Reproducibility issues in neuroscience and neuroimaging

Jean-Baptiste Poline

MNI, Ludmer Center, BIC, McGill
HWNI, UC Berkeley

Part I: Reproducibility: background

Part II : Etiology of Irreproducibility

Part III : Some therapeutic proposals

Part I: Reproducibility: more than just fMRI

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

Forensic Analysis

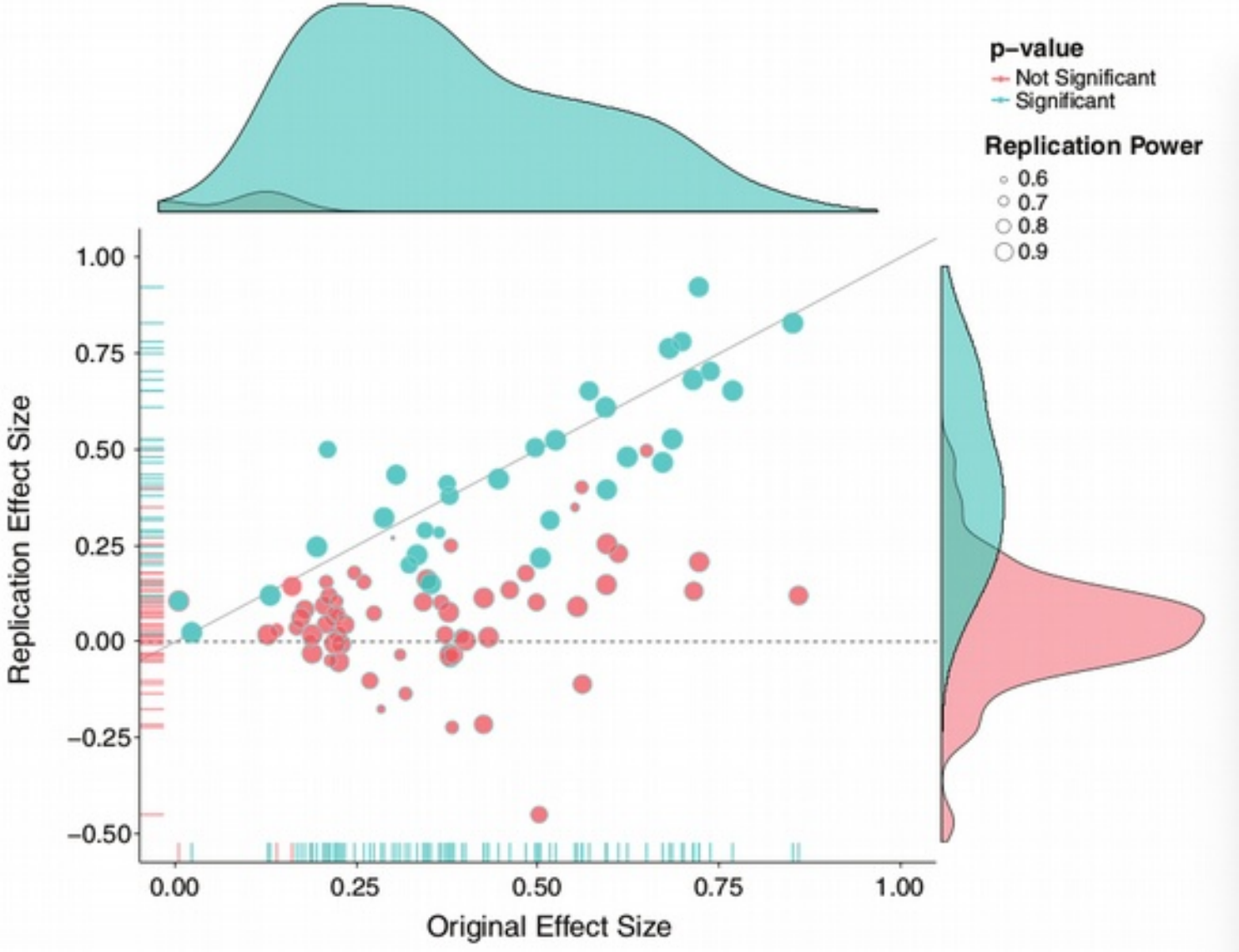
- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, "Forensic analysis", Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug, use expression profiles to build a model that predicts patient response

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."

