Introduction to Reproducibility in the Life Sciences

Jean-Baptiste Poline

MNI, Brain Imaging Centre, McGill, Montreal
HWNI, UC Berkeley

Part I: Reproducibility: background

Part II : Etiology of Irreproducibility

Part III :Some therapeutic proposals

Part I: Reproducibility: background

Part II : Etiology of Irreproducibility

Part III :Some therapeutic proposals

Amgen replication

- 53 papers examined at Amgen in preclinical cancer research
- Papers were selected that described something completely new and in very high impact factor journals
- **Scientific findings were confirmed in only 6 (11%)**

Begley and Ellis, Nature, 2012

JAMA Psychiatry 2017

Altered Brain Activity in Unipolar Depression Revisited Meta analyses of Neuroimaging Studies

Veronika I. Müller, PhD, Edna C. Cieslik, PhD, Ilinca Serbanescu, MSc, Angela R. Laird, PhD, Peter T. Fox, MD, and Simon B. Eickhoff, MD

RESULTS—In total, 57 studies with 99 individual neuroimaging experiments comprising in total 1058 patients were included; 34 of them tested cognitive and 65 emotional processing. Overall analyses across cognitive processing experiments ($P > .29$) and across emotional processing experiments ($P > .47$) revealed no significant results. Similarly, no convergence was found in analyses investigating positive (all $P > .15$), negative (all $P > .76$), or memory (all $P > .48$) processes. Analyses that restricted inclusion of confounds (eg, medication, comorbidity, age) did not change the results.

Imaging Genetics GWAS

Stein et al., 2012, Nature Genetics, study of the hippocampal volume in more than 10k+7k subjects

Previously identified candidate polymorphisms associated with hippocampal volume in general showed little association within our meta-analysis :(

Stein et al, Nat. Gen. 2013

Credibility Crisis


Los Angeles Times | BUSINESS

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH STYLE TRAV

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.


Science

AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

All Science Journals

AAAS NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Article Views
- Summary
- Full Text
- Full Text (PDF)

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints
- View PubMed Citation

Related Content

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

‹ Prev | Table of Contents | Next ›

Read Full Text to Comment (8)

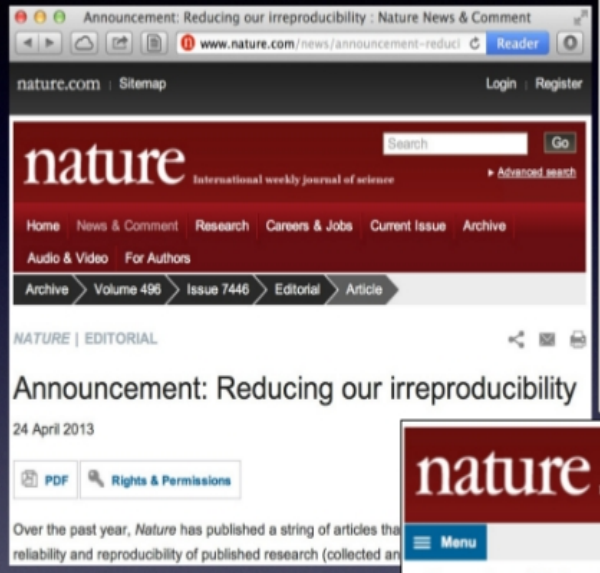
EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

Science advances on a foundation of trusted d
approach that scientists use to gain confidenc
community was shaken by reports that a troub
not reproducible. Because confidence in result
community, we are announcing new initiatives
Science. For preclinical studies (one of the targ
recommendations of the U.S. National Institute
increasing transparency.* Authors will indicate
handling (such as how to deal with outliers), wh
ensure a sufficient signal-to-noise ratio, wheth
experimenter was blind to the conduct of the e
guidelines.


Announcement: Reducing our irreproducibility : Nature News & Comment

www.nature.com/news/announcement-reduci Reader

nature.com | Sitemap Login | Register

nature International weekly journal of science

Search Go

Advanced search

Home News & Comment Research Careers & Jobs Current Issue Archive

Audio & Video For Authors

Archive › Volume 496 › Issue 7446 › Editorial › Article

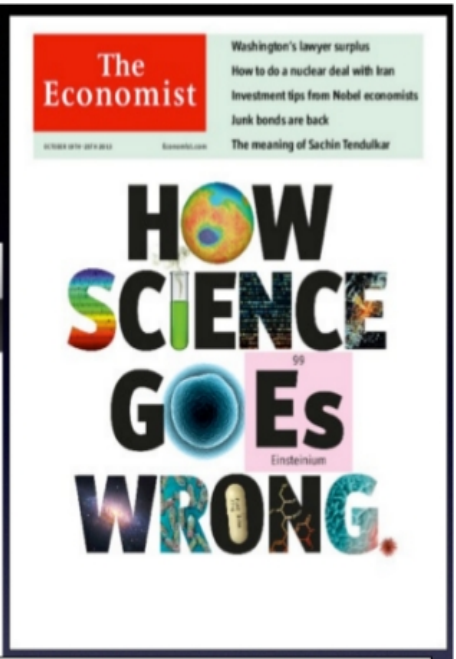
NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

PDF Rights & Permissions

Over the past year, *Nature* has published a string of articles tha
reliability and reproducibility of published research (collected an


The Economist

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.


nature International weekly journal of science

Menu Advanced search Search Go

archive › volume 483 › issue 7391 › editorials › article

NATURE | EDITORIAL

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

PDF Citation Reprints Rights & permissions Article metrics

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.


TheScientist

EXPLORING LIFE. INSPIRING INNOVATION

NIH Tackles Irreproducibility

The federal agency speaks out about how to improve the quality of scientific research.

By Jef Akst | January 28, 2014

Funding agencies reaction

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.

Collins and Tabak. 2014. Nature 505: 612–13.

The problem is widespread

Essay

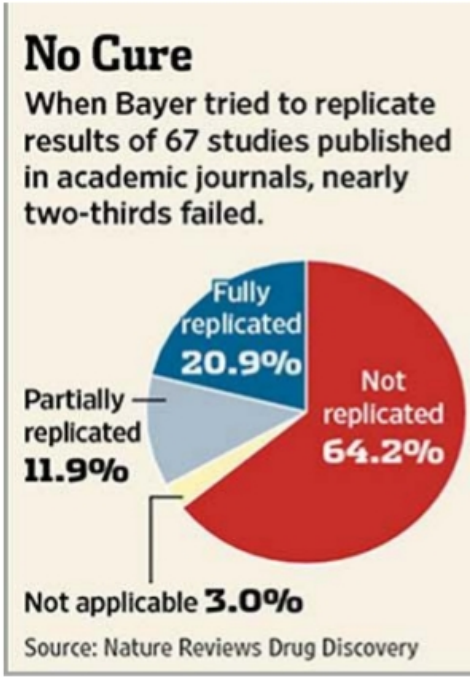
Why Most Published Research Findings Are False

John P. A. Ioannidis

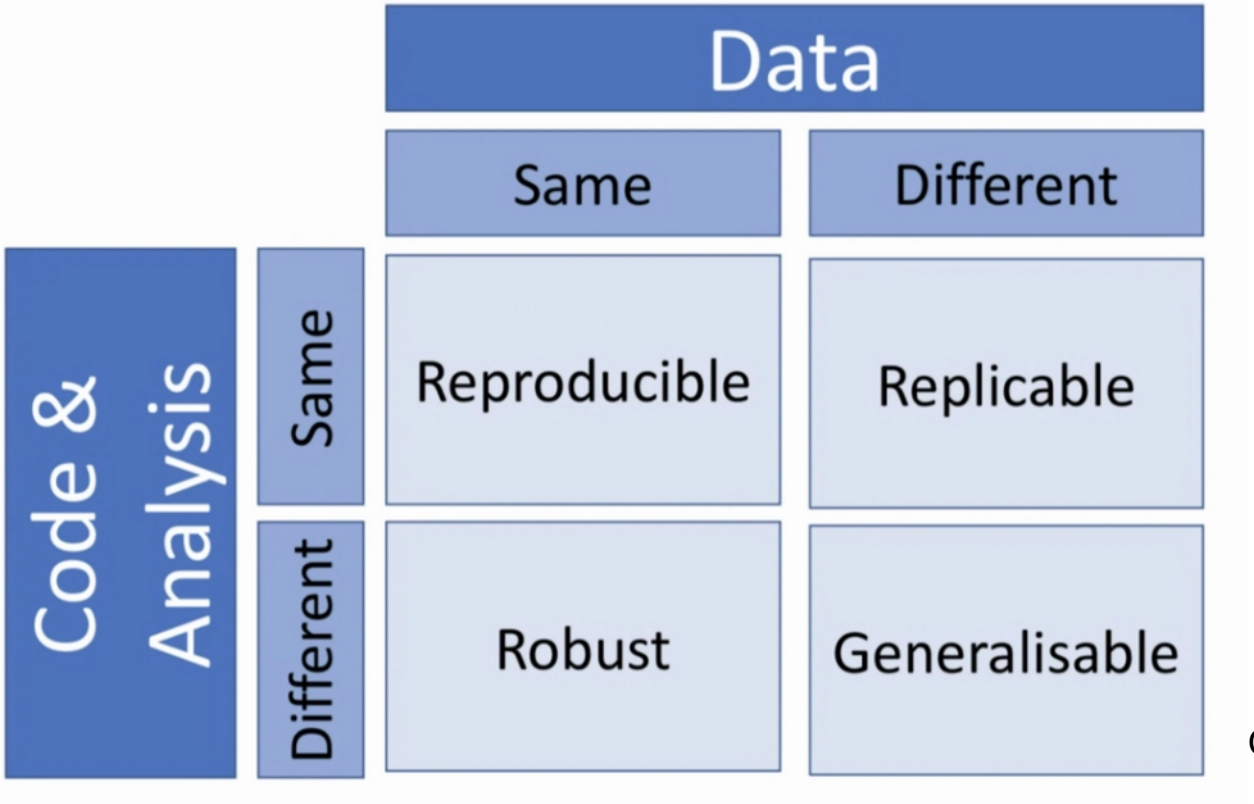
2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

"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*



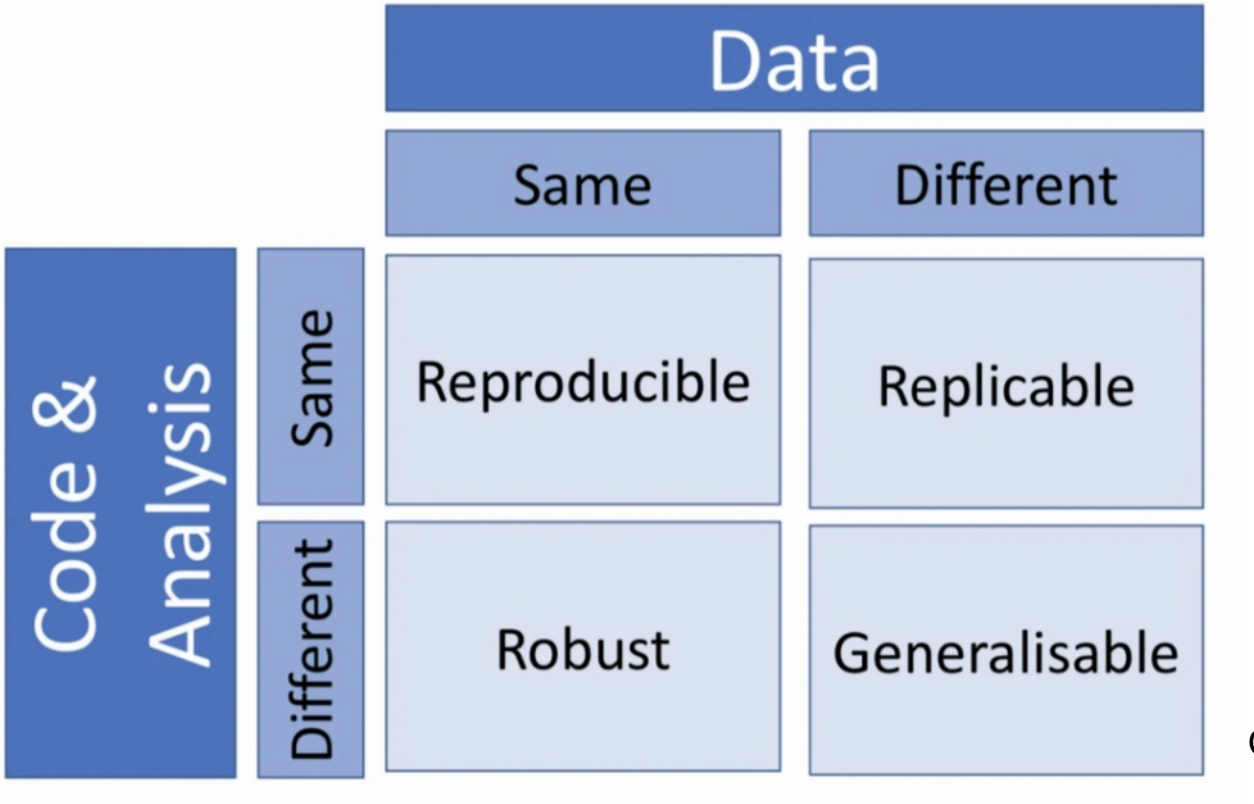
Some Definitions



Credit: J.Pineau

- My (prefered) way of thinking about reproducibility: **only talk about generalizability** across … (Data, Software, Time, Scanner, Stimuli, … etc)

Some Definitions



Credit: J.Pineau

- My (prefered) way of thinking about reproducibility: **only talk about generalizability** across … (Data, Software, Time, Scanner, Stimuli, … etc)

