What is the reproducibility crisis in science and what can we do about it?

Dorothy V. M. Bishop

Professor of Developmental Neuropsychology

University of Oxford

@deevybee

wellcome trust

UNIVERSITY OF OXFORD

What is the problem?

Essay

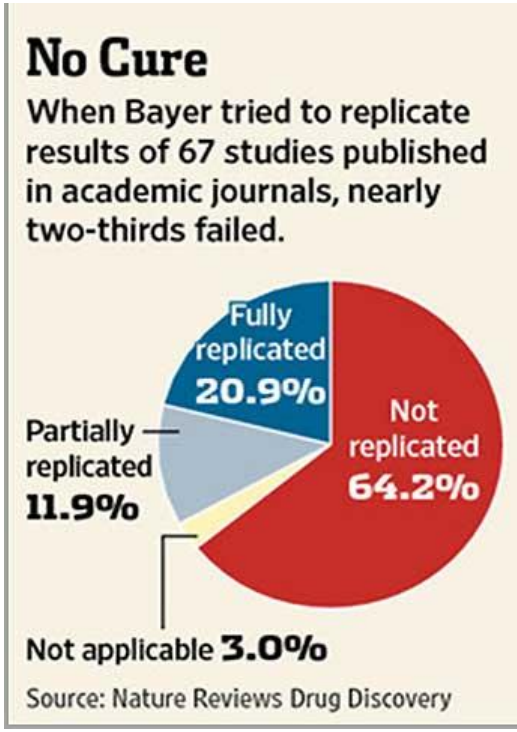
Why Most Published Research Findings Are False

John P. A. Ioannidis

2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

"There is increasing concern about the reliability of biomedical research, with recent articles suggesting that up to 85% of research funding is wasted."

Bustin, S. A. (2015). The reproducibility of biomedical research: Sleepers awake! *Biomolecular Detection and Quantification*



THE LANCET

Online First | Current Issue | All Issues | Special Issues | Multimedia | Information for Authors

All Content | Search | Advanced Search

Research: increasing value, reducing waste

Published: January 8, 2014

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Arch

News & Comment > News > 2015 > May > Article

NATURE | NEWS

First results from psychology's largest reproducibility test

The four horsemen of the Apocalypse

HARKing

Low power

P-hacking

Publication bias

Historical timeline: concerns about reproducibility

1956
De Groot

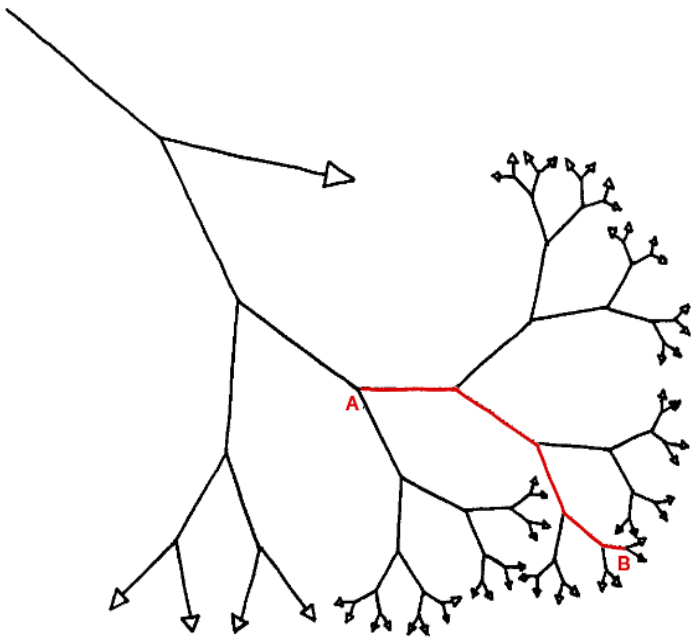
Failure to distinguish between hypothesis-testing and hypothesis-generating (exploratory) research
-> misuse of statistical tests

P-hacking

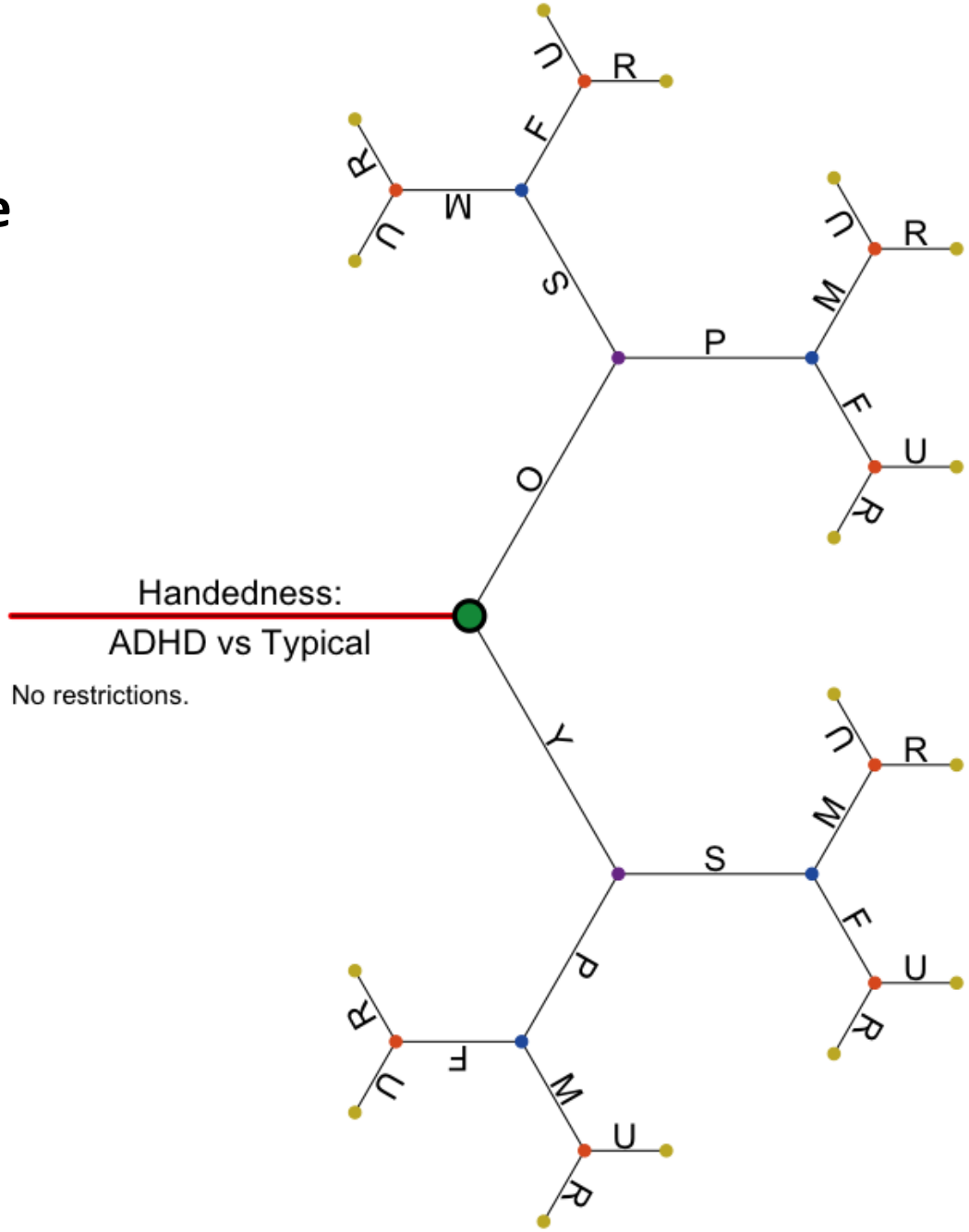
Gelman A, and Loken E. 2013.
The garden of forking paths

"El jardín de senderos que se bifurcan"