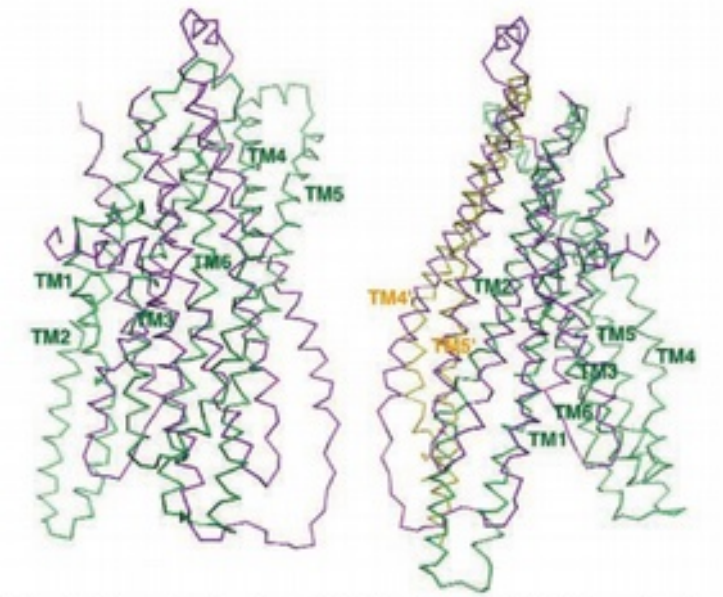
B. Nosek, Estimating the reproducibility of psychological science, Science 2015



* 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

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

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

Genetics and Imaging genetics

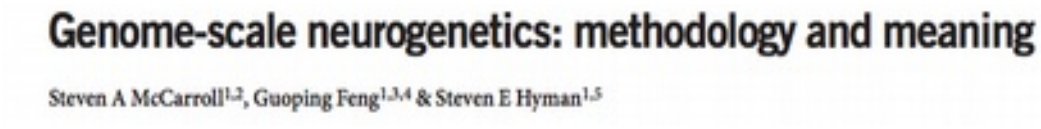

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing studies in human genetics: design and interpretation

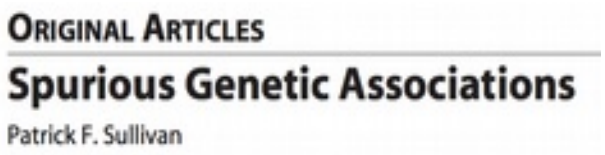
David B. Goldstein[1], Andrew Allen[1,2], Jonathan Keebler[2], Elliott H. Margulies[3], Steven Petrou[4,5], Slavé Petrovski[1,6] and Shamil Sunyaev[7]

"Human genomes have a high level of '**narrative potential**' to provide compelling but statistically poorly justified connections between mutations and phenotypes."

"A critical challenge for biologists […] will be avoiding premature hypotheses born of biological plausibility and '**Just So**' stories."


Genome-scale neurogenetics: methodology and meaning

Steven A McCarroll[1,2], Guoping Feng[1,3,4] & Steven E Hyman[1,5]


ORIGINAL ARTICLES

Spurious Genetic Associations

Patrick F. Sullivan

"Findings from single association studies constitute '**tentative knowledge**' and must be interpreted with exceptional caution.

<u>Biological plausibility is not a substitute for statistical significance or experimental validation</u>

Who said that ...

Much of the scientific literature, perhaps half, may simply be untrue.

In their quest for *telling a compelling story*, scientists too often sculpt data to fit their preferred theory of the world. Or they retrofit hypotheses to fit their data.

Our love of “significance” pollutes the literature with many a statistical fairytale. We reject important confirmations.

Brian Nosek	David Eidelman	Richard Horton
Open Science	Dean Faculty of Medicine	Lancet Editor in Chief

Who said that ...

Much of the scientific literature, perhaps half, may simply be untrue.

Richard Horton
Lancet Editor in Chief

In their quest for telling a compelling story, scientists too often sculpt data to fit their preferred theory of the world. Or they retrofit hypotheses to fit their data.

Our love of "significance" pollutes the literature with many a statistical fairytale. We reject important confirmations.

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

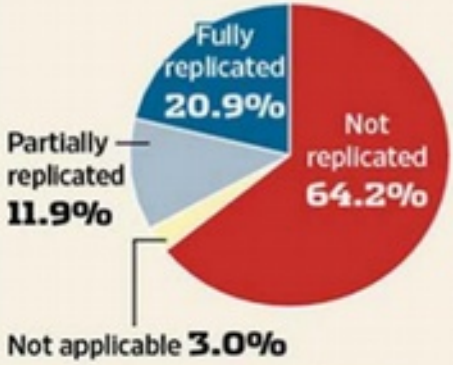
2005. PLoS Medicine, 2(8), e124. doi: 10.1371/journal.pmed.0020124

"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*

No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.