Part I: Reproducibility: background

Part II : Etiology of Irreproducibility

Part III :Some therapeutic proposals

Three causes

1. Poor statistical procedures

2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

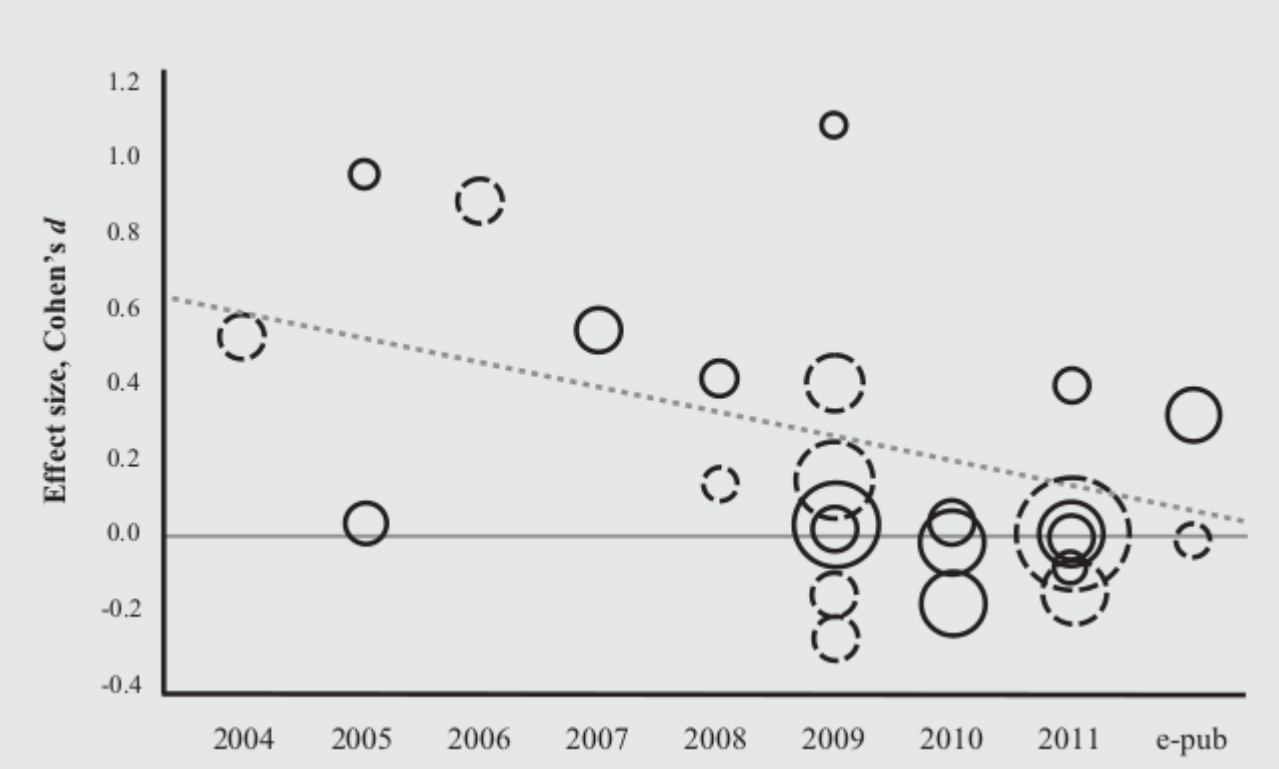
Three causes

1. Poor statistical procedures
    - Small Ns and effect sizes
    - Power issues
    - P-hacking

2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

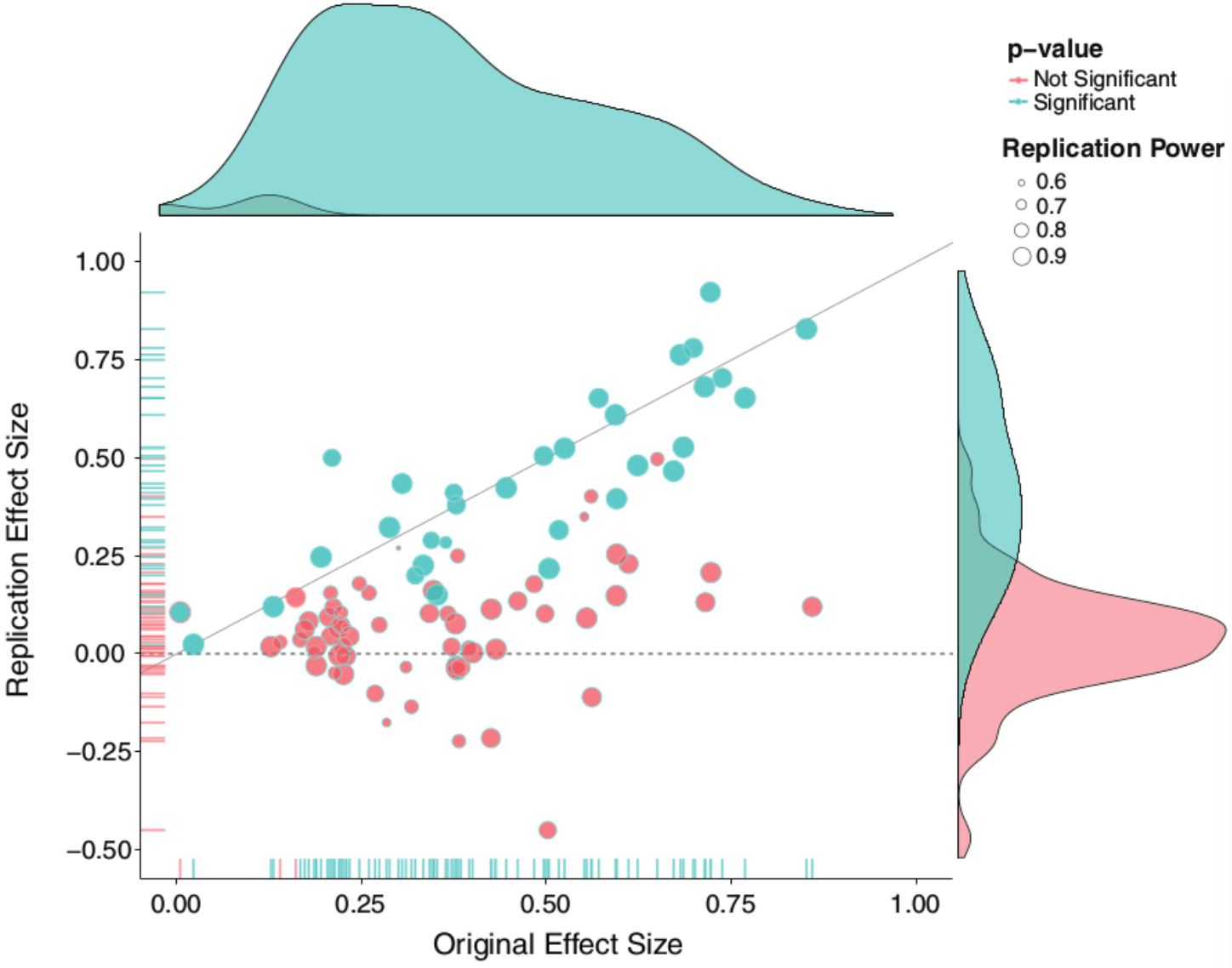
Statistics: ~~The~~ One problem



Molendijk, 2012: BDNF and hippocampal volume

See also : Mier, 2009: COMT and DLPFC

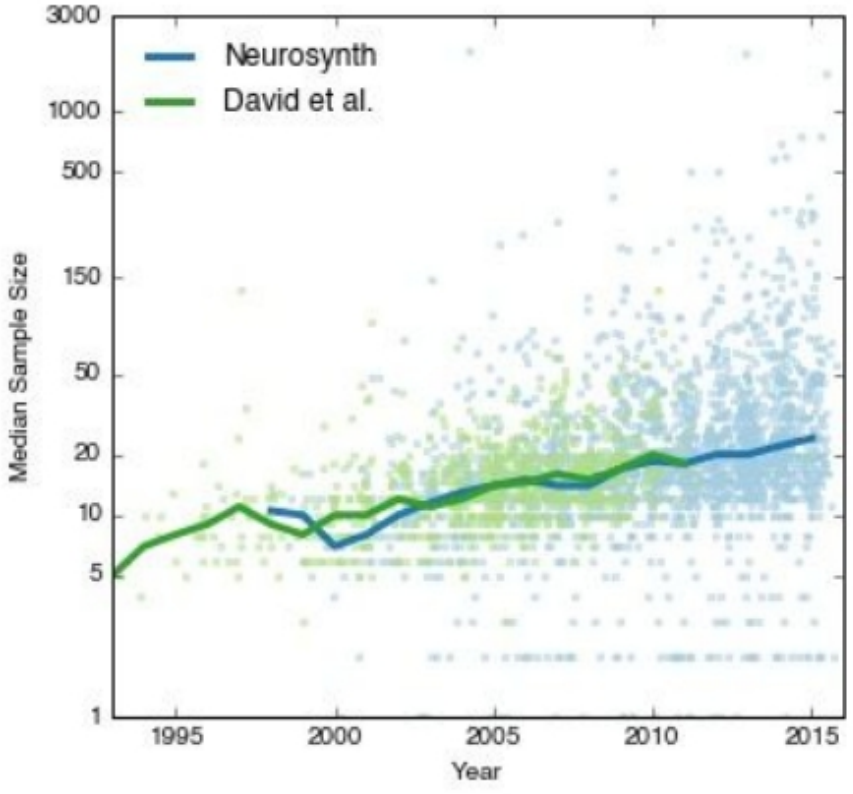
B. Nosek, Estimating the reproducibility of psychological science, Science 2015

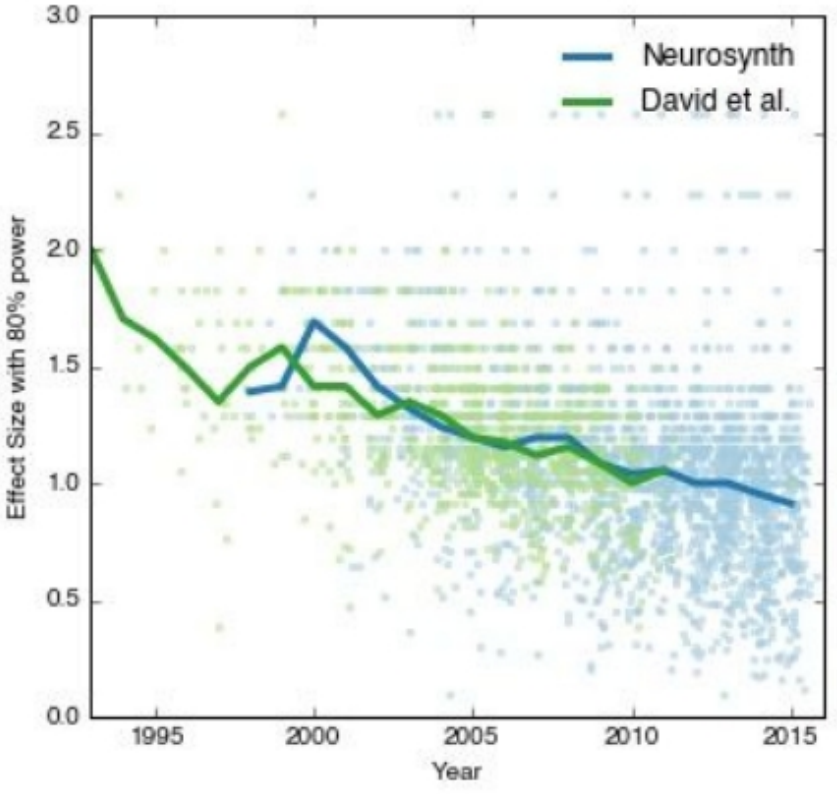


* The mean **effect size** (r) of the replication effects (M r = 0.197, SD = 0.257) **was half the magnitude** of the mean effect size of the original effects (M r = 0.403, SD = 0.188)

* **39%** of effects were rated to have replicated the original effect

Feeling the Future





Poldrack et al., PNAS, 2016

Feeling the Future

Paradigm	Intersection mask	mask size (vox)	Cohen D			BOLD		
			P10	median	P90	P10	median	P90
MOTOR	Bilateral Precentral Gyrus	12894	0.158	0.628	1.070	0.505	2.707	8.582
	Bilateral Supplementary motor cortex	3418	0.211	0.716	1.197	0.911	4.033	12.510
	Left putamen	1532	0.114	0.513	0.864	0.586	2.388	4.318
	Right putamen	1437	-0.008	0.369	0.749	-0.045	1.696	3.609
WM	Bilateral Middle frontal gyrus	7116	0.101	0.474	0.837	0.130	0.986	2.504
EMOTION	Left amygdala	1133	0.265	0.534	1.065	0.516	1.198	3.379
	Right amygdala	1082	0.308	0.645	1.140	0.581	1.350	3.557
GAMBLING	Left accumbens	455	0.138	0.310	0.461	0.369	0.849	1.440
	Right accumbens	417	0.141	0.332	0.488	0.373	0.981	1.618