The Garden of Forking Paths
by Jorge Luis Borges



Large population database used to explore link between ADHD and handedness

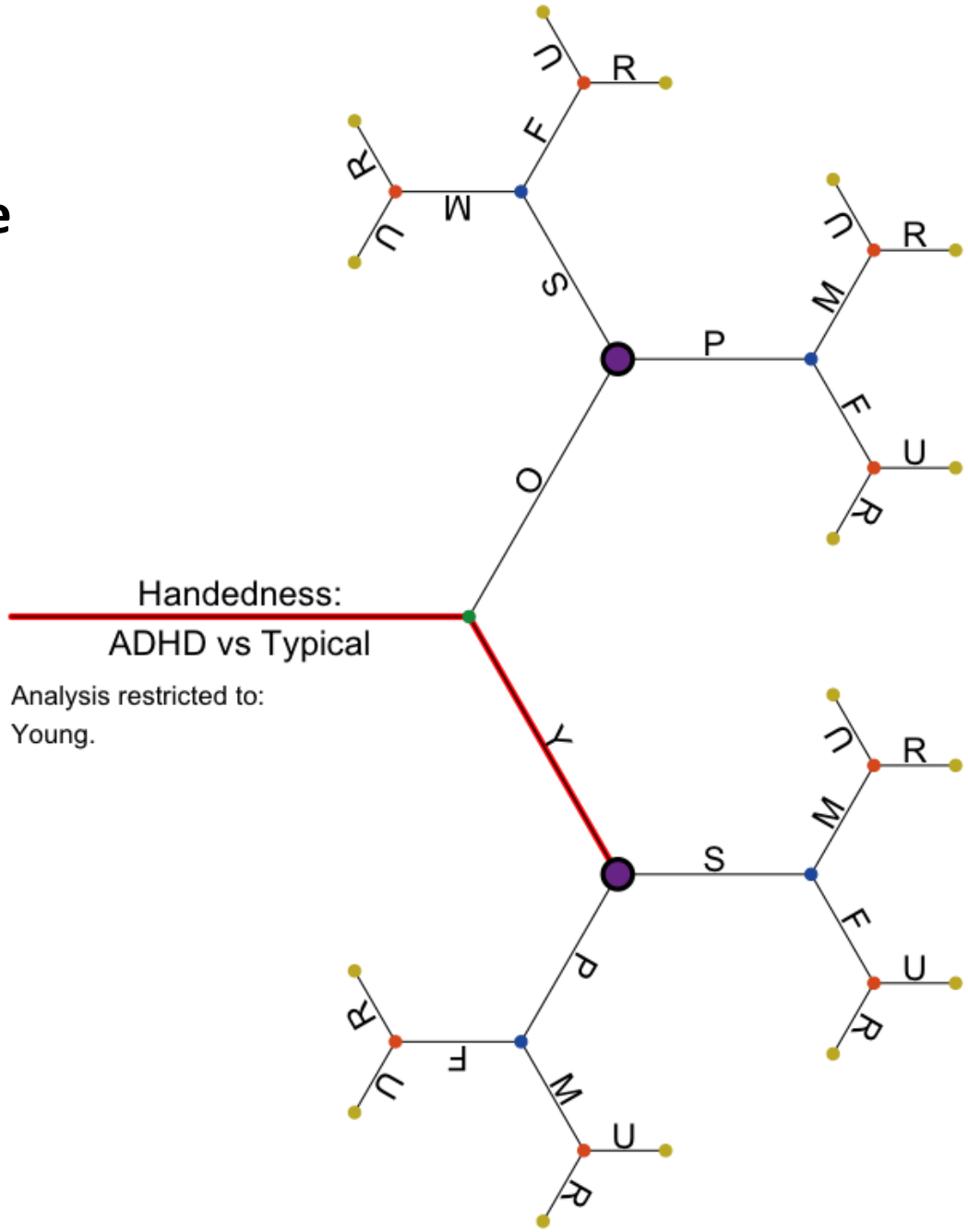


1 contrast

Probability of a ‘significant’ p-value < .05 = .05

https://figshare.com/articles/The_Garden_of_Forking_Paths/2100379

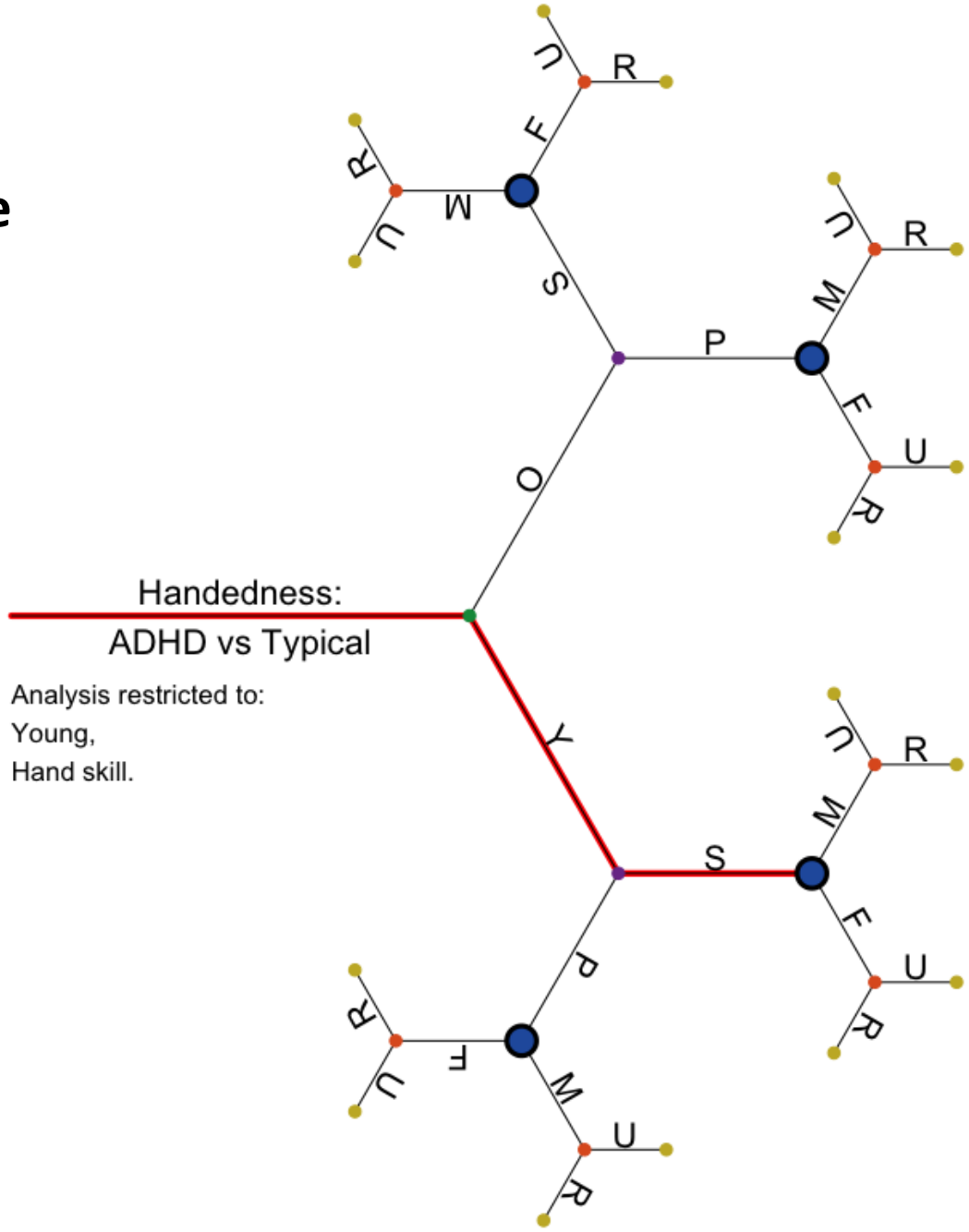
Large population database used to explore link between ADHD and handedness



Focus just on Young subgroup:
2 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .10