Source: Nature Reviews Drug Discovery

THE LANCET

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

All Content | Search | Advanced Search

Research: increasing value, reducing waste

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

Part I: Reproducibility: case studies

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

2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

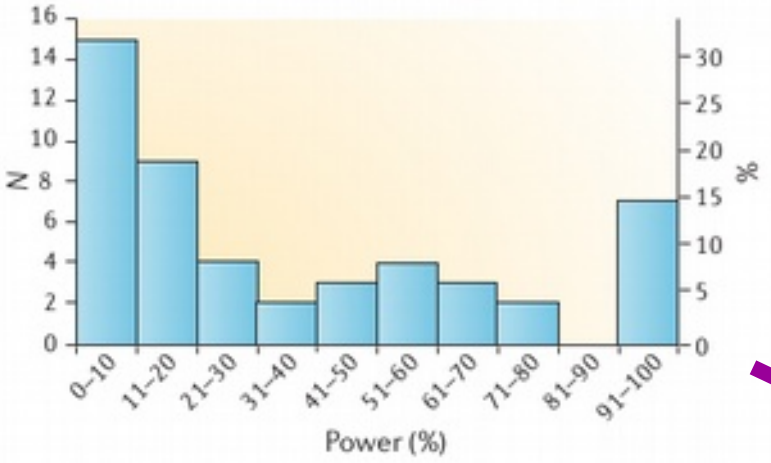
Power issues

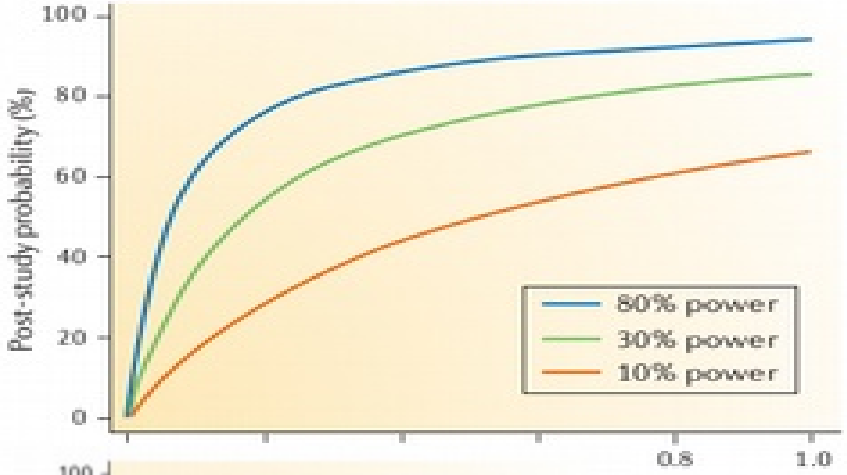
Open access, freely available online

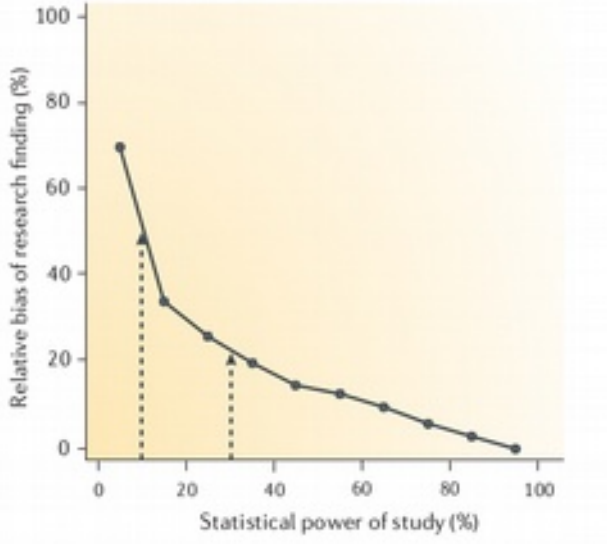
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