With effect size = 0.5 => Power ~ 30%

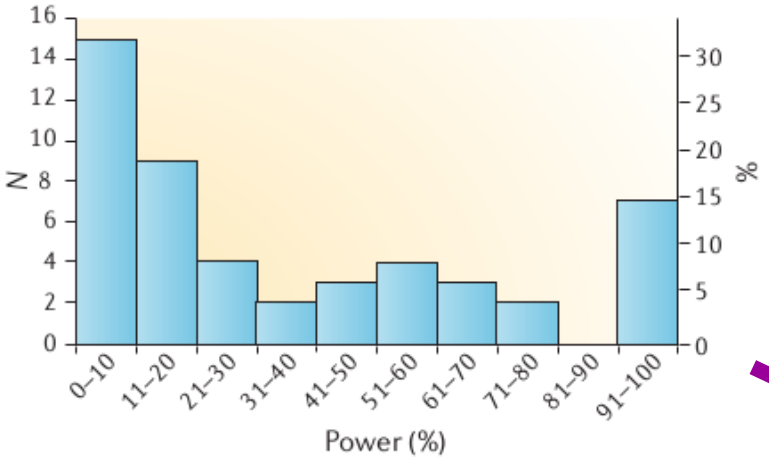
Power issues

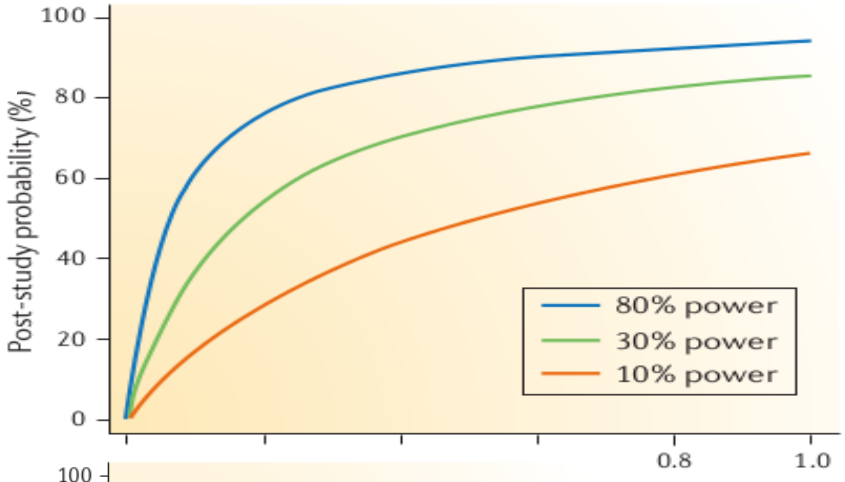
Open access, freely available online

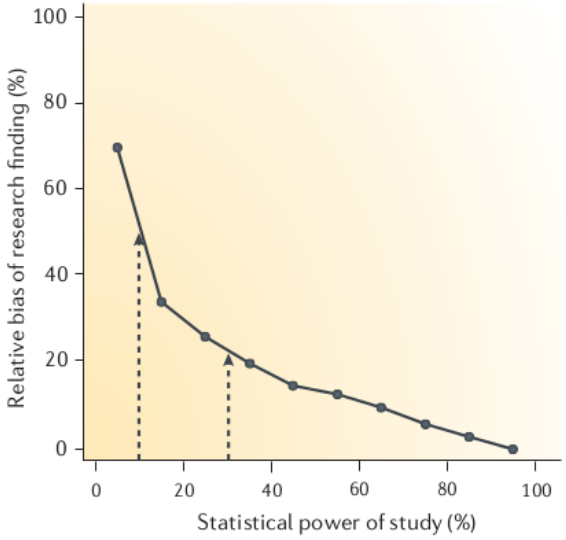
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis







Button et al., NNR, 2013

Positive Predictive Value

Open access, freely available online

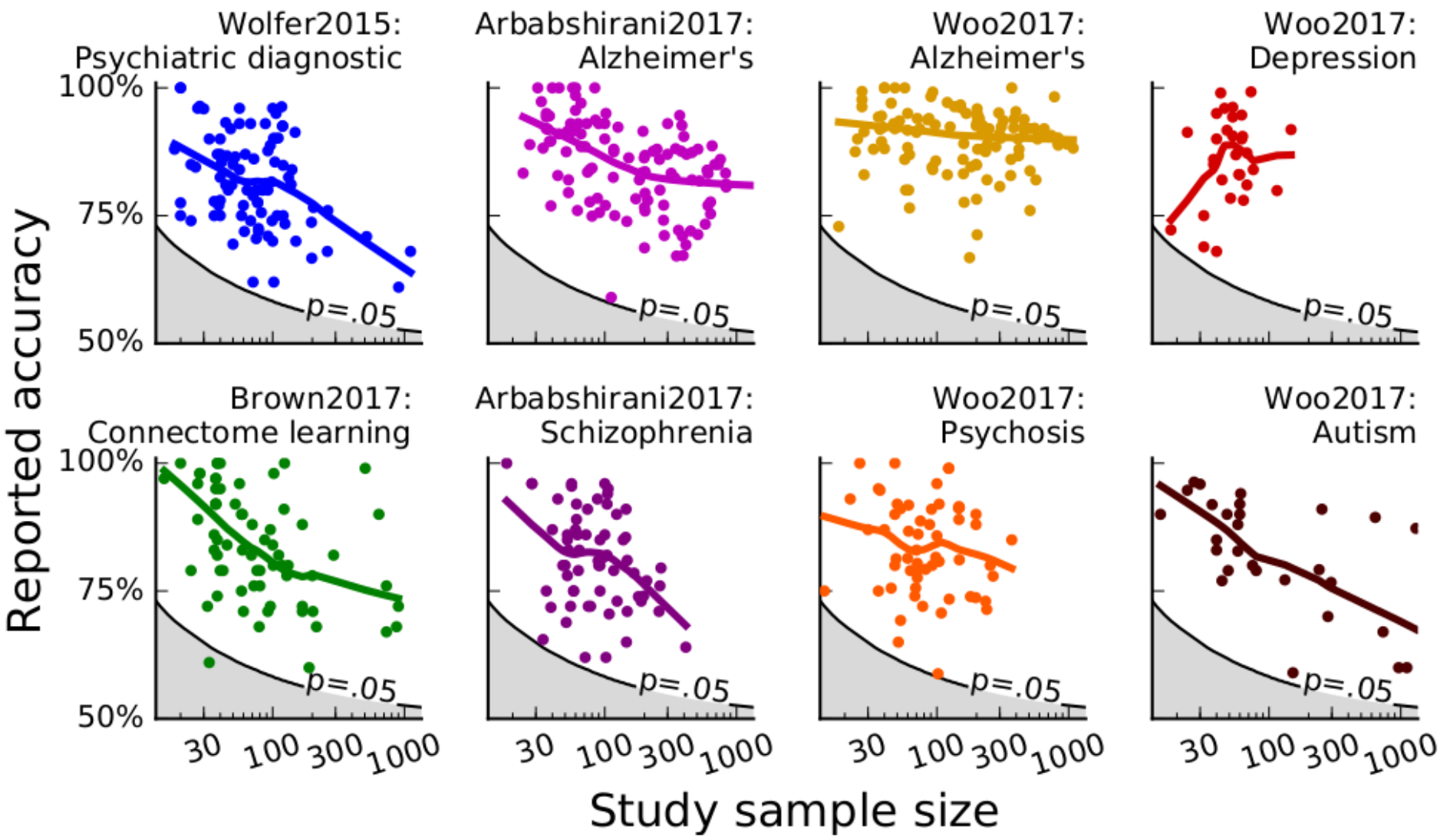
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- Positive Predictive Value : The probability that the alternative hypothesis is true knowing that the test is significant

Sample size issue in ML



Varoquaux, 2017, PsyArXiv, Neuroimage, Cross validation failure

Ioannidis 2005

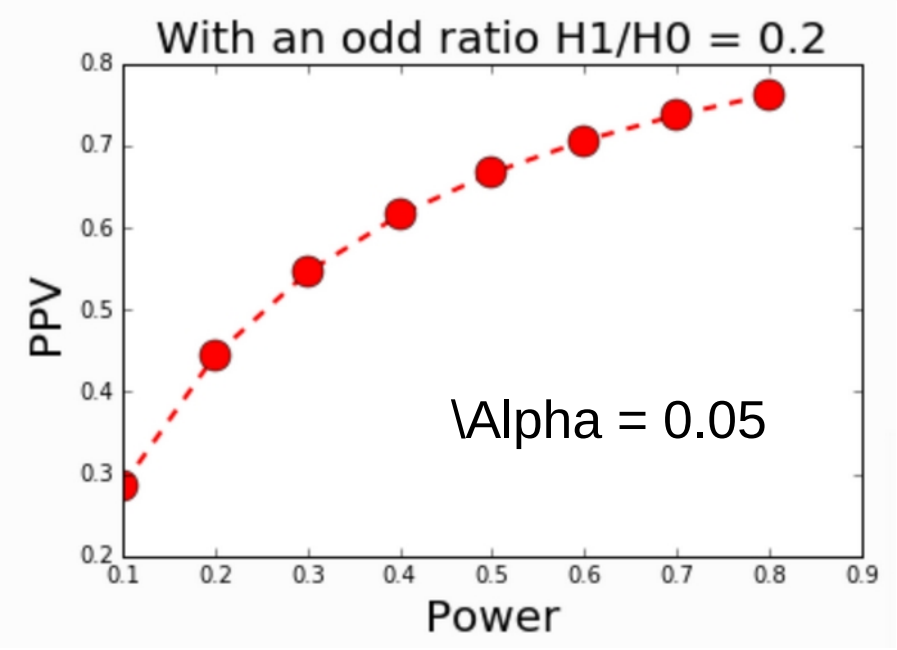
Open access, freely available online

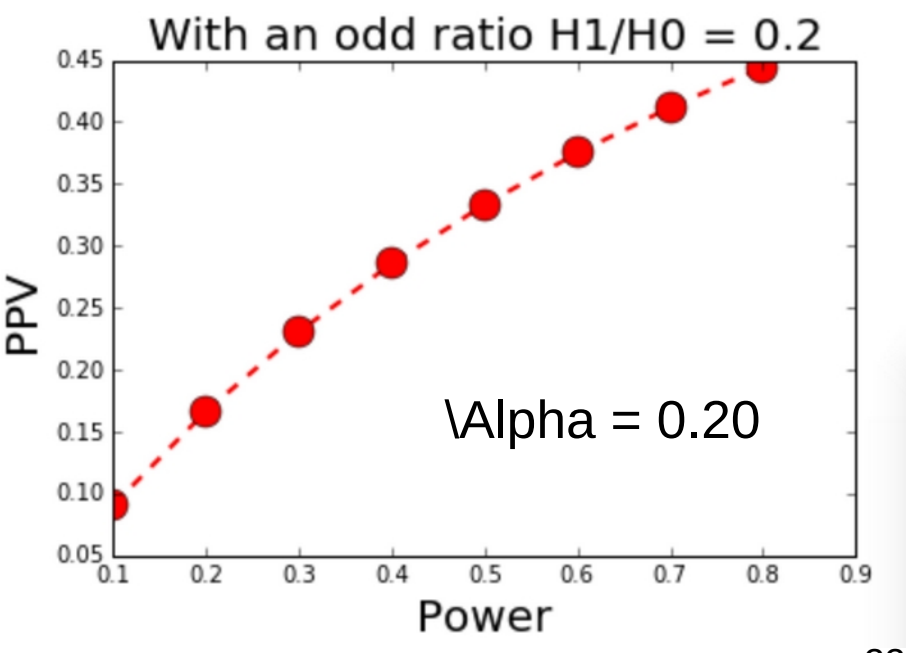
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

$$P(H_A \mid T_S) = \frac{WR}{WR + \alpha}$$





P-hacking

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]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

P-hacking - Harking

- Usually performed "in good faith"
  - I have forgotten a covariable
  - I haven't excluded some outliers
  - I should have included an interaction
  - I should only test subjects with X
  - …
- The "cost of seeing data"
  - Proper statistical test for inference require that you are "blind" to the data
- Harking : Hypothesising after seeing the results
  - You find something a little unexpected
  - You look in the literature : there are a few papers related
  - You reformulate your hypothesis based on these papers

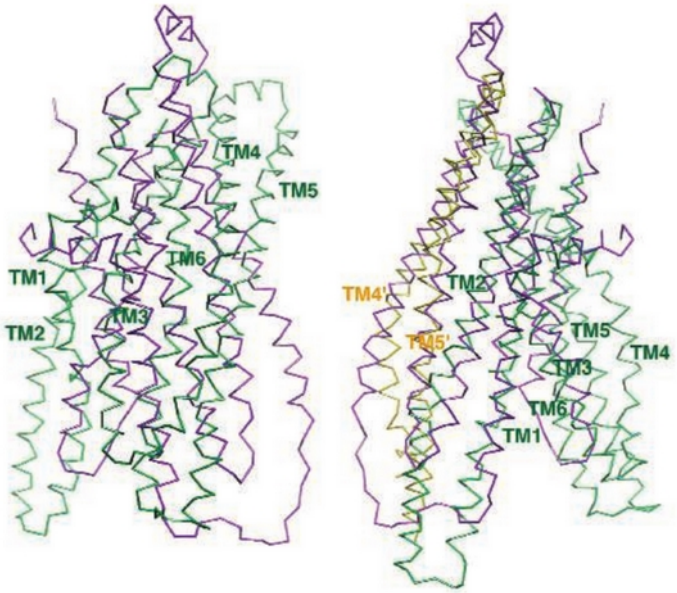
Three causes

1. Poor statistical procedures

2. Issues in data and software
    - Research software issues: motivation and examples
    - Across OS / implementation
    - Across sofware and parametrization
    - Across pipeline: “analytical flexibility”
    - Issues in data

3. A cultural issue: Publication practices and research incentives

Protein structure flip



Flipping fiasco. The structures of MsbA (purple) and Sav1866 (green) overlap little (*left*) until MsbA is inverted (*right*).

- G. Chang: 3 Science, 1 PNAS, 1 J Mol Biol retracted
- "… a homemade data-analysis program had flipped two columns of data...",
- "… inherited from another lab..."
- The code was distributed and used by others

Forensic Analysis

- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, "Forensic analysis", Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug, use expression profiles to build a model that predicts patient response

Baggerly and Coombes Forensic:

"***with poor documentation and irreproducibility even well meaning investigator may argue for drug that are contraindicated to some patients***"

"**the most common errors are simple**
(e.g.,row or column offsets); conversely,
the most simple errors are common."