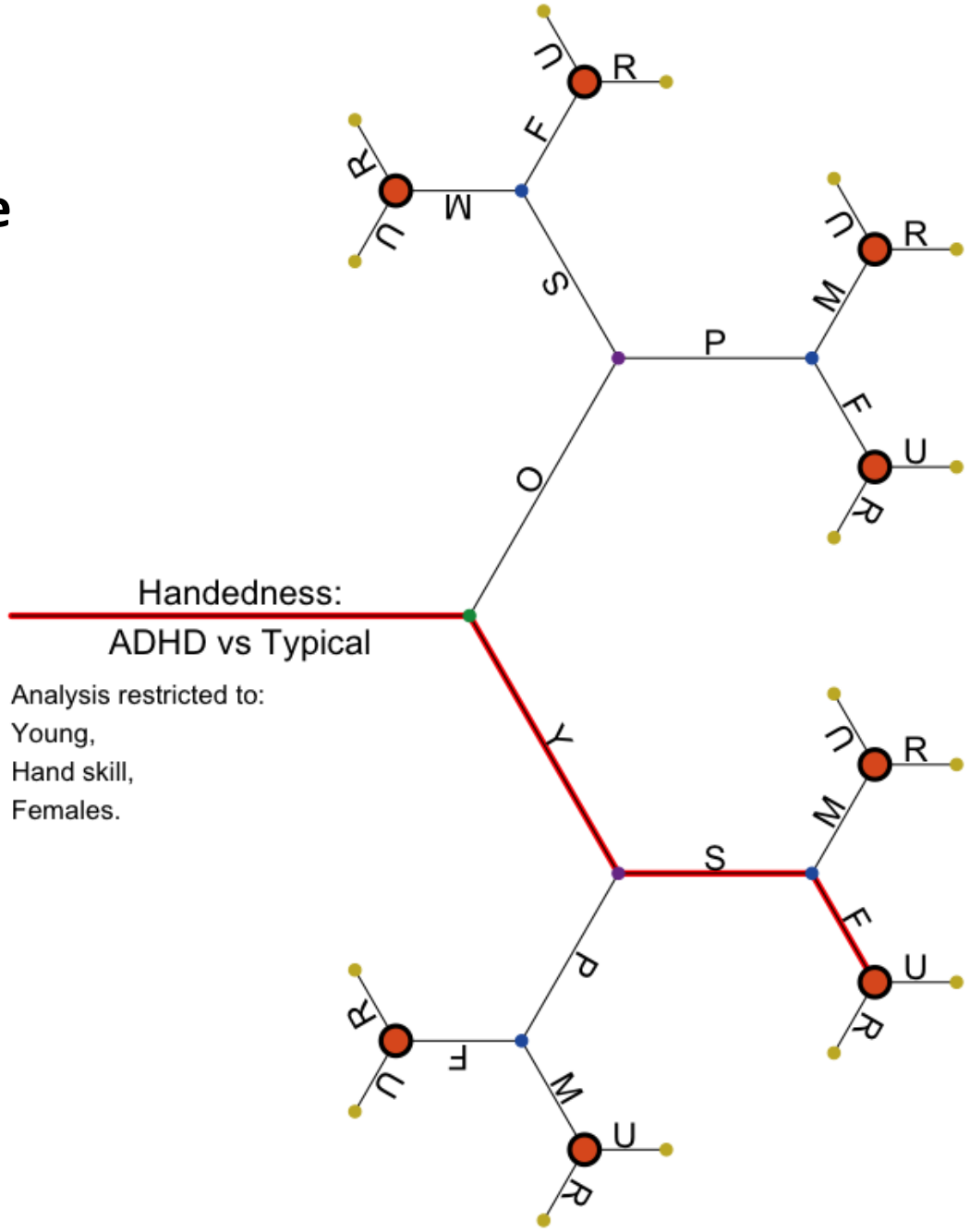
Large population database used to explore link between ADHD and handedness



Focus just on Young on measure of hand skill:
4 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .19

Large population database used to explore link between ADHD and handedness



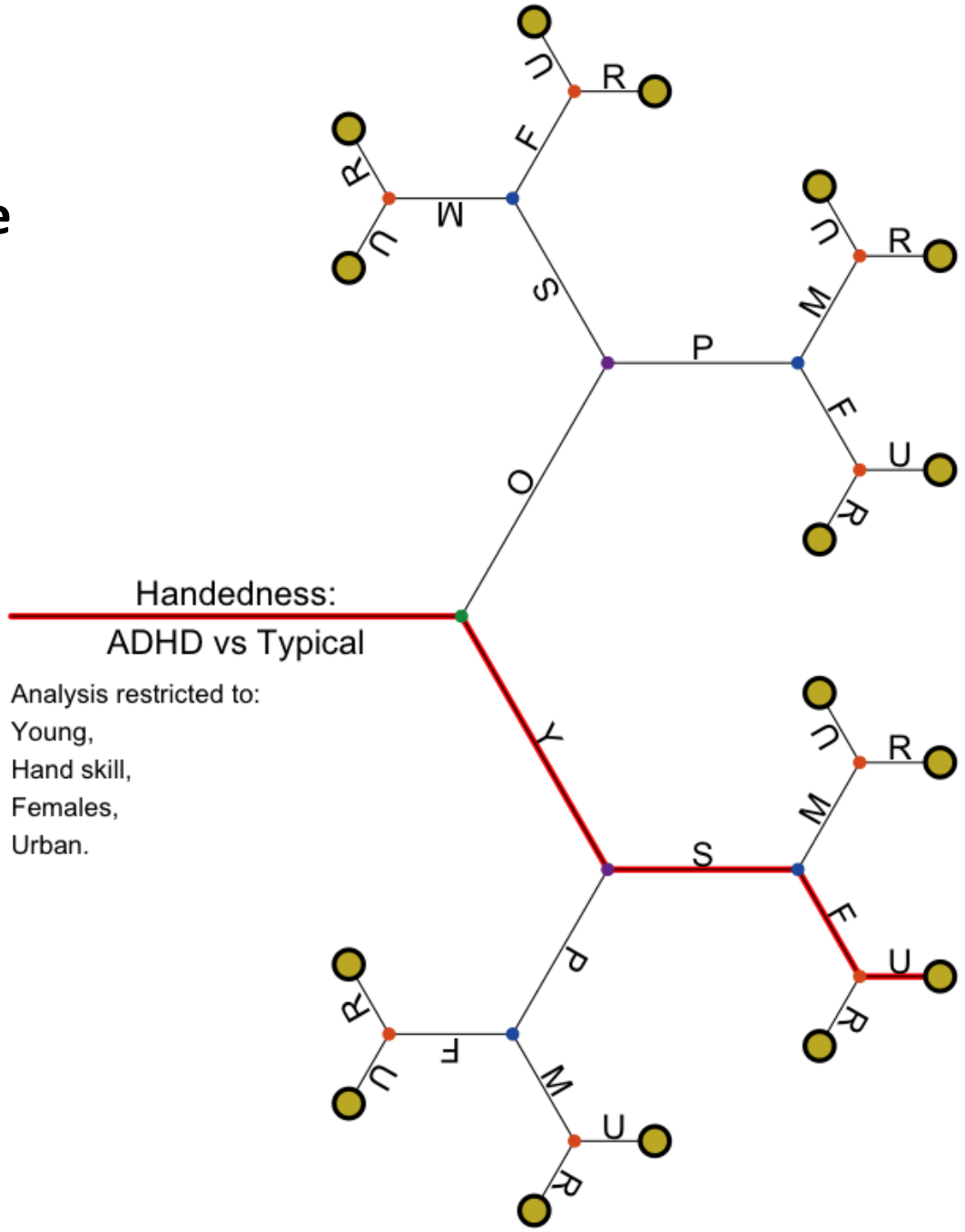
Focus just on Young, Females on measure of hand skill: 8 contrasts at this level

Probability of a ‘significant’ p-value < .05 = .34

Large population database used to explore link between ADHD and handedness

Focus just on Young, Urban, Females on measure of hand skill: 16 contrasts at this level

Probability of a 'significant' p-value < .05 = .56

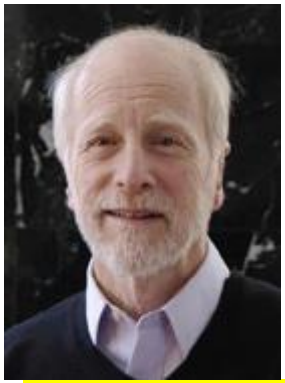


Publication bias

1956 De Groot

1975 Greenwald

1979 Rosenthal



The “file drawer” problem

Prejudice against the null

“As it is functioning in at least some areas of behavioral science research, the research-publication system may be regarded as a device for systematically generating and propagating anecdotal information.”