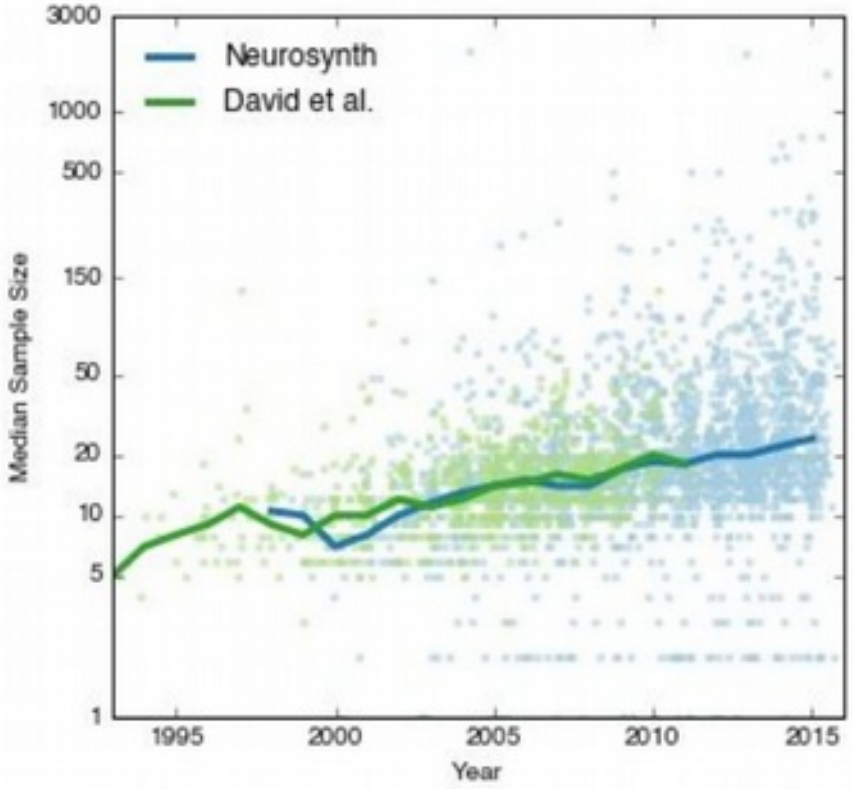


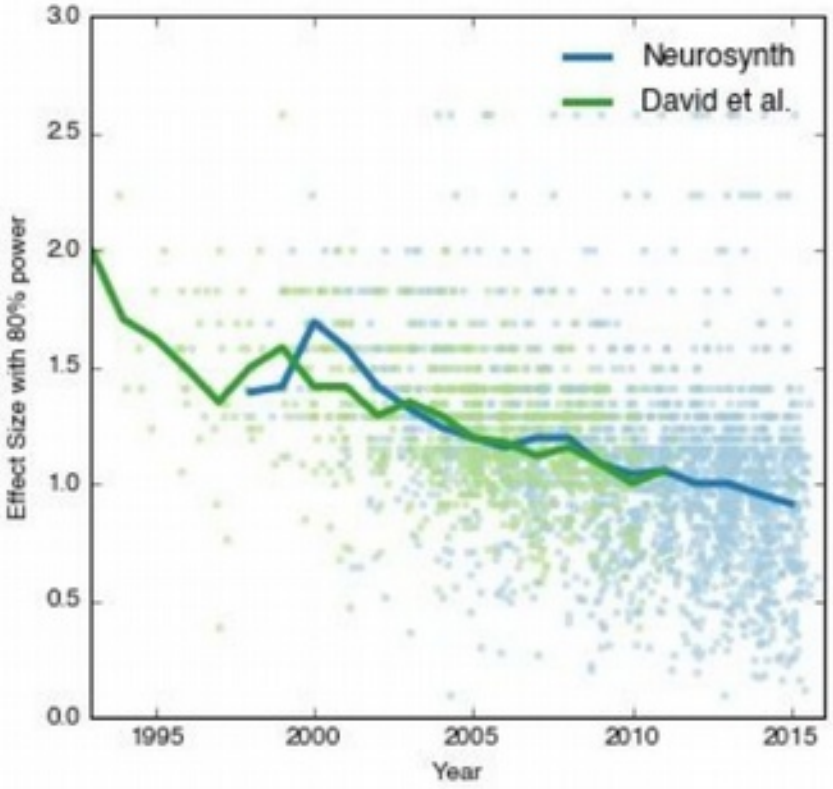




Button et al., NNR, 2013

Feeling the Future





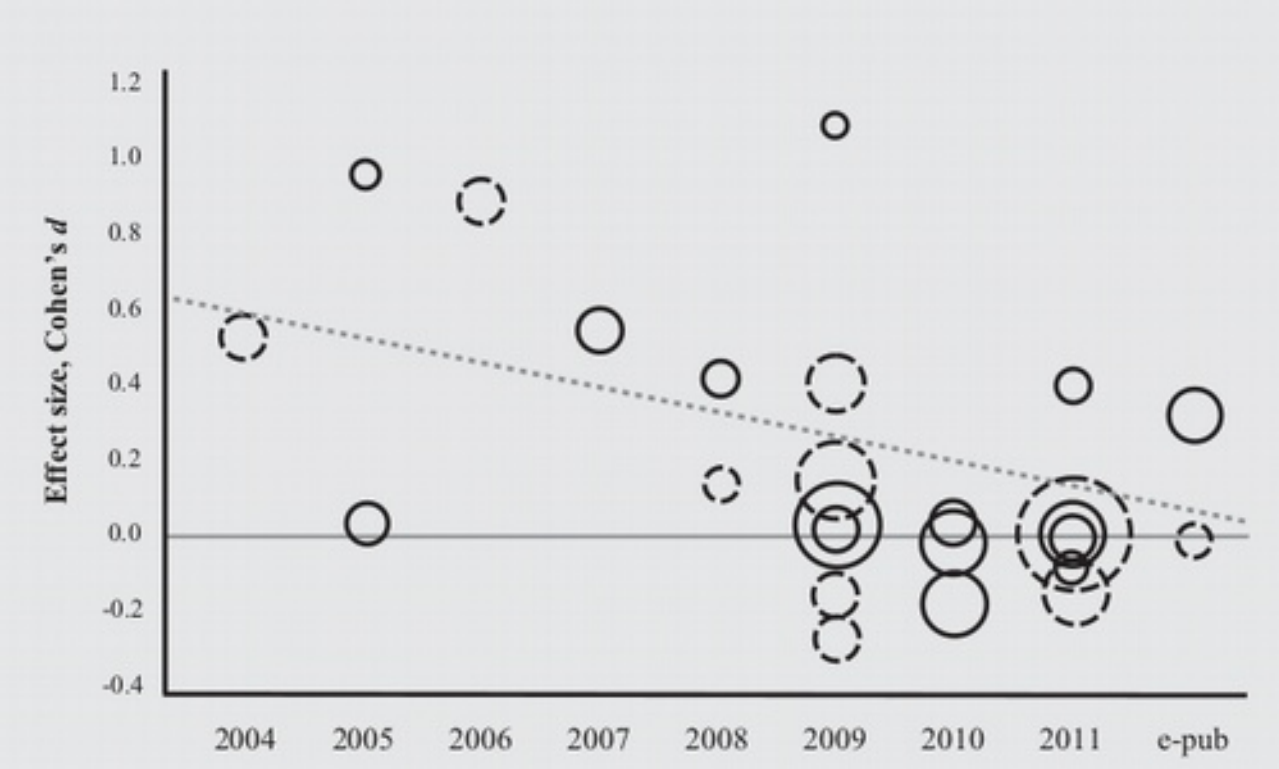
Poldrack et al., PNAS, 2016