Across OS
Across implementations

Same software different OS version

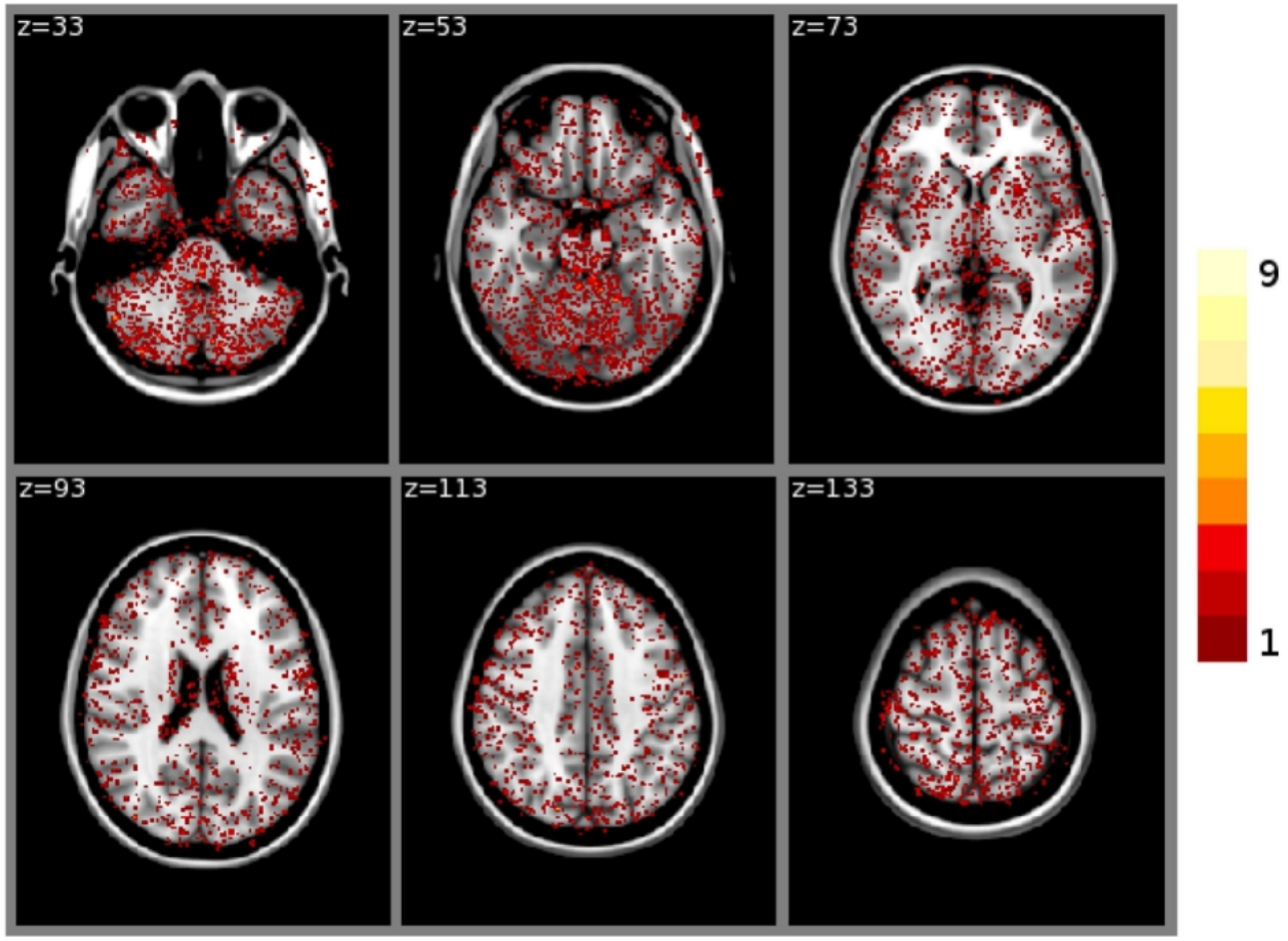
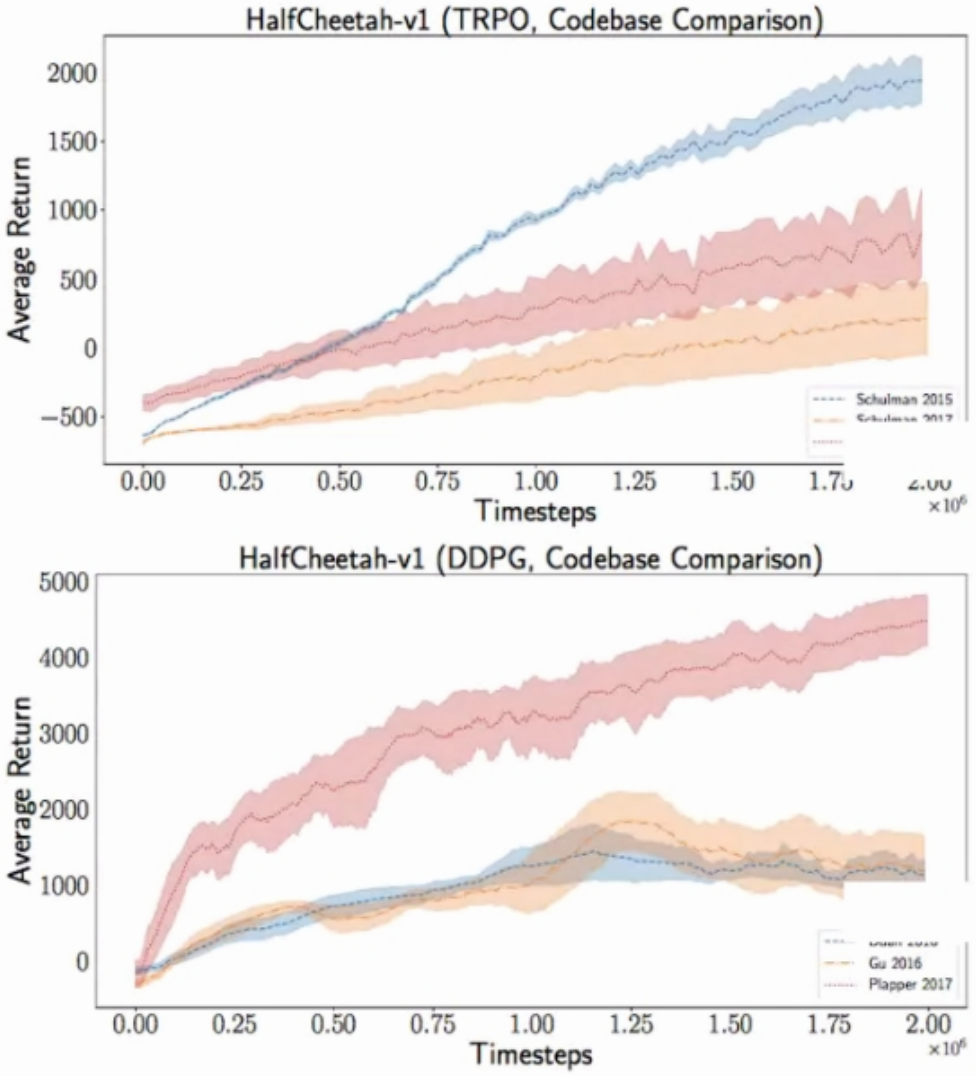


FIGURE 2 | Sum of binarized differences between cortical tissue classifications obtained on cluster A and cluster B (FSL FAST, build 1, *n* = 150 subjects). All binarized differences were resampled to the default MNI152 volume template.

Glatard et al, 2015, F. in Neuroinformatics

Example in machine learning

- Same algorithm
  - Top graph: TRPO
  - Bottom graph: DDPG
- Same domain
- Simulation environment
- Different implementations

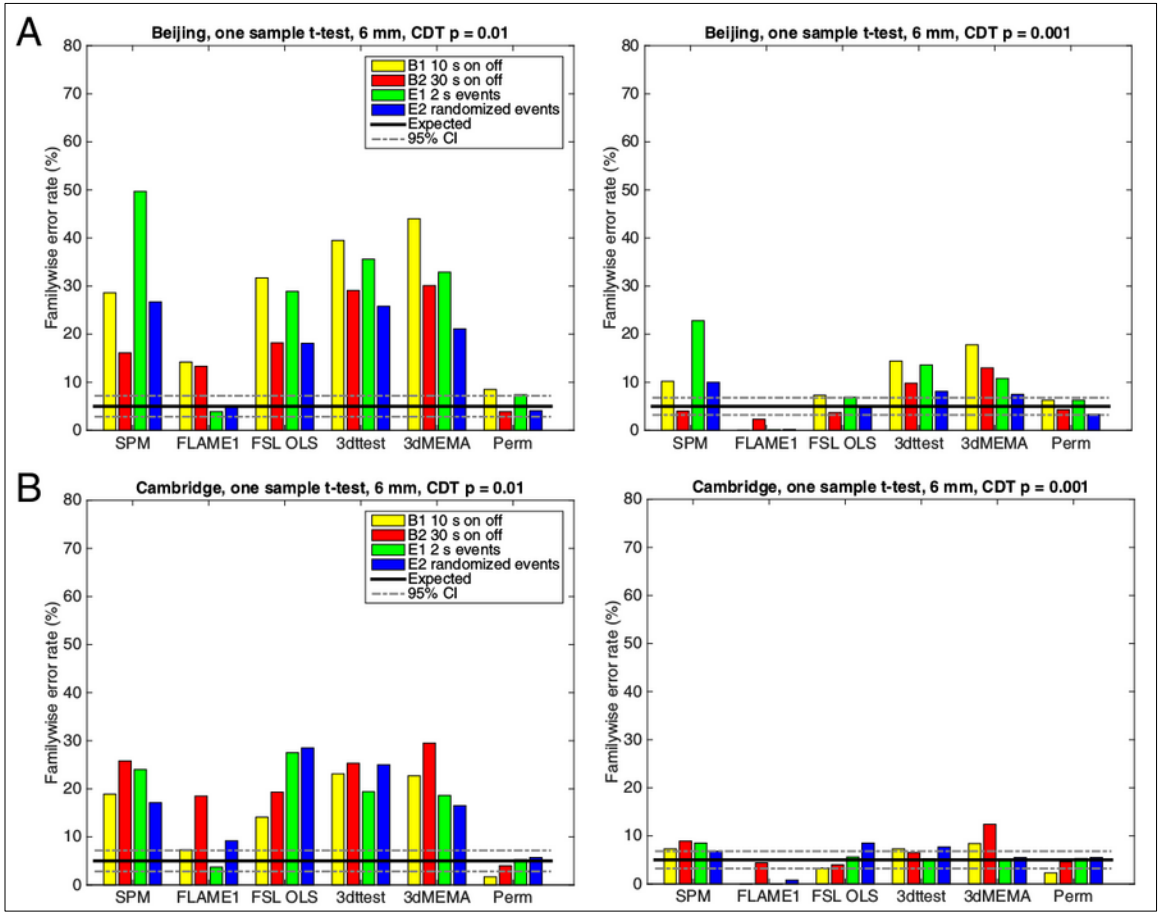


Credit : Joelle Pineau

Across parametrizations of software

and across software

“Cluster failure”? Or RFT misuse?



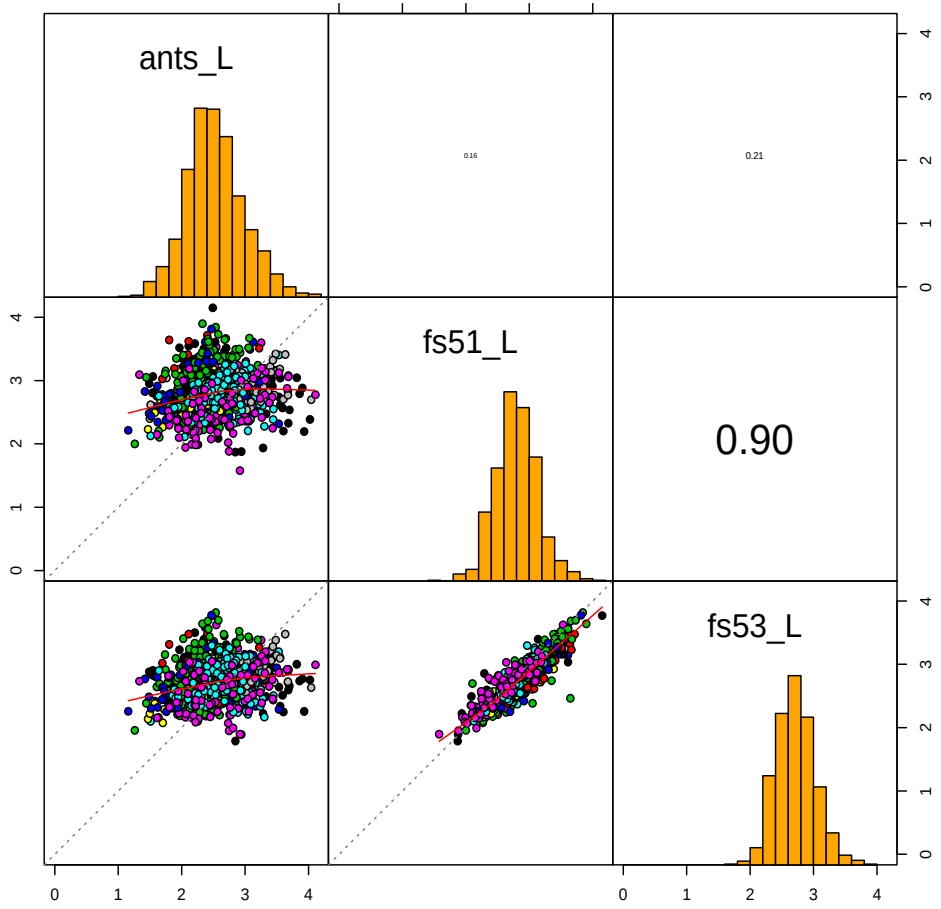
Eklund et al., PNAS, 2016 :

- Low threshold issue
- High threshold issue with Paradigm E1 ?
- Ad hoc procedure leads to around 70% FPR

-
- Estimated 3,500 papers affected by low threshold ?
- But 13000 w/o multiple comparisons ?