Low power

1956 De Groot

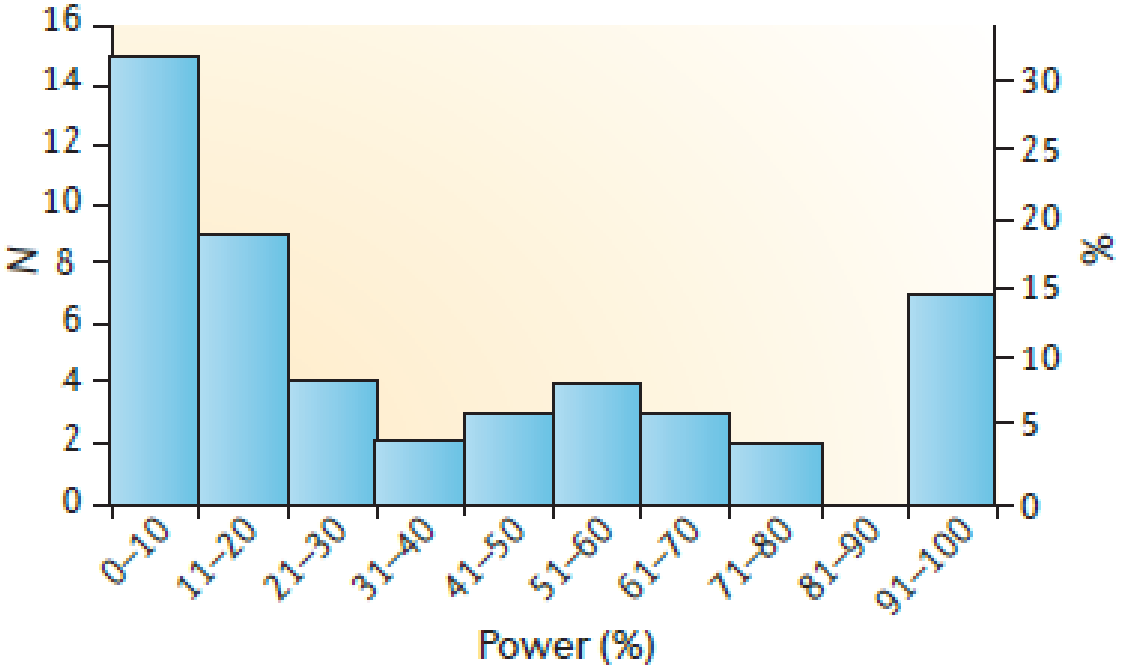
1975 Greenwald

1979 Rosenthal

1987 Newcombe

"Small studies continue to be carried out with little more than a blind hope of showing the desired effect. Nevertheless, papers based on such work are submitted for publication, especially if the results turn out to be statistically significant."

Button KS et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14:365-376.*



Median power of studies included in neuroscience meta-analyses

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

Explicitly advises HARKing!

Writing the Empirical Journal Article

Daryl J. Bem

The Compleat Academic: A Practical Guide for the Beginning Social Scientist, 2nd Edition. Washington, DC: American *Psychological* Association, 2004.

Which Article Should You Write?

There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).

re Data Analysis: Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something— anything —interesting.

"This book provides invaluable guidance that will help new academics plan, play, and ultimately win the academic career game."

Why is this making headlines now?

"It really is striking just for how long there have been reports about the poor quality of research methodology, inadequate implementation of research methods and use of inappropriate analysis procedures as well as lack of transparency of reporting. All have failed to stir researchers, funders, regulators, institutions or companies into action". Bustin, 2014

- Increase in studies quantifying the problem
- Concern from those who use research:
  - Doctors and Patients
  - Pharma companies
- Social media

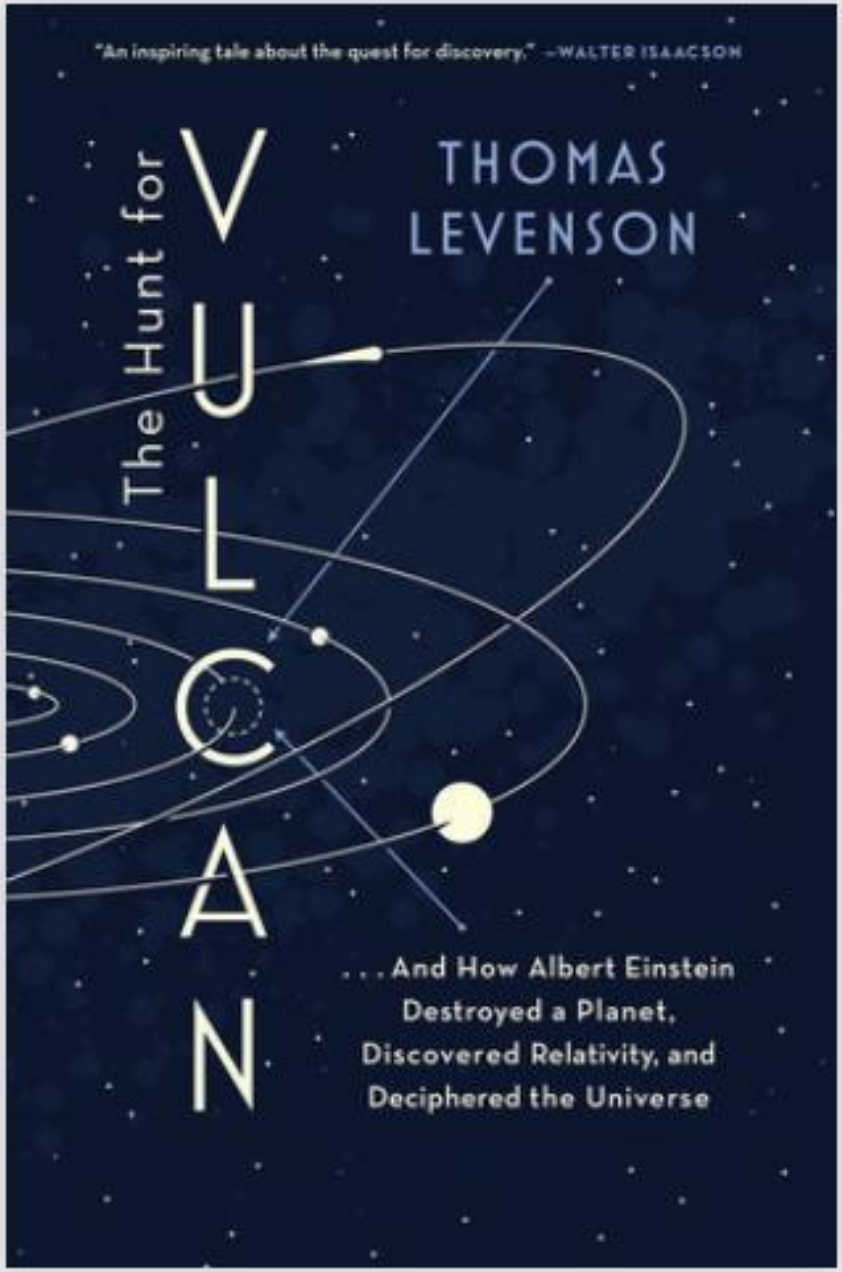
What is the solution?