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%

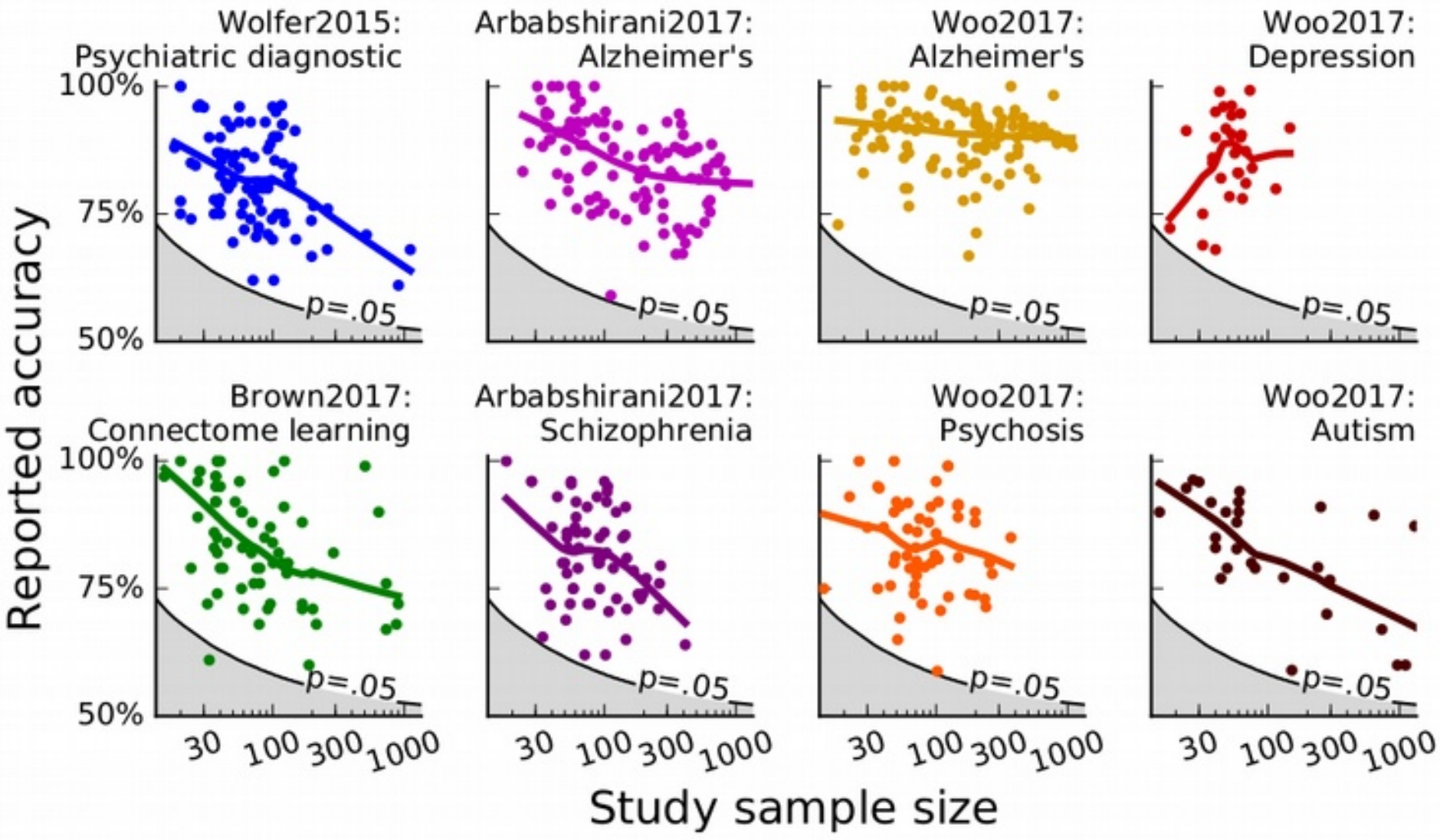
Statistics: ~~The~~ One problem



Molendijk, 2012: BDNF and hippocampal volume

See also : Mier, 2009: COMT and DLPFC