ANTS – FS5.1- FS5.3

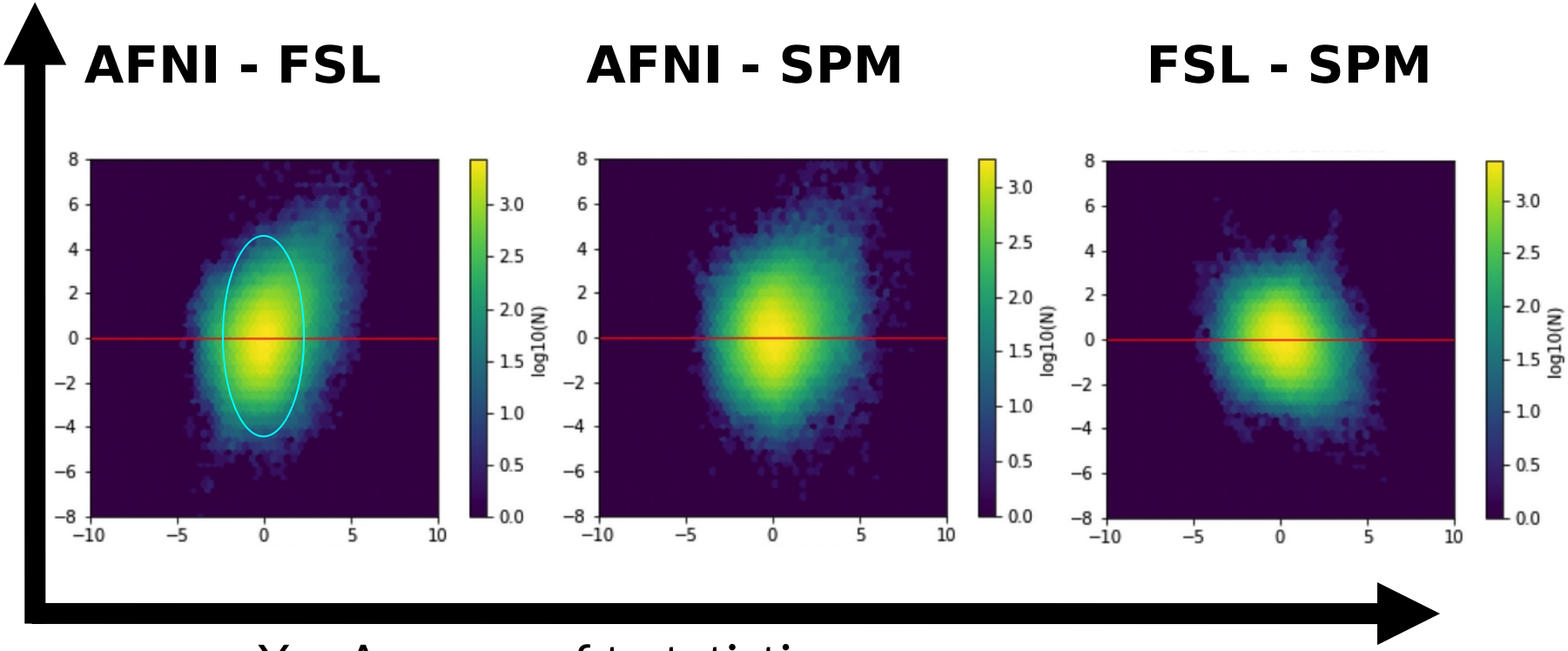
Size of the left caudal anterior Cingulate



Dickie E et al., 2017

How close are results on the same dataset?

Y = Diff. of t-statistics



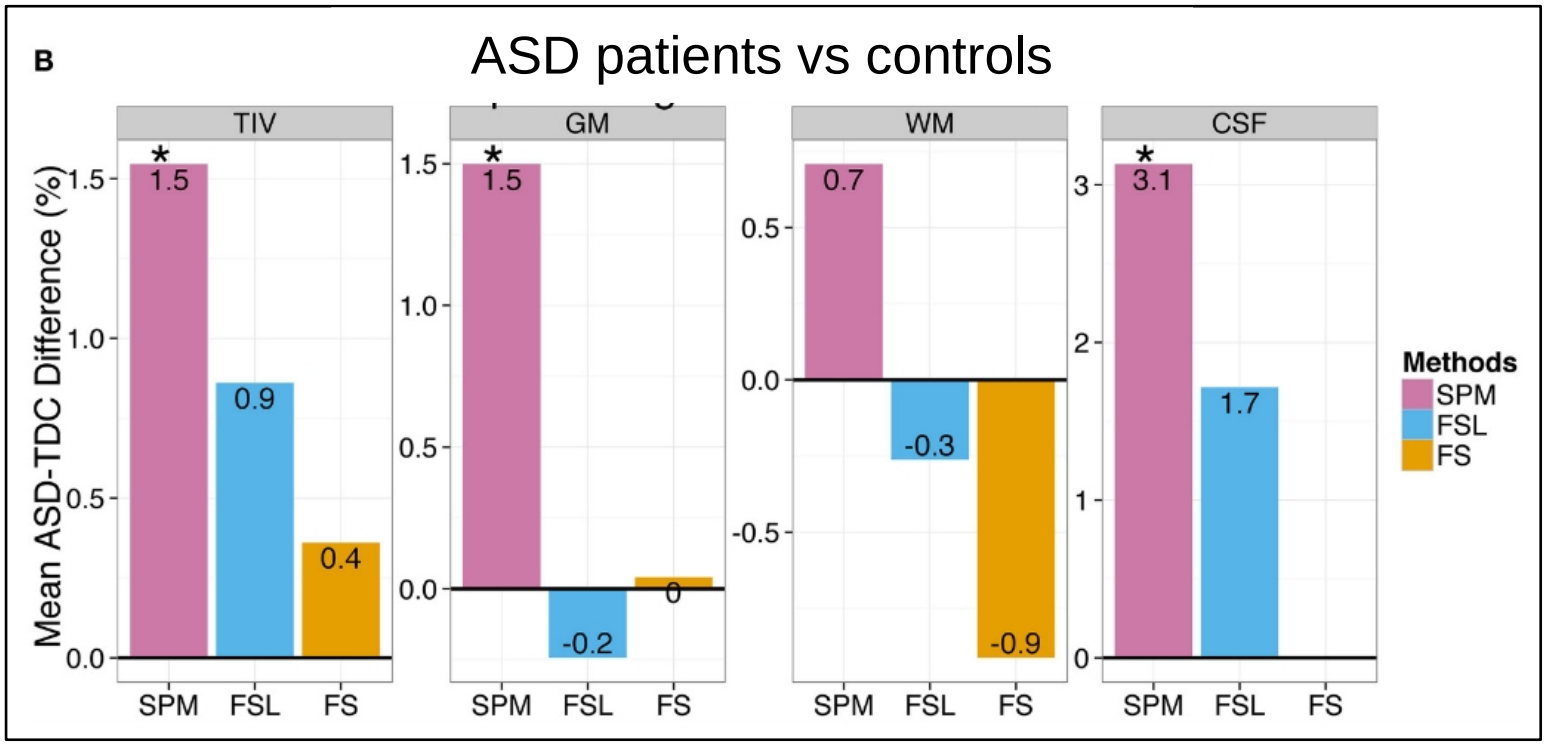
X = Average of t-statistics

▷ Plots similar to expected variation if **independent** was fed into each package

Alex Bowring, Camille Maumet, Thomas Nichols

Software, version, OS

G. Katuwal, f. in Brain Imaging Methods, 2016



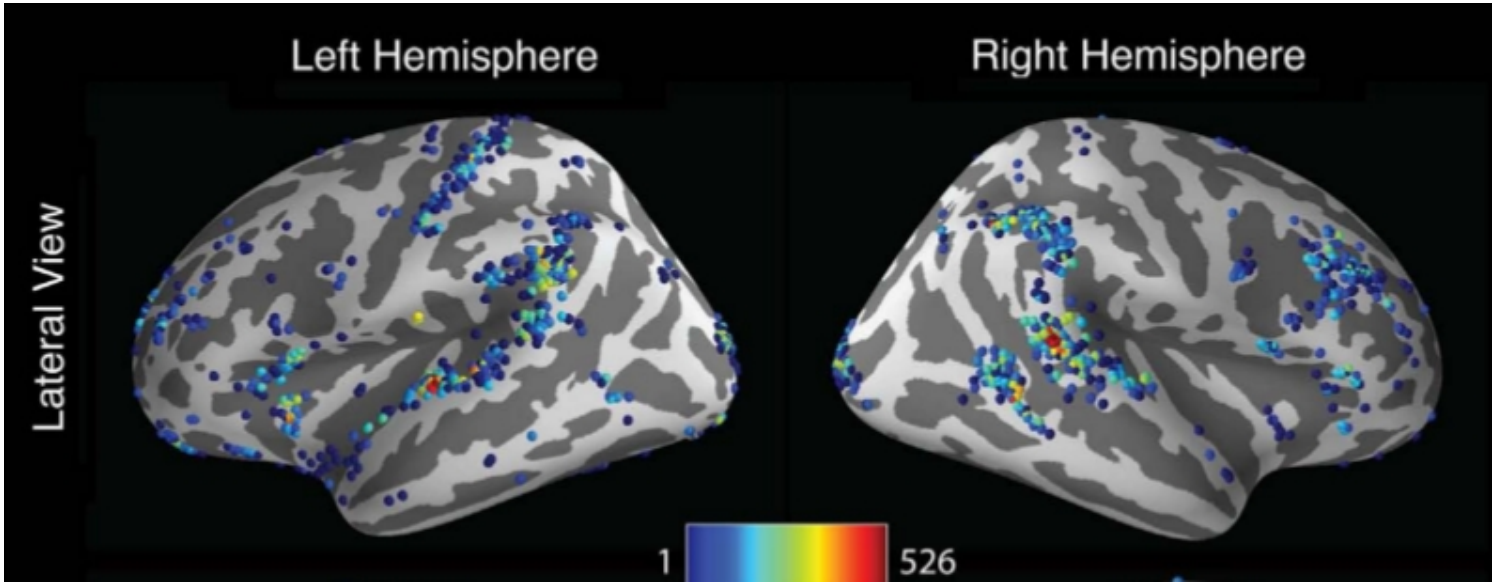
- Change from FSL to SPM?
- Change from v.1.12 to v.2.1 ?
- Change from cluster A to cluster B? Glatard et. al., finsc, 2015

Across possible pipelines

Estimating analytical flexibility of fMRI

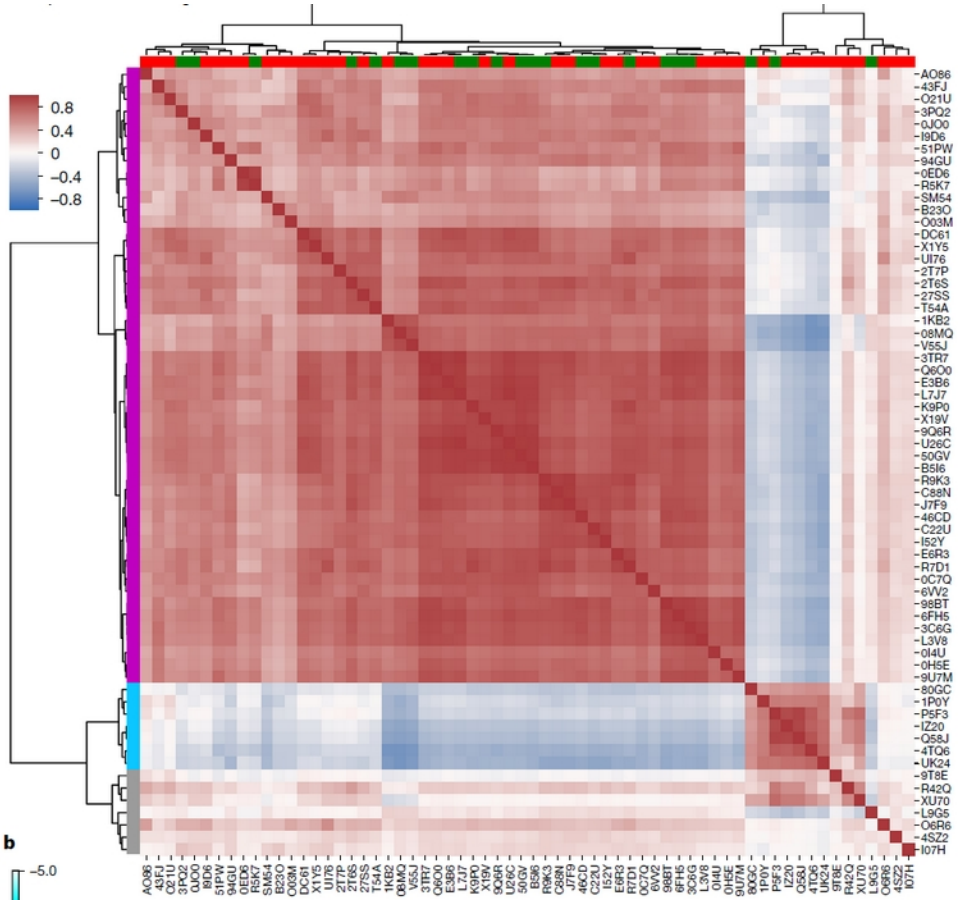
J. Carp, f. Neuroscience, 2012

- A **single** event-related fMRI experiment to a large number of unique analysis procedures
- Ten analysis steps for which multiple strategies appear in the literature : **6,912 pipelines**
- Plotting the maximum peak