Complex problem: needs to be attacked from multiple directions

- Researchers
- Journals
- Institutions
- Funders

Problems caused by researchers: 1

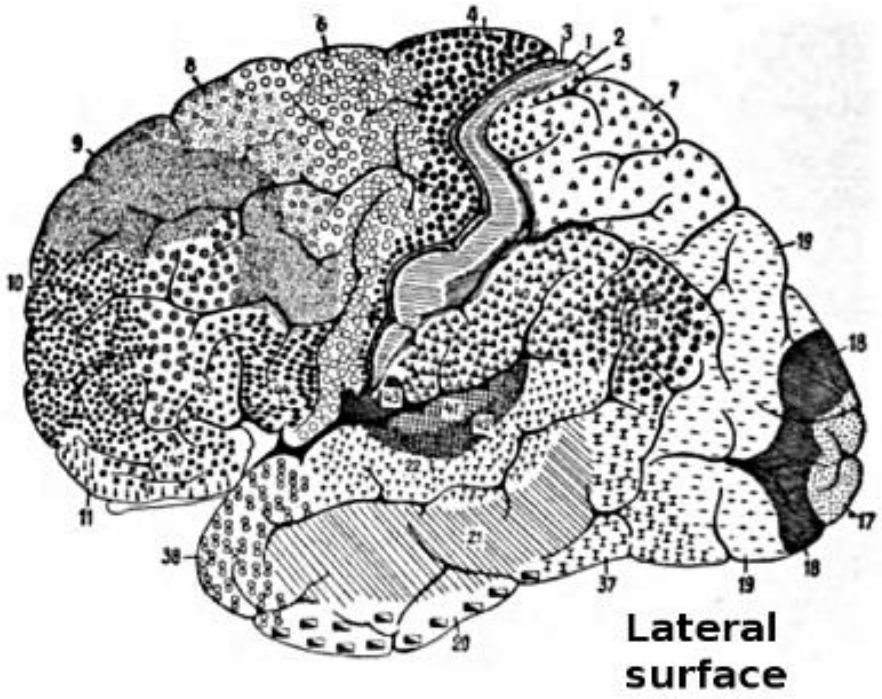
Failure to appreciate power of ‘the prepared mind’
Natural instinct is to look for consistent evidence, not disproof



"The self-deception comes in that over the next 20 years, people believed they saw specks of light that corresponded to what they thought Vulcan should look during an eclipse: round objects crossing the face of the sun, which were interpreted as transits of Vulcan."

Seeing things in complex data requires skill

Brodmann areas, 1909



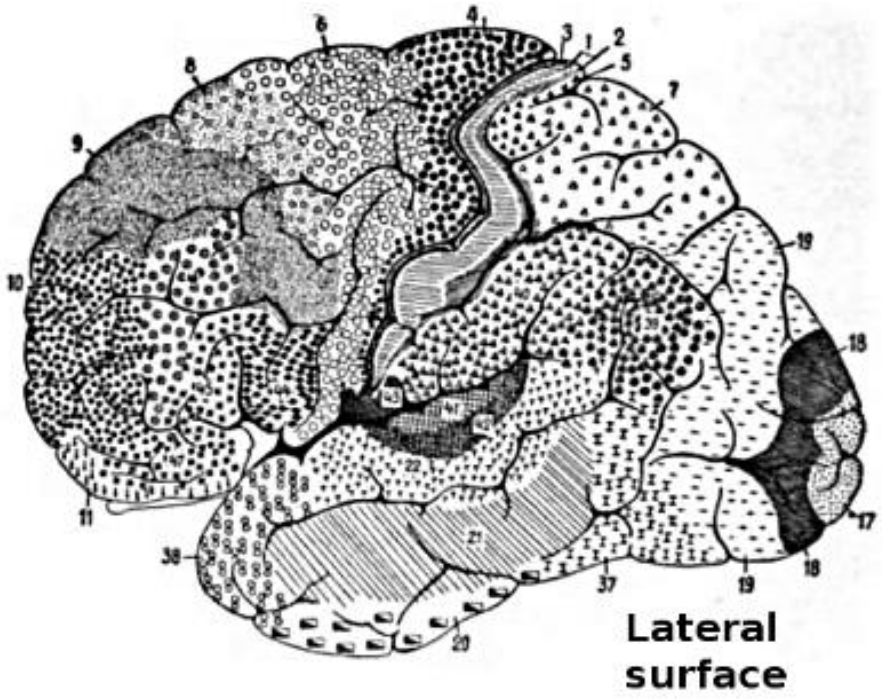
Bailey and von Bonin (1951) noted problems in Brodmann's approach — lack of observer independency, reproducibility and objectivity

Yet have stood test of time: still used today

Discusses failure so replicate studies on preferential looking in babies – role of experimenter expertise

Seeing things in complex data requires skill

Brodmann areas, 1909



Bailey and von Bonin (1951) noted problems in Brodmann's approach — lack of observer independency, reproducibility and objectivity

Yet have stood test of time: still used today

Or pareidolia

Special expertise or Jesus in toast? How to decide

- Eradicate subjectivity from methods
- Adopt standards from industry for checking/double-checking
- Automate data collection and analysis as far as possible
- Make recordings of methods (e.g. Journal of Visualised Experiments)
- Make data and analysis scripts open

Problems caused by researchers: 2

Failure to understand statistics (esp. p-values and power)

http://deevybee.blogspot.co.uk/2016/01/the-amazing-significo-why-researchers.html

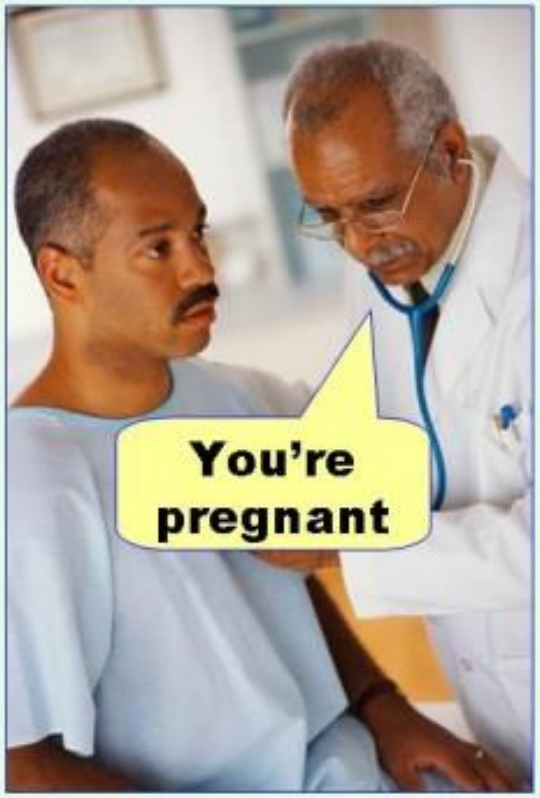
False-Positive Psychology

Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

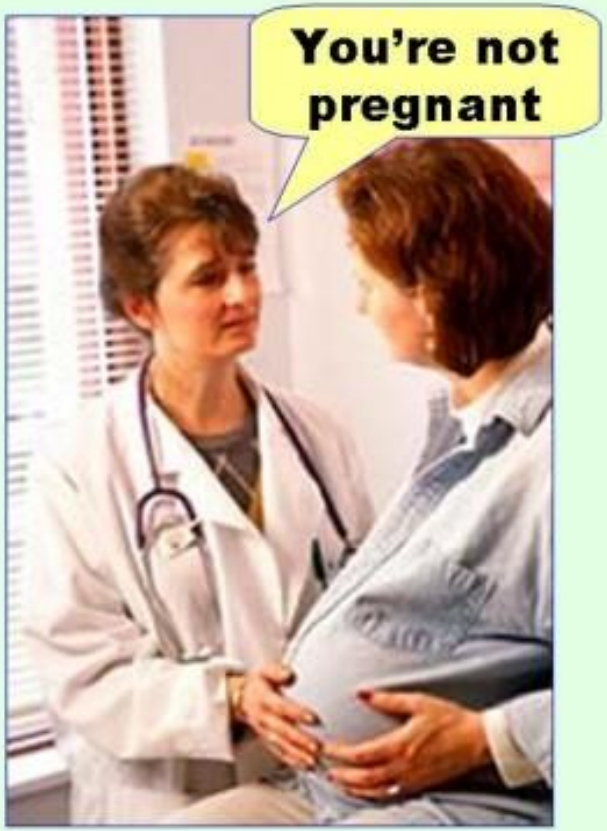
Joseph P. Simmons[1],
Leif D. Nelson[2] and
Uri Simonsohn[1]