Sample size issue in ML



Ioannidis 2005

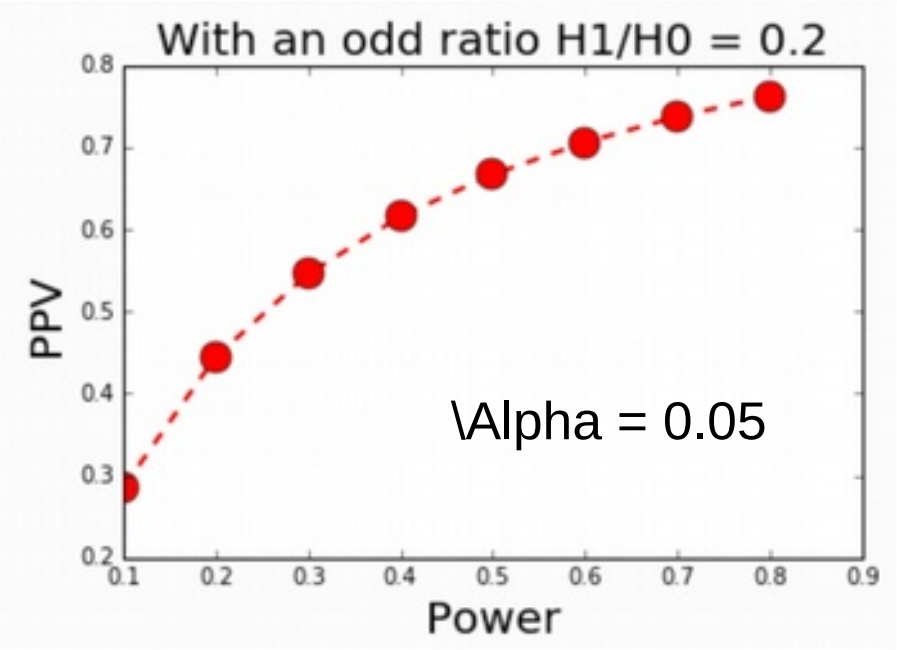
Open access, freely available online

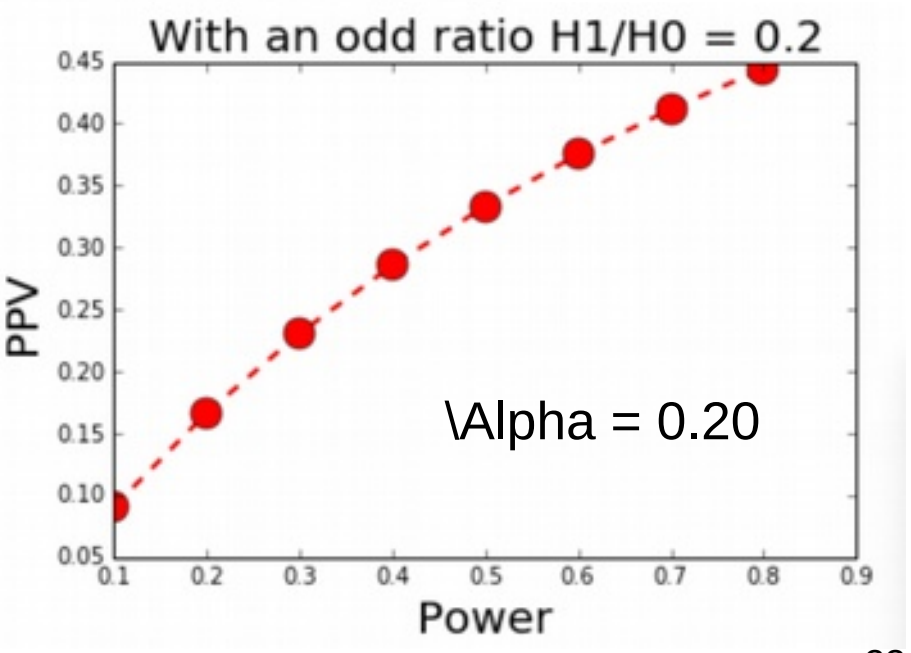
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