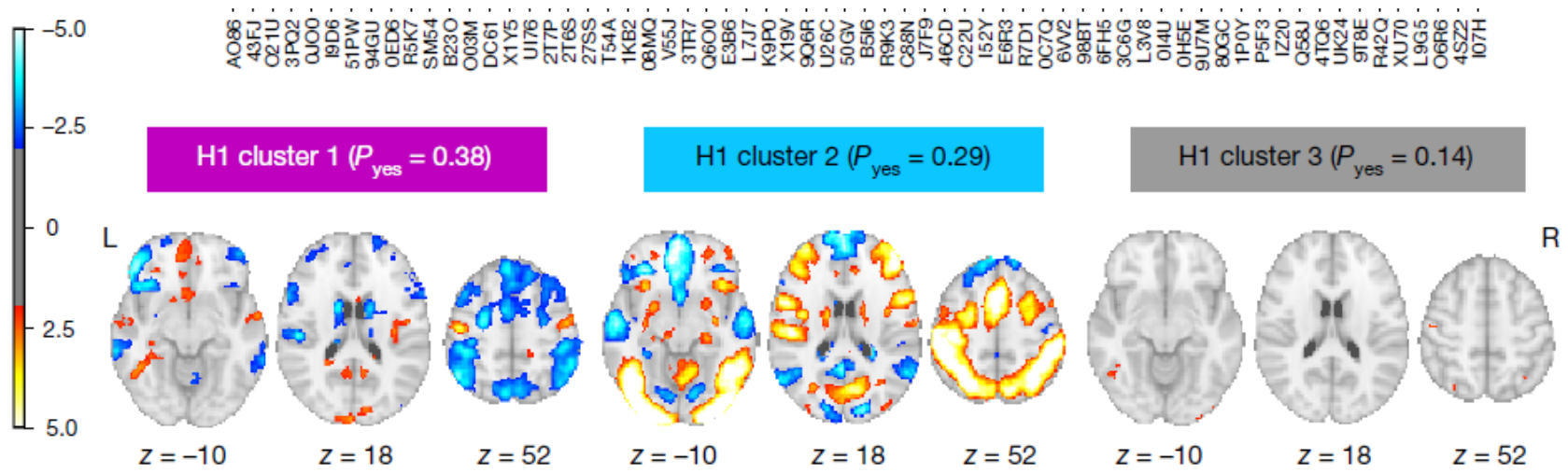


Analytical flexibility: Botvinik-Nezer, 2020

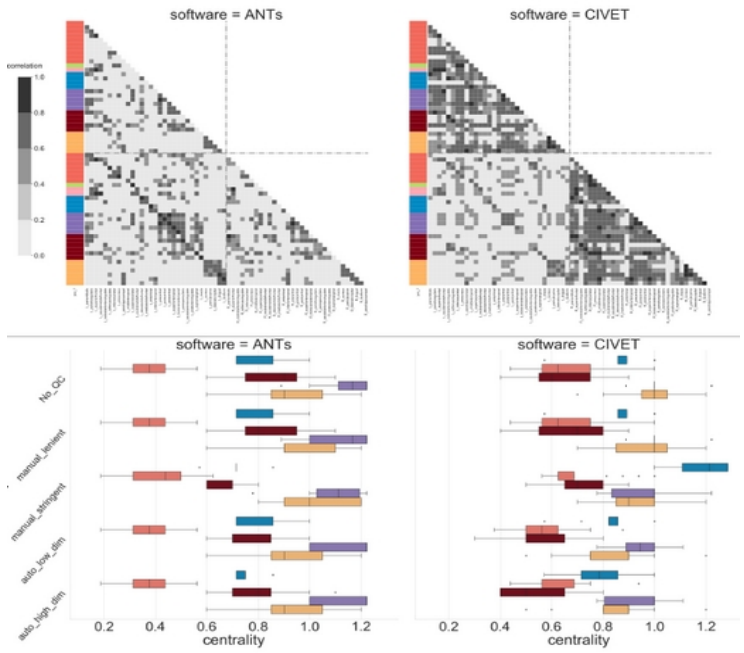
- 70 independent teams analyzing *the same* fMRI dataset
- No team had the same pipeline
- Results show three "clusters"
- Even within clusters decision to reject H0 varies



Analytical flexibility



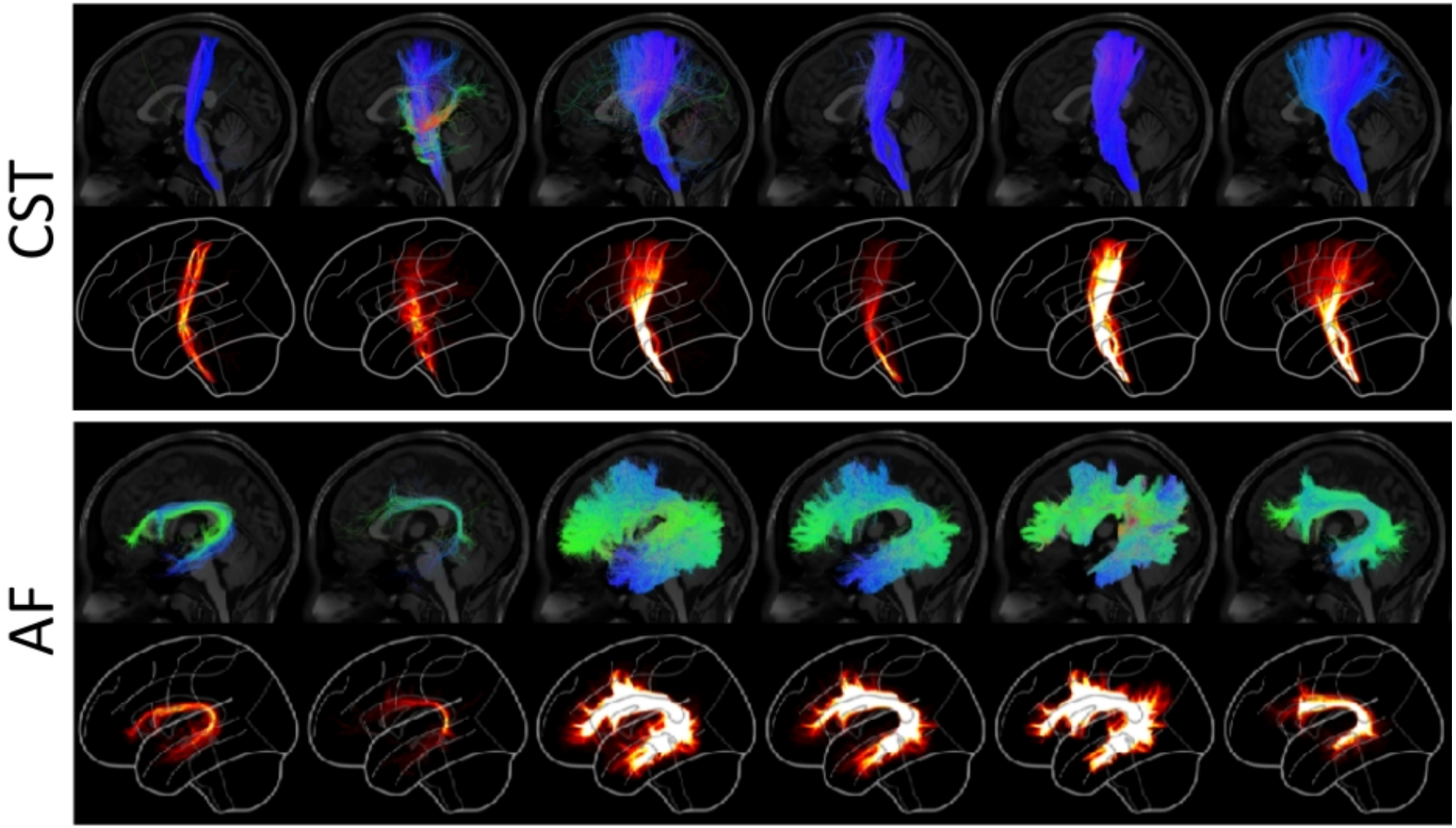
Analytical flexibility: N. Bhagwat, 2020



Diffusion pipelines

What happens when 42 groups dissect 14 white matter bundles on the same dataset?

KG Schilling … M Descoteaux 2020 (~150 authors)



Cause: data issues

- A less rare case than usually thought !
- No license
- Database not containing what it describes
- Wrong QC – QC unreliable
- Headers of files are not correct (cf the Left/Right issue)
- Provenance of data is lost
- SAM1 SAM2 SAM3:
  https://www.youtube.com/watch?v=N2zK3sAtr-4

- From HCP:

"With the releases of FreeSurfer 7.X, there have been some regressions in surface placement performance when running FreeSurfer inside the HCP Pipelines. At this time, *I would recommend sticking with FreeSurfer 6.0* while we get these issues sorted out."

Three causes

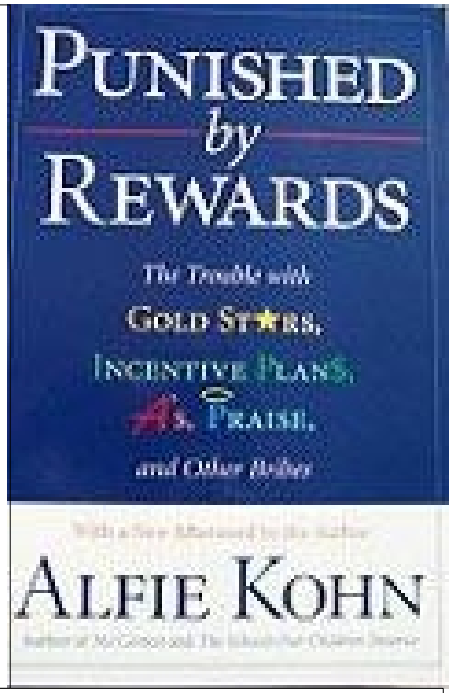
1. Poor statistical procedures

2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

Misplaced incentives

- Publication = the only "currency" for researchers, universities
- The high competition incites researchers to keep data and code as "assets" and to get as many authorships as possible
- The current incentive system promotes poorly reproducible research



ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

The natural selection of bad science

Paul E. Smaldino[1] and Richard McElreath[2]

[1]Cognitive and Information Sciences, University of California, Merced, CA 95343, USA

[2]Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

BMJ bioethical issues in biostatistical consulting

Objectives US national pilot study to

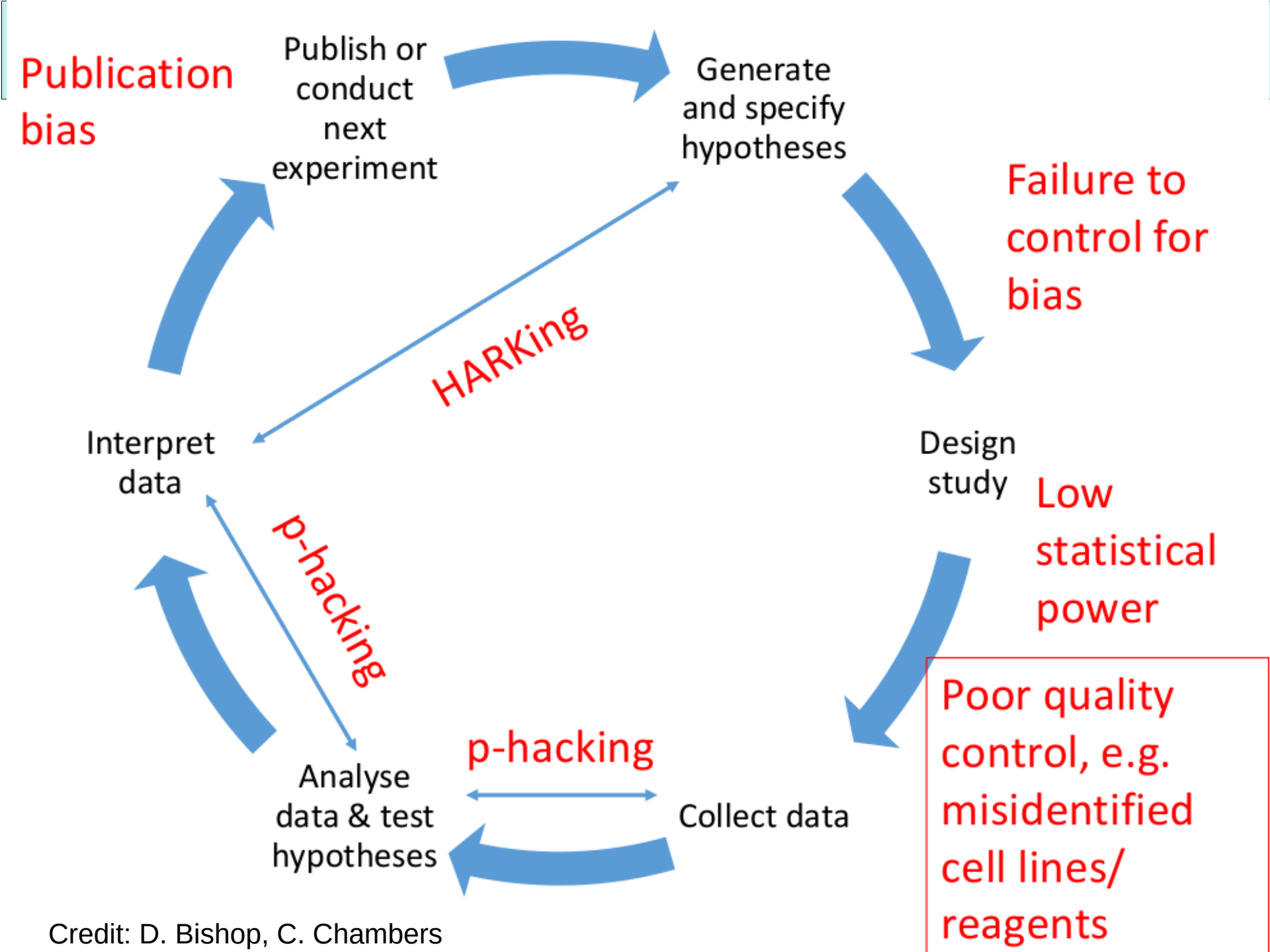
(1) test the feasibility of online administration of the Bioethical Issues in Biostatistical Consulting (BIBC) Questionnaire

(2) determine the prevalence and relative severity of a broad array of bioethical violations requests that are presented to biostatisticians by investigators seeking biostatistical consultations; and

(3) establish the sample size needed for a full-size phase II study.