P-hacking -> huge risk of type I error

Type I error
(false positive)



Type II error
(false negative)



Solutions

a. Using simulated datasets to give insight into statistical methods

BishopBlog

Ramblings on academic-related matters. For information on my res
http://psyweb.psy.ox.ac.uk/oscci/. Twin analysis blog: http://dbte
time-frequency analysis blog: bishoptechbits.blogspot.com/ . For

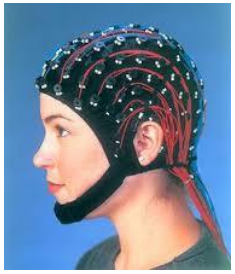
Wednesday, 5 October 2011

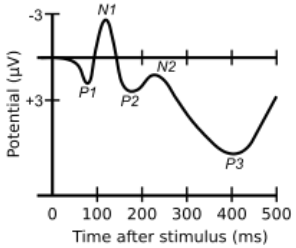
The joys of inventing data

Have I gone over to the dark side? Cracked under pressure from the REF to resort to fabrication of results to secure that elusive Nature paper? Or had my brain addled by so many requests for information from ethics committees that I've just decided that its easier to be unethical? Well readers will be reassured to hear that none of these things is true. What I have to say concerns the benefits of made-up data for helping understand how to analyse real data.

Illustrated with field of ERP/EEG

- Flexibility in analysis in terms of:
  - Electrodes
  - Time intervals
  - Frequency ranges
  - Measurement of peaks
  - etc, etc





- Often see analyses with 4- or 5-way ANOVA (group x side x site x condition x interval)

- Standard stats packages correct p-values for N levels WITHIN a factor, but not for overall N factors and interactions

Cramer AOJ, et al 2016. Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review 23:640-647*

4 way anova

run	group	task	side	loc	group: task	group: side	group: loc	task: side	task: loc	side: loc	group: task: side	group: task: loc	group: side: loc	task: side: loc	group: task: side: loc
1	0.056	0.119	0.896	0.371	0.052	0.469	0.588	0.825	0.299	0.979	0.183	0.117	0.391	0.780	**0.024**
2	0.931	0.576	0.391	0.637	0.463	0.419	0.691	0.442	0.608	0.360	0.424	0.977	0.281	0.583	0.990
3	0.161	0.565	0.430	**0.025**	0.642	0.743	**0.040**	0.268	0.955	0.130	0.784	0.763	0.516	0.296	0.921
4	0.142	0.812	0.935	0.762	0.454	0.105	0.521	0.497	0.861	0.413	0.191	0.405	0.948	0.073	0.185
5	0.393	0.486	0.955	0.889	0.963	0.688	0.482	0.104	0.250	0.779	0.237	0.224	0.770	0.310	0.176
6	0.584	0.755	0.350	0.063	0.768	0.481	0.098	0.365	0.062	0.733	0.372	**0.004**	0.521	0.178	0.263
7	0.742	0.071	0.481	0.471	0.152	0.091	0.254	0.485	0.537	0.609	0.601	0.598	0.545	0.359	0.744
8	0.215	0.594	0.572	0.659	0.551	0.930	0.937	0.220	0.642	0.171	0.118	0.126	**0.004**	0.138	0.143
9	0.642	**0.032**	**0.031**	0.458	0.250	**0.010**	0.676	0.988	0.230	0.289	0.833	0.437	0.843	0.161	0.134
10	0.564	0.707	**0.008**	0.930	0.516	**0.014**	0.730	0.211	**0.023**	0.790	0.465	**0.031**	0.581	0.236	0.314
11	0.779	0.616	0.234	0.880	0.765	0.090	0.521	0.592	0.291	0.174	0.896	0.244	**0.018**	**0.009**	0.184
12	0.703	**0.027**	0.619	0.158	**0.024**	0.551	0.383	0.109	0.969	0.262	0.276	0.372	0.445	0.249	0.508
13	**0.013**	0.641	0.560	0.171	0.672	0.995	0.184	0.585	0.688	**0.025**	0.683	0.755	**0.047**	0.450	0.537
14	0.866	0.060	0.520	0.462	0.238	0.404	0.279	0.637	0.718	0.950	0.646	0.959	0.504	0.189	0.283
15	0.331	0.606	0.089	**0.020**	0.921	0.313	**0.050**	0.122	0.203	0.470	0.091	0.757	0.441	0.594	0.663

Each row shows p-value outputs from a 4 way ANOVA applied to a new set of random data
See http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html

Solutions

b. Distinguish exploration from hypothesis-testing analyses

- Subdivide data into exploration and replication sets.
- Or replicate in another dataset

Replication in Genome-Wide Association Studies

Peter Kraft, Eleftheria Zeggini and John P. A. Ioannidis
Statistical Science
Vol. 24, No. 4 (November 2009), pp. 561-573

Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/25681332
Page Count: 13

Solutions

c. Preregistration of analyses

Science
Head quarters

Psychology's 'registration revolution'

Moves to uphold transparency are not only making psychology more scientific – they are harnessing our knowledge of the mind to strengthen science

Chris Chambers

Tuesday 20 May 2014 07.15 BST

Problems caused by researchers. 3

- Reluctance to collaborate with competitors
- Reluctance to share data
- Fabricated data

Solutions to these may require changes to incentive structures, which leads us to....

Problems caused by journals

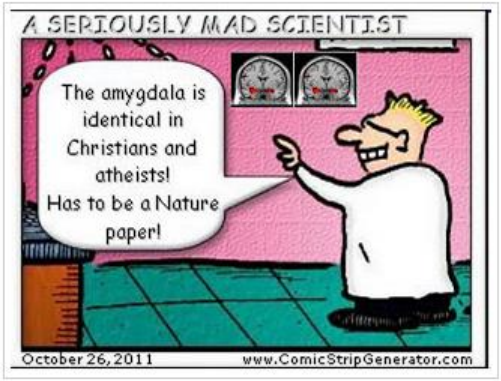
- More concern for newsworthiness than methods
  - Won't publish replications (or failures to replicate)
    - Won't publish 'negative' findings

Thursday, 19 January 2012

Novelty, interest and replicability

Wednesday, 26 October 2011

Accentuate the negative



Sunday, 10 March 2013

High-impact journals: where newsworthiness trumps methodology

BishopBlog

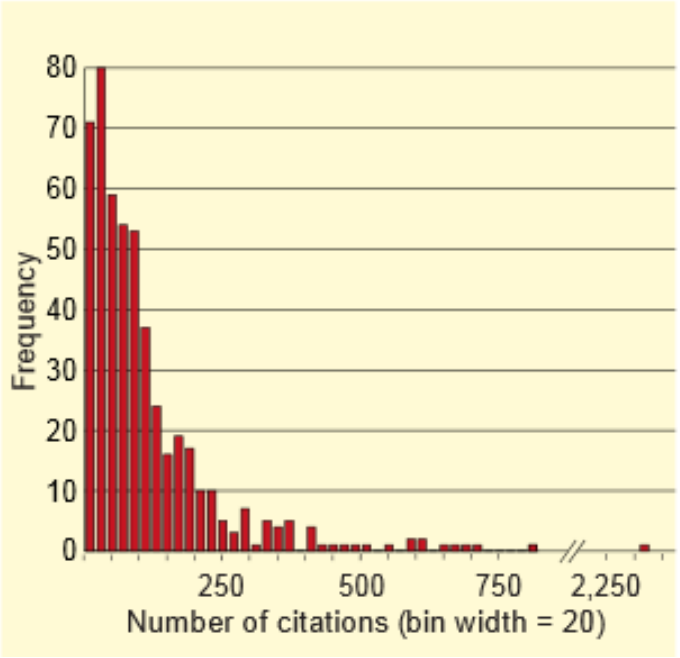
http://deevybee.blogspot.co.uk/

Problems caused by institutions

- Reward according to journal impact factor
- Reward those with most grant income

