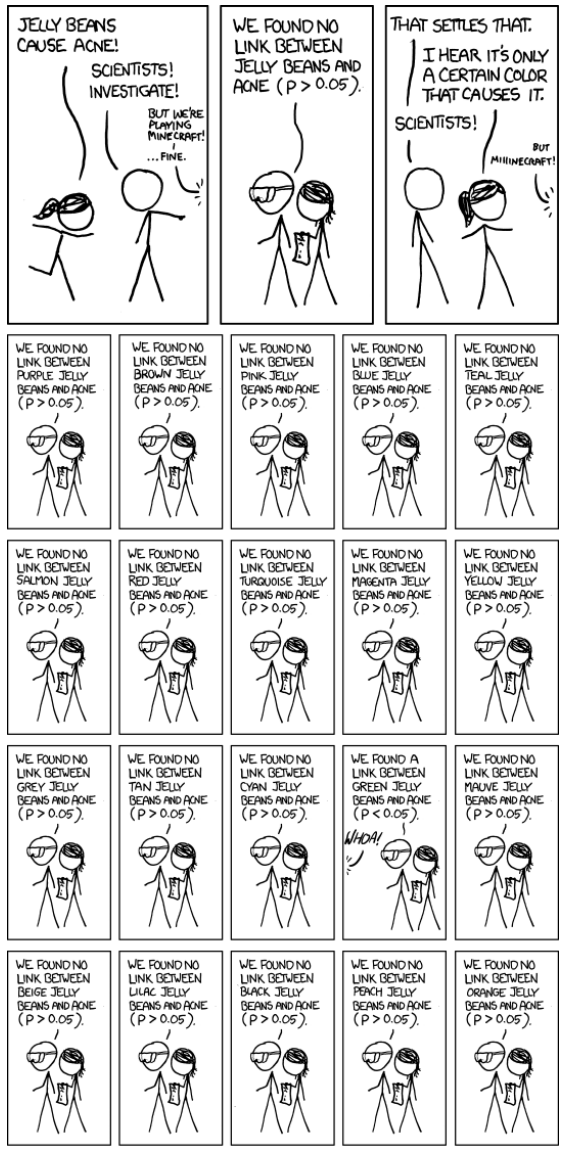
### *Study 2: musical contrast and chronological rejuvenation*

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted $M = 20.1$ years) rather than to "Kalimba" (adjusted $M = 21.5$ years), $F(1, 17) = 4.92$, $p = .040$.

# P-hacking




xkcd.com/882

# Researcher degrees of freedom

**The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time***

Andrew Gelman† and Eric Loken‡

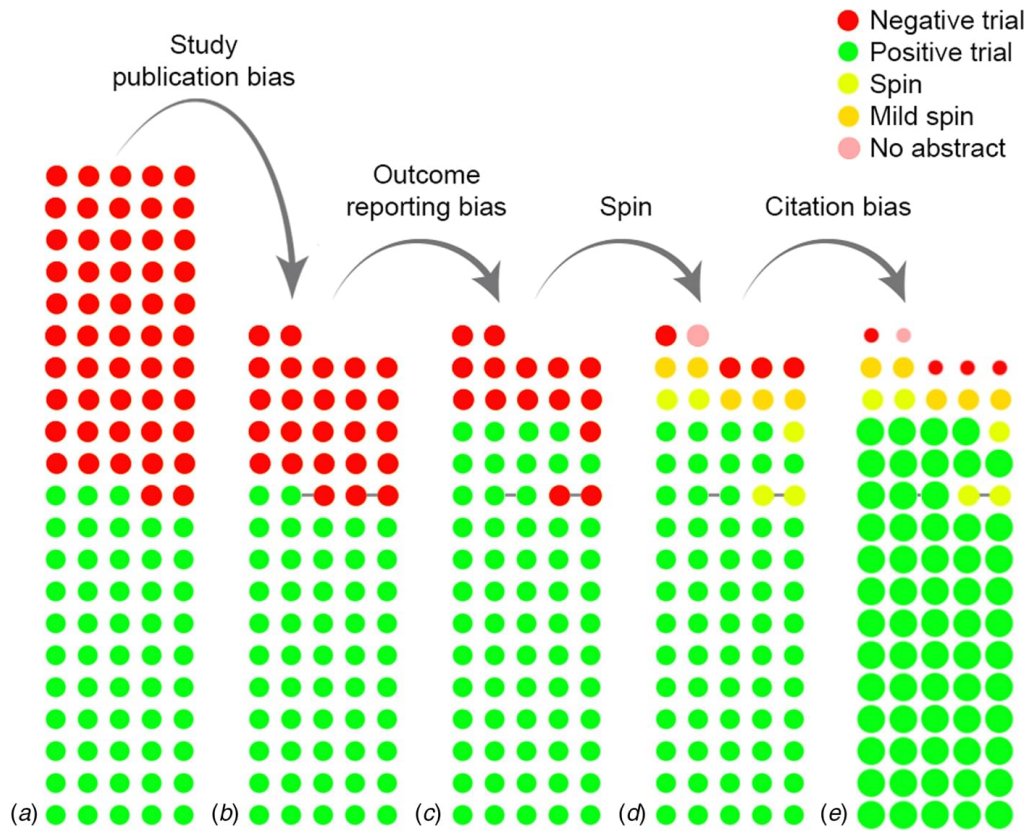
14 Nov 2013

**Abstract**

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

# Publication / citation bias

While only 50% of FDA-registered studies on antidepressants find positive results, but 95% of publications report positive findings.




bit.ly/depressionspin

# Robustness

Opinion

## No, You Can't Ignore Email. It's Rude.

Being overwhelmed is no excuse. It's hard to be good at your job if you're bad at responding to people.

**By Adam Grant**
Dr. Grant is an organizational psychologist.

Feb. 15, 2019

When researchers compiled a huge database of the digital habits of teams at Microsoft, they found that the clearest warning sign of an ineffective manager was being slow to answer emails. Responding in a timely manner shows that you are conscientious — organized, dependable, and hardworking. And that matters. In a comprehensive analysis of people in hundreds of occupations, conscientiousness was the single best personality predictor of job performance. (It turns out that people who are rude online tend to be rude offline, too.)