Conclusion**: clear evidence** that researchers make requests of their biostatistical consultants that are rated as **severe violations** and occur **frequently**

Wang et al. 2017. BMJ Open 7 (11): 2017.



Credit: D. Bishop, C. Chambers

Part I: Reproducibility: background

Part II : Etiology of Irreproducibility

Part III : Some therapeutic proposals

What can we do ?

- Improve training
- Develop better tools and standards – make these tools that could change the culture
- Change the incentives

P-hacking - solutions ?

- **Pre-registration**
- Ban p-values
- Change \alpha
- Complement with other statistics

Significance

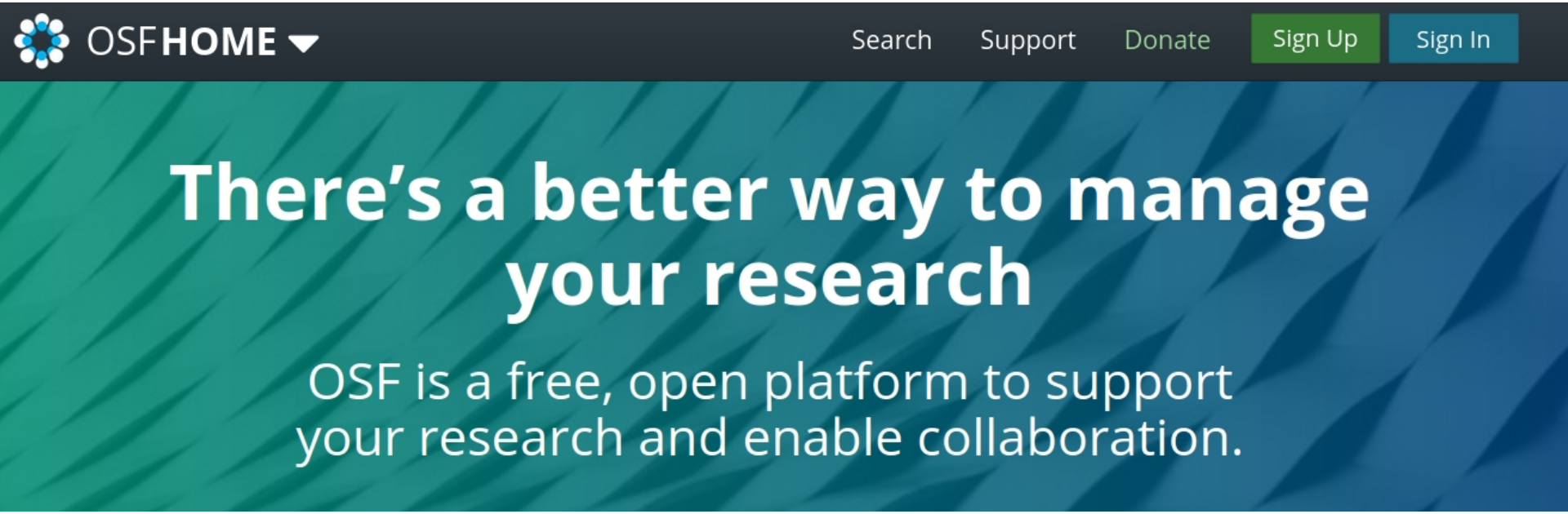
The lack of reproducibility of scientific research undermines public confidence in science and leads to the misuse of resources when researchers attempt to replicate and extend fallacious research findings. Using recent developments in Bayesian hypothesis testing, a root cause of nonreproducibility is traced to the conduct of significance tests at inappropriately high levels of significance. Modifications of common standards of evidence are proposed to reduce the rate of nonreproducibility of scientific research by a factor of 5 or greater.

Johnson, V.E. (2013). Revised standards for statistical evidence. PNAS 110, 19313–19317.

Pre-registration

- Promoted by many, in particular Chris Chambers
- Simple: explain in details your methods and hypotheses, what you will test
- This is reviewed, and if accepted results will be published
- This does not preclude exploratory analyses
- Do it for yourselve - eg use OSF (time stamped)
- There is almost no reason to not do it

Open Solution: Preregistration with OSF



COMMENT · 10 SEPTEMBER 2019

What's next for Registered Reports?

Reviewing and accepting study plans before results are known can counter perverse incentives. Chris Chambers sets out three ways to improve the approach.

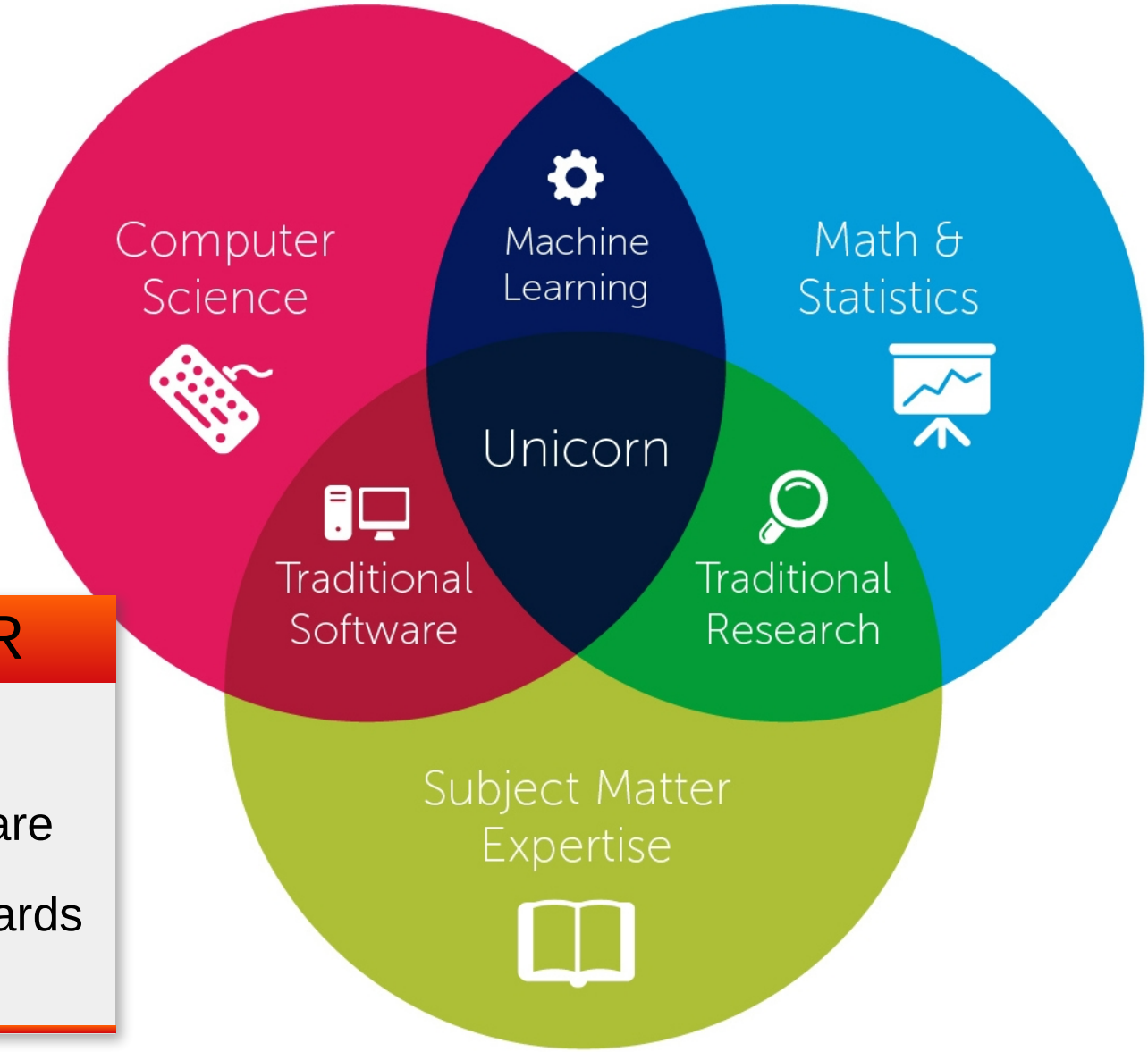
Chris Chambers

Open Software and Pipelines

- The rise of Git (GitHub, GitLab, ...) repositories
  - Many are open, collaborative
  - From lab-based “one software” to collaborative open source
- The rise of “FAIR” containers
  - BIDS
- The (slow) rise of standardized provenance information
  - NIDM - W3C prov

Ethical Scholarly communications
Epistemiology / lessons from the past
How to collaborate and teach