Problems with journal impact factors

- Impact factor not a good indication of the citations for individual articles in the journal, because distribution very skewed
- Typically, around half the articles have very few citations



N citations for sample of papers in Nature

http://www.dcscience.net/colquhoun-nature-impact-2003.pdf

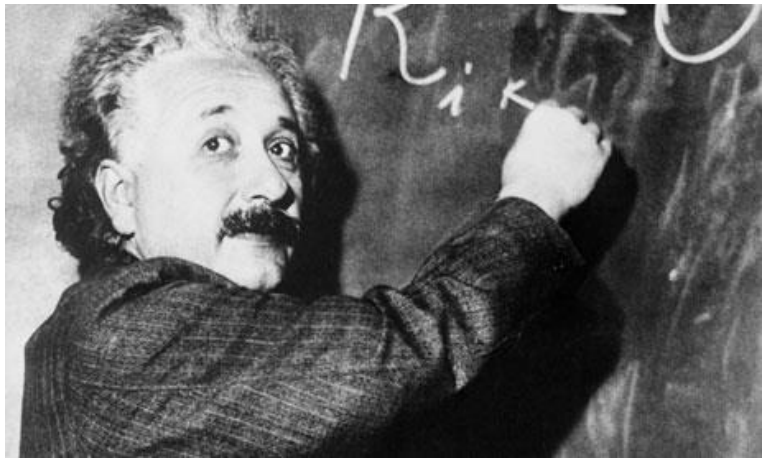
Income to institution increases with the amount of funding and so….

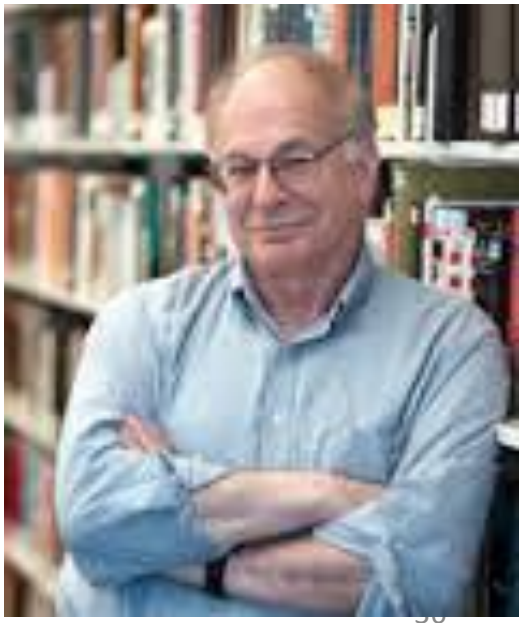
- The system encourages us to assume that:
  - Big grant is better than small grant
  - Many grants are better than one grant

"This is Dr Bagshaw, discoverer of the infinitely expanding research grant“

This is counterproductive because

- Amount of funding needed to do research is not a proxy for value of that research
- Some activities intrinsically more expensive
- Does not make sense to disfavour research areas that cost less





Daniel Kahneman

Furthermore….

- Desperate scramble for research funds leads to researchers being overcommitted -> poorly conducted studies
- Ridiculous amount of waste due to the ‘academic backlog’

Solutions for institutions

"It is time to remedy a flawed bibliometric-based assessment for young scientists."

Marcia McNutt
Science 2014 • VOL 346 ISSUE 6214

- Consider 'bang for your buck' rather than amount of grant income
- Reward research reproducibility over impact factor in evaluation
- Reward those who adopt open science practices

Scientific rigor and the art of motorcycle maintenance

Marcus Munafò, Simon Noble, William J Browne, Dani Brunner, Katherine Button, Joaquim Ferreira, Peter Holmans, Douglas Langbehn, Glyn Lewis, Martin Lindquist, Kate Tilling, Eric-Jan Wagenmakers & Robi Blumenstein

The reliability of scientific research is under scrutiny. A recently convened working group proposes cultural adjustments to incentivize better research practices.

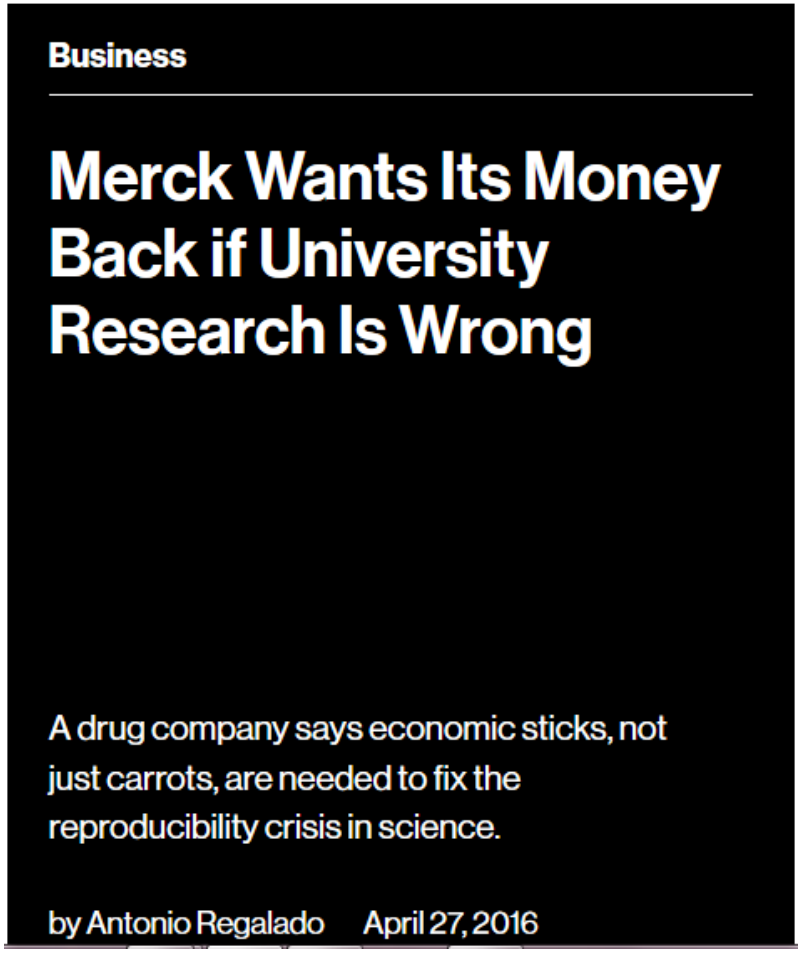
Nat Biotech, 32(9), 871-873. doi: 10.1038/nbt.3004

Problems caused by funders

- Don’t require that all data reported
  - Though growing interest in data sharing
- No interest in funding replications
- No interest in funding systematic reviews

Solutions

Funding contingent on adoption of reproducible practices



Business

Merck Wants Its Money Back if University Research Is Wrong

A drug company says economic sticks, not just carrots, are needed to fix the reproducibility crisis in science.

by Antonio Regalado April 27, 2016

https://www.technologyreview.com/s/601348/merck-wants-its-money-back-if-university-research-is-wrong/