Data Science



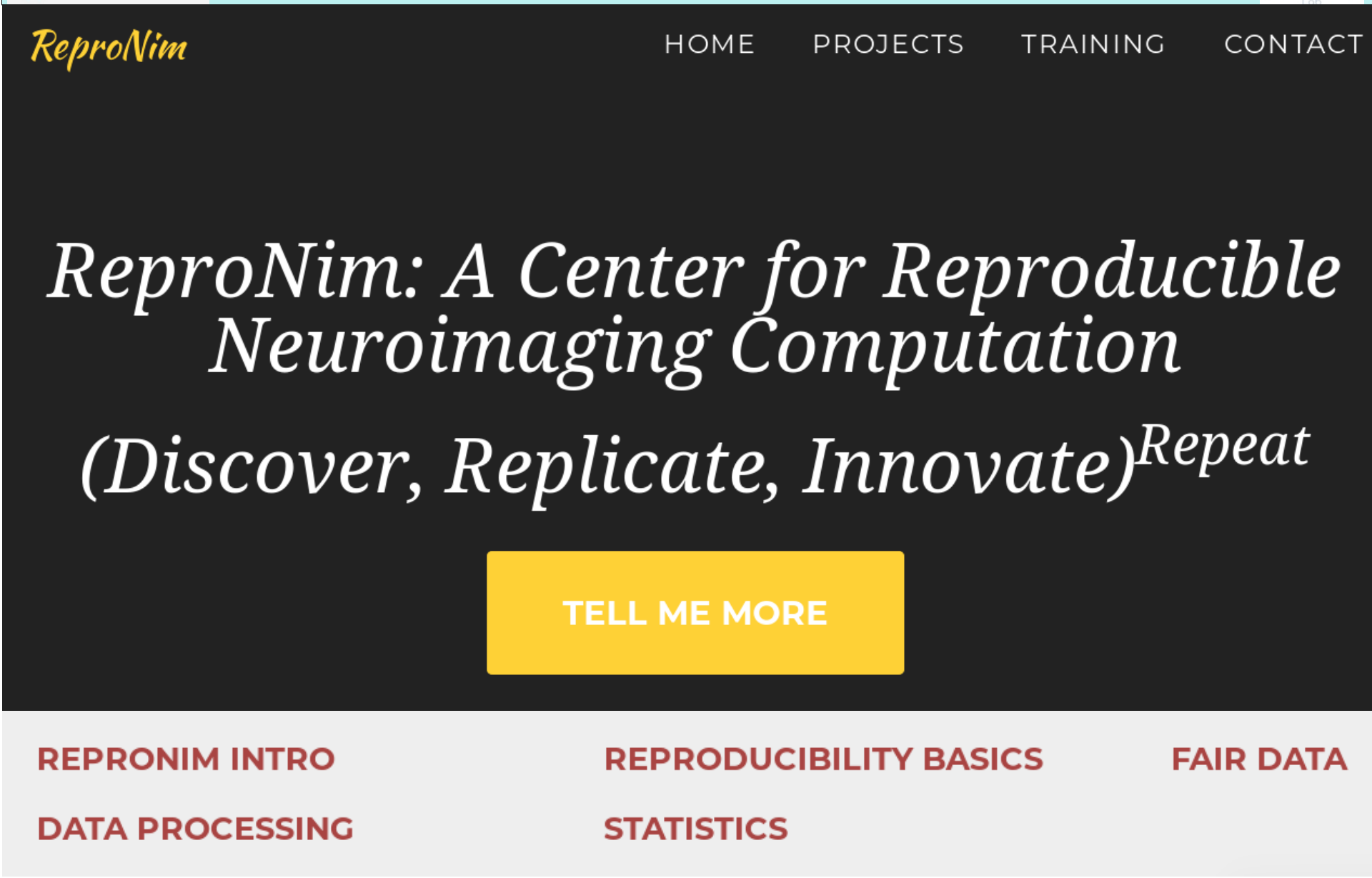
FAIR

- Data
- Software
- Standards reuse

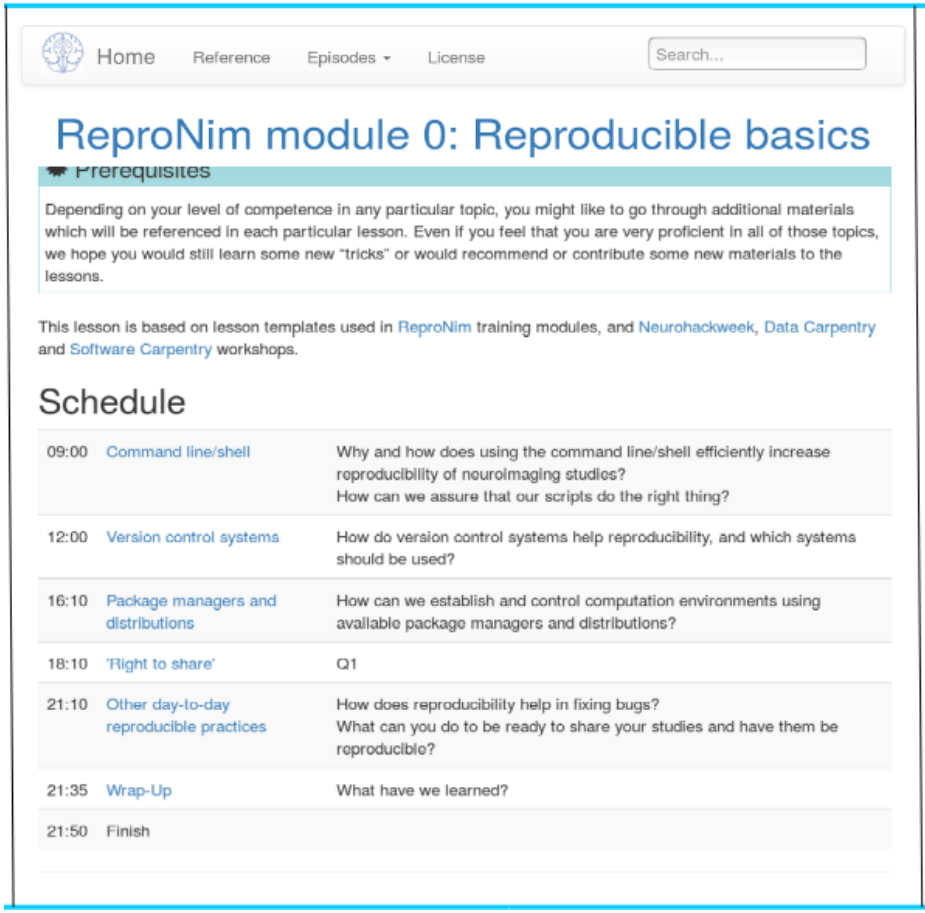
Science and Technology

Research

Training: NIH P41 ReproNim



Training



Home Reference Episodes License Search...

ReproNim module 0: Reproducible basics

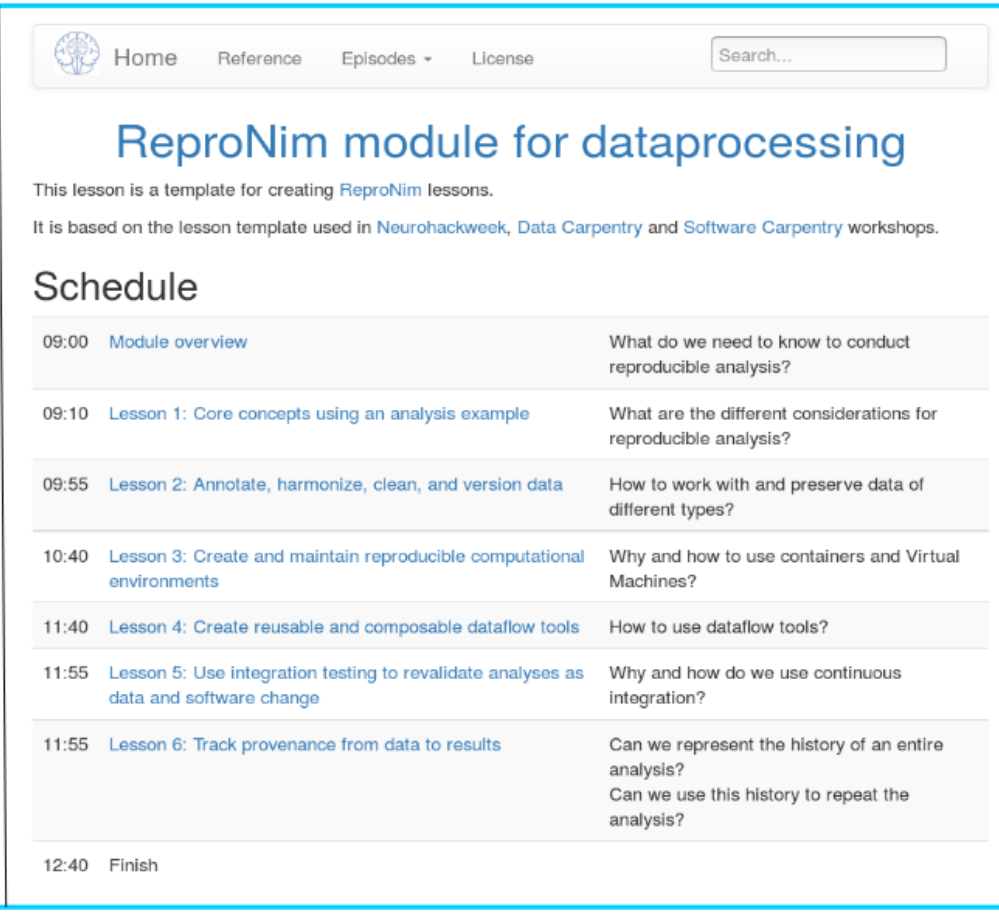
Prerequisites

Depending on your level of competence in any particular topic, you might like to go through additional materials which will be referenced in each particular lesson. Even if you feel that you are very proficient in all of those topics, we hope you would still learn some new "tricks" or would recommend or contribute some new materials to the lessons.

This lesson is based on lesson templates used in ReproNim training modules, and Neurohackweek, Data Carpentry and Software Carpentry workshops.

Schedule

09:00	Command line/shell	Why and how does using the command line/shell efficiently increase reproducibility of neuroimaging studies? How can we assure that our scripts do the right thing?
12:00	Version control systems	How do version control systems help reproducibility, and which systems should be used?
16:10	Package managers and distributions	How can we establish and control computation environments using available package managers and distributions?
18:10	'Right to share'	Q1
21:10	Other day-to-day reproducible practices	How does reproducibility help in fixing bugs? What can you do to be ready to share your studies and have them be reproducible?
21:35	Wrap-Up	What have we learned?
21:50	Finish	



Home Reference Episodes License Search...

ReproNim module for dataprocessing

This lesson is a template for creating ReproNim lessons.

It is based on the lesson template used in Neurohackweek, Data Carpentry and Software Carpentry workshops.

Schedule

09:00	Module overview	What do we need to know to conduct reproducible analysis?
09:10	Lesson 1: Core concepts using an analysis example	What are the different considerations for reproducible analysis?
09:55	Lesson 2: Annotate, harmonize, clean, and version data	How to work with and preserve data of different types?
10:40	Lesson 3: Create and maintain reproducible computational environments	Why and how to use containers and Virtual Machines?
11:40	Lesson 4: Create reusable and composable dataflow tools	How to use dataflow tools?
11:55	Lesson 5: Use integration testing to revalidate analyses as data and software change	Why and how do we use continuous integration?
11:55	Lesson 6: Track provenance from data to results	Can we represent the history of an entire analysis? Can we use this history to repeat the analysis?
12:40	Finish	

Changing the publication model

Reproducibility: A tragedy of errors, Allison et al, 2016, Nature

SCIENTIFIC DATA

OPEN

Comment: High-quality science requires high-quality open data infrastructure

Susanna-Assunta Sansone[1], Patricia Cruse[2] & Mark Thorley[3]

Received: 26 January 2018
Accepted: 29 January 2018
Published: 27 February 2018

Resources for data management, discovery and (re)use are numerous and diverse, and more specifically we need data resources that enable the FAIR principles[1] of Findability, Accessibility, Interoperability and Reusability of data.

Example

My paper concludes:

- Increase in resting state connectivity between Right Superior Temporal Gyrus and the Right Superior Frontal Gyrus in subjects with autism, and this connectivity correlated with diagnostic severity.

Credit : D. Kennedy

What statistic? (covariates, corrections)

What data? (MR parameters)
What analysis? (software and parameters)

My paper concludes:

- Increase in resting state connectivity between Right Superior Temporal Gyrus and the Right Superior Frontal Gyrus in subjects with autism, and this connectivity correlated with diagnostic severity.

What subject characteristics? (age, gender, SES, genetics, environment, etc.)

What measure?

What anatomic framework? (atlas)

Credit : D. Kennedy

My paper actually concludes:

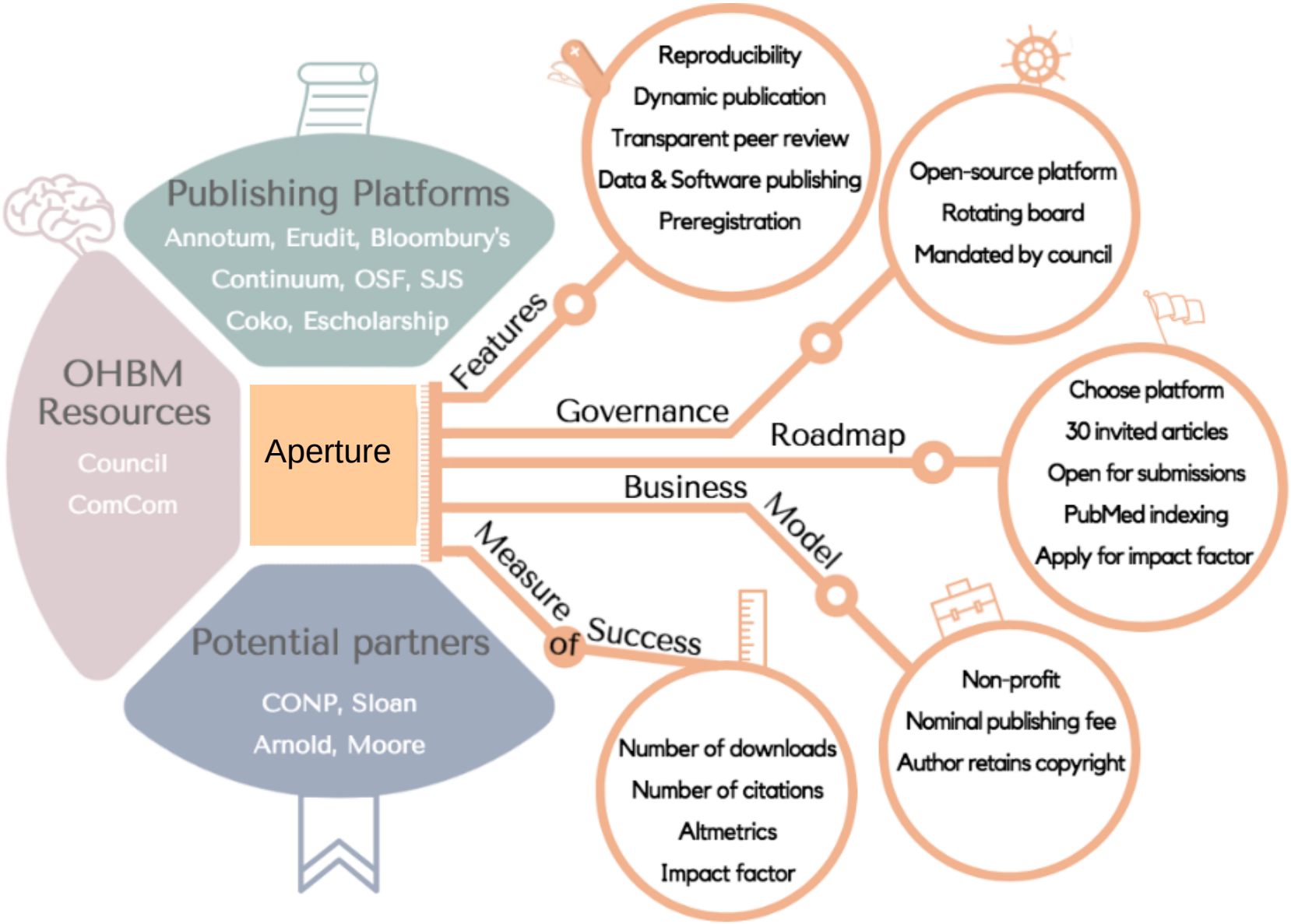
- Using a paired T-test, covarying for age, gender and handedness using cluster-size FWE correction, we saw an increase (P<.01) in regional seed-based using CONN resting state connectivity between AAL regions of the Right Superior Temporal Gyrus and the Right Superior Frontal Gyrus in 40 subjects with autism (age 14+-5, 19M/11F, IQ 90+- 10, ADOS 20+-5), and this connectivity correlated (Pearson, P<.05) with diagnostic severity as measured by the social subscale of the ADI.

Credit : D. Kennedy

Publishing: what do we need

- Publish reusable research objects
  - Data first !
  - Software, workflow, analyses
  - Jupyter notebooks, hybrid objects
  - Pre-registered report
- Vetting objects
  - By experts
  - By community based (alt)metrics
- Make published research object machine readable as much as possible

OHBM Open Publishing Initiative



Thank you

- Lab@McGill**: https://neurodatascience.github.io/**
- **McGill** colleagues: S. Brown, T. Glatard, G. Kiar, A. Evans, C. Greenwood, A. DeGuise and others
- **ReproNim** colleagues: D. Kennedy, D. Keator, S. Ghosh, M. Martone, J. Grethe, M. Hanke, Y. Halchenko
- **Berkeley** colleagues: S. Van der Walt, M. Brett, J. Millman, Dan Lurie, M. D'Esposito, et al
- **Pasteur** colleagues: G. Dumas, R. Toro, T. Bourgeron, and others
- **Paris** colleagues: B. Thirion, G. Varoquaux, V. Frouin, et al
- **Funders**: HBHL, HBHL NeuroHub, NIH, NIMH, NSERC, Compute Canada