NIH plans to enhance reproducibility

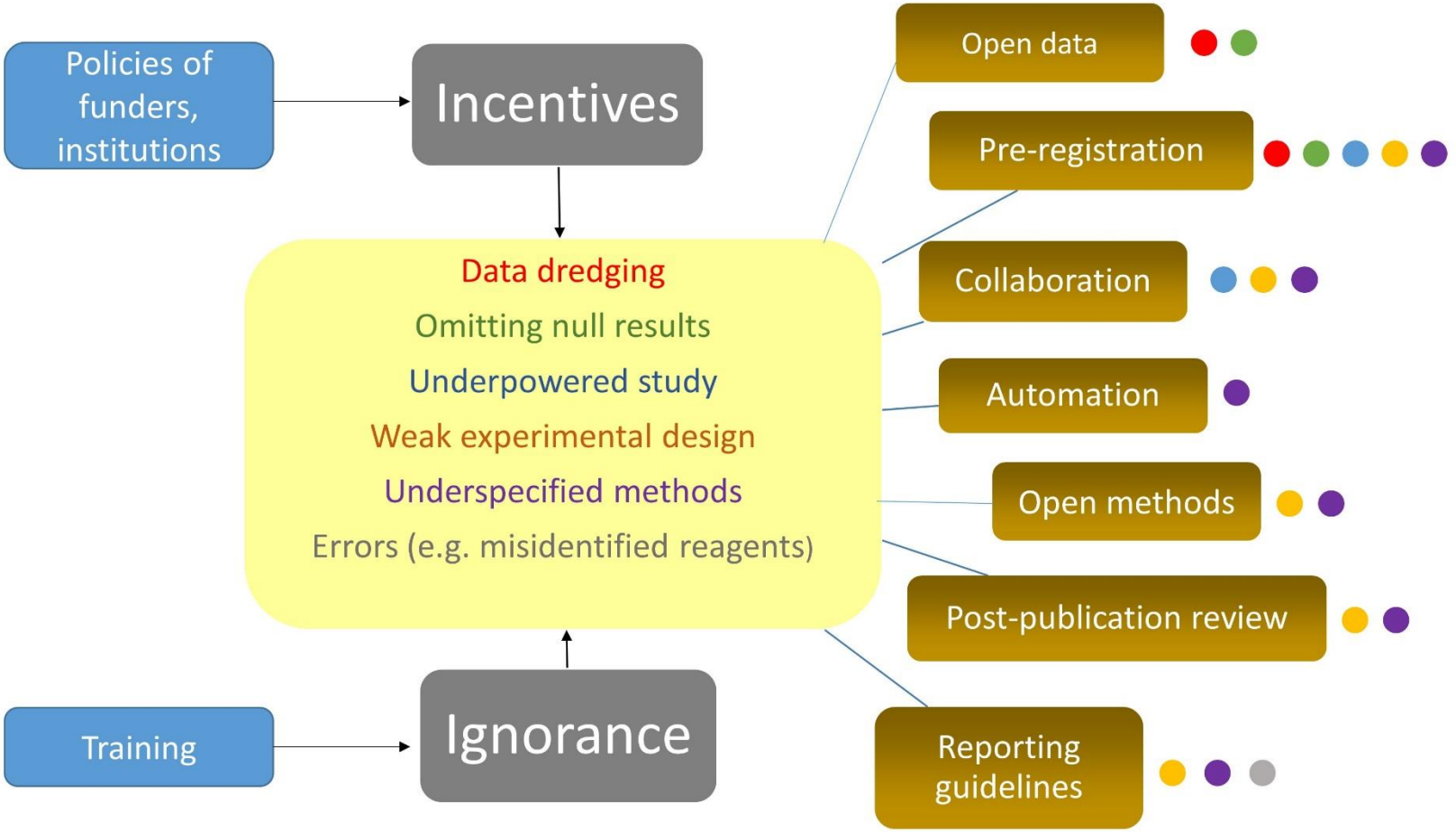
Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

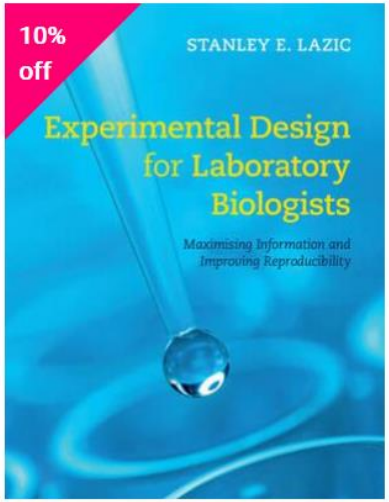
A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring[1,2]. As leaders of the US National Institutes of Health (NIH), we share this concern and here explore some of the significant interventions that we are planning.

Science has long been regarded as 'self-correcting', given that it is founded on the replication of earlier work. Over the long term, that principle remains true. In the shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today's researchers to reproduce others' findings.

Let's be clear: with rare exceptions, we have no evidence to suggest that irreproducibility is caused by scientific misconduct. In 2011, the Office of Research Integrity of the US Department of Health and Human Services pursued only 12 such cases[3]. Even if this represents only a fraction of the actual problem, fraudulent papers are vastly

Academy of Medical Sciences, 2015
Report on Reproducibility and Reliability of Biomedical Research

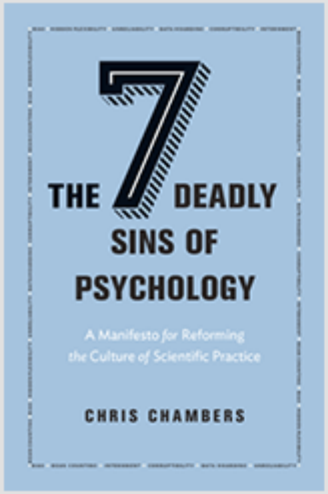




Coming very soon!

Experimental Design for Laboratory Biologists : Maximising Information and Improving Reproducibility

Stanley E Lazic



The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice

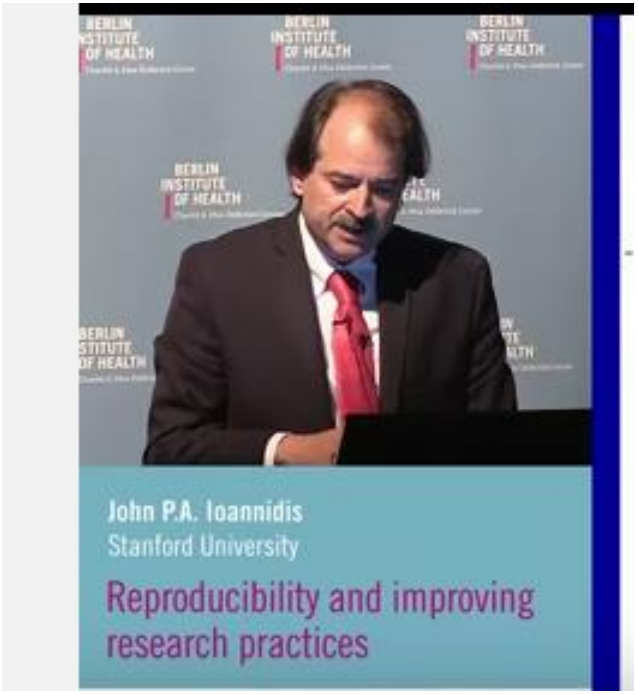
Chris Chambers

Hardcover | April 2017 | **$29.95** | **£22.95** | ISBN: 9780691158907
288 pp. | 6 x 9 | 8 halftones. 9 line illus.

Add to Shopping Cart

eBook | ISBN: 9781400884940 |
Our eBook editions are available from these online vendors

Endorsements



https://www.youtube.com/watch?v=xGLF6olIZYY

1st BHA Annual Special Lecture:
John P. A. Ioannidis

YouTube GB

Free Coursera lectures

Improving your statistical inferences

Daniel Lakens
Associate Professor
Department of Human-Technology Interaction
Eindhoven University of Technology

https://www.coursera.org/learn/statistical-inferences