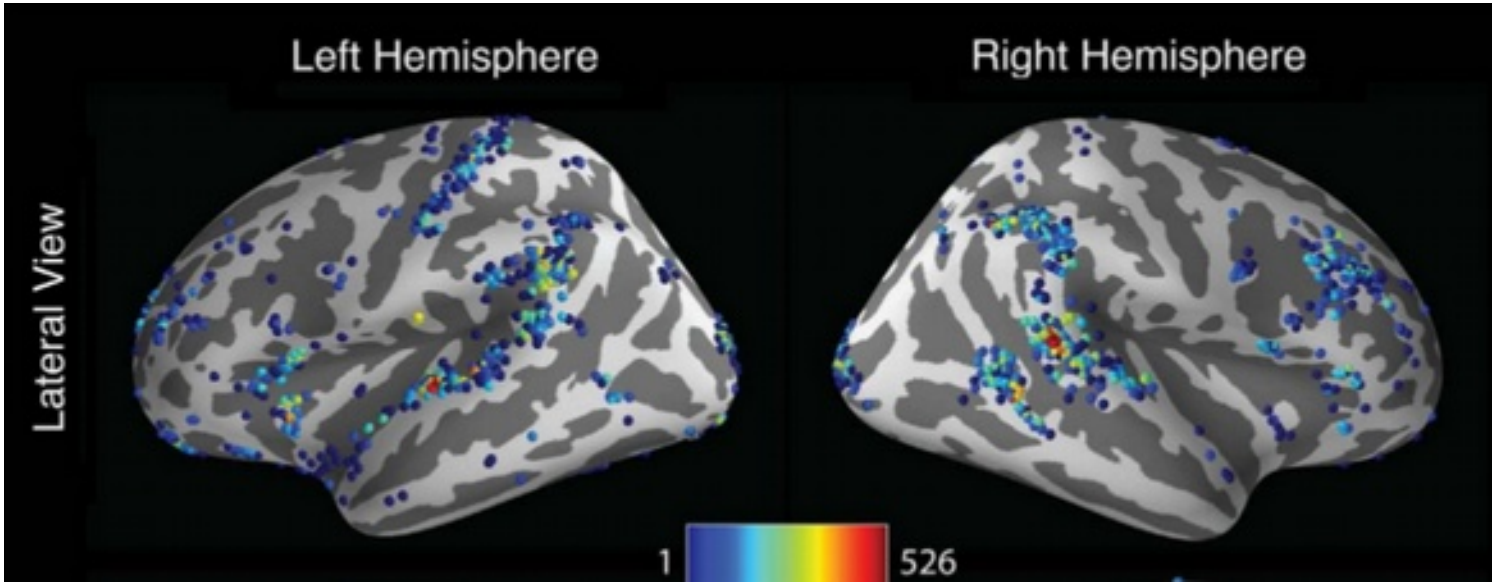




Estimating analytical flexibility of fMRI

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



Three causes

1. Poor statistical procedures

2. Issues in data and software

3. A cultural issue: Publication practices and research incentives

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

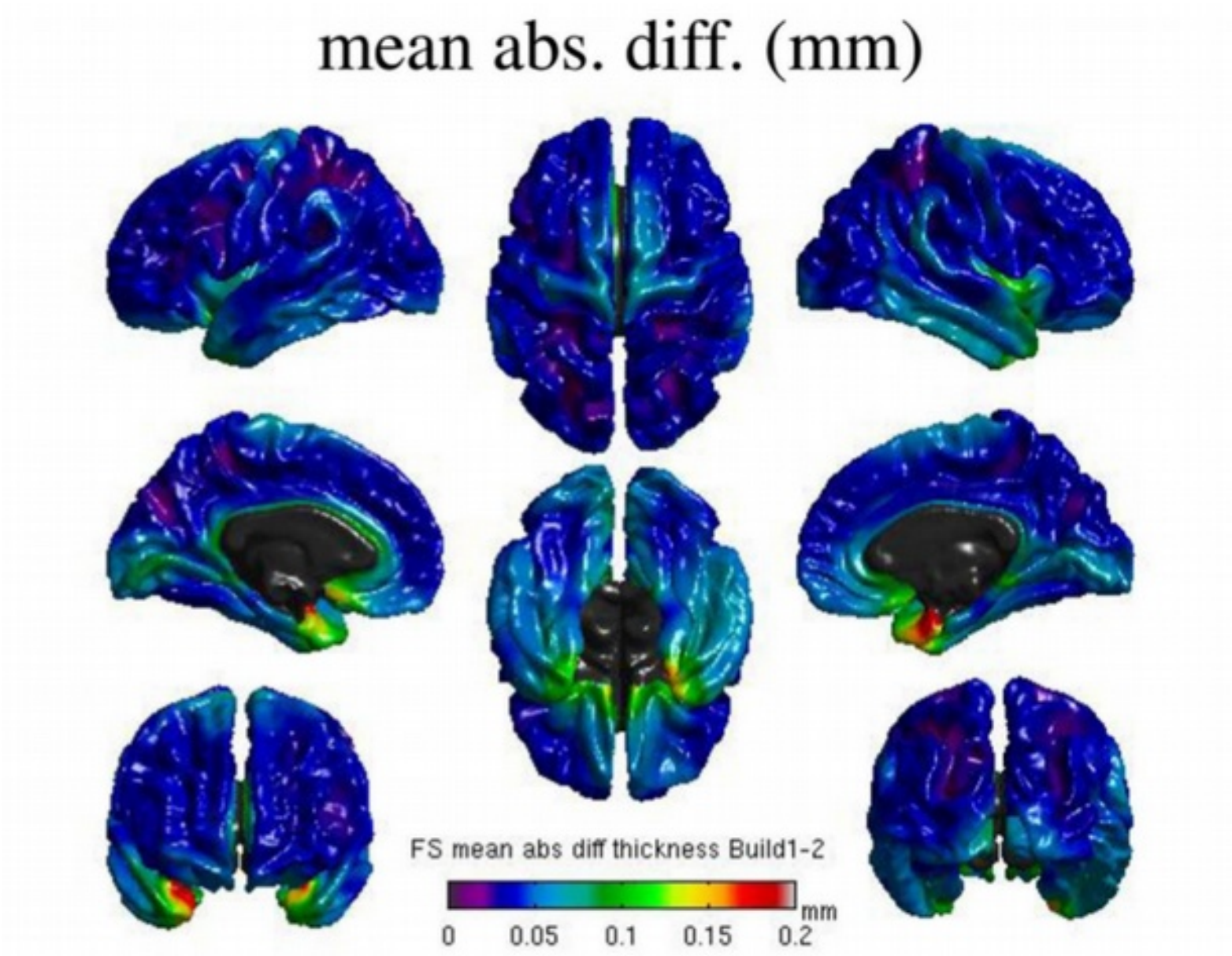
(2) determine the prevalence and relative severity of a broad array of bioethical violations requests that are presented to biostatisticians by investigators seeking biostatistical consultations; and

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

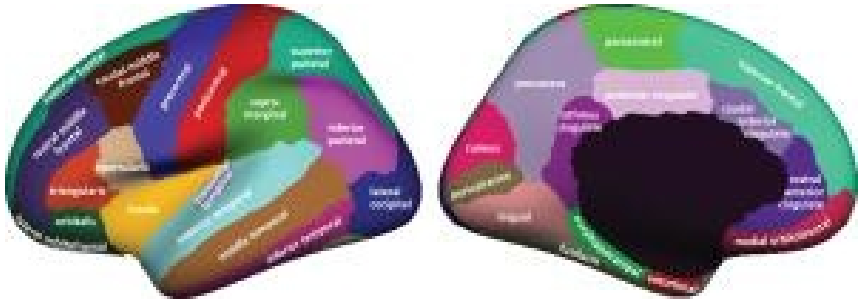
Conclusion**: clear evidence** that researchers make requests of their biostatistical consultants that are rated as **severe violations** and occur **frequently**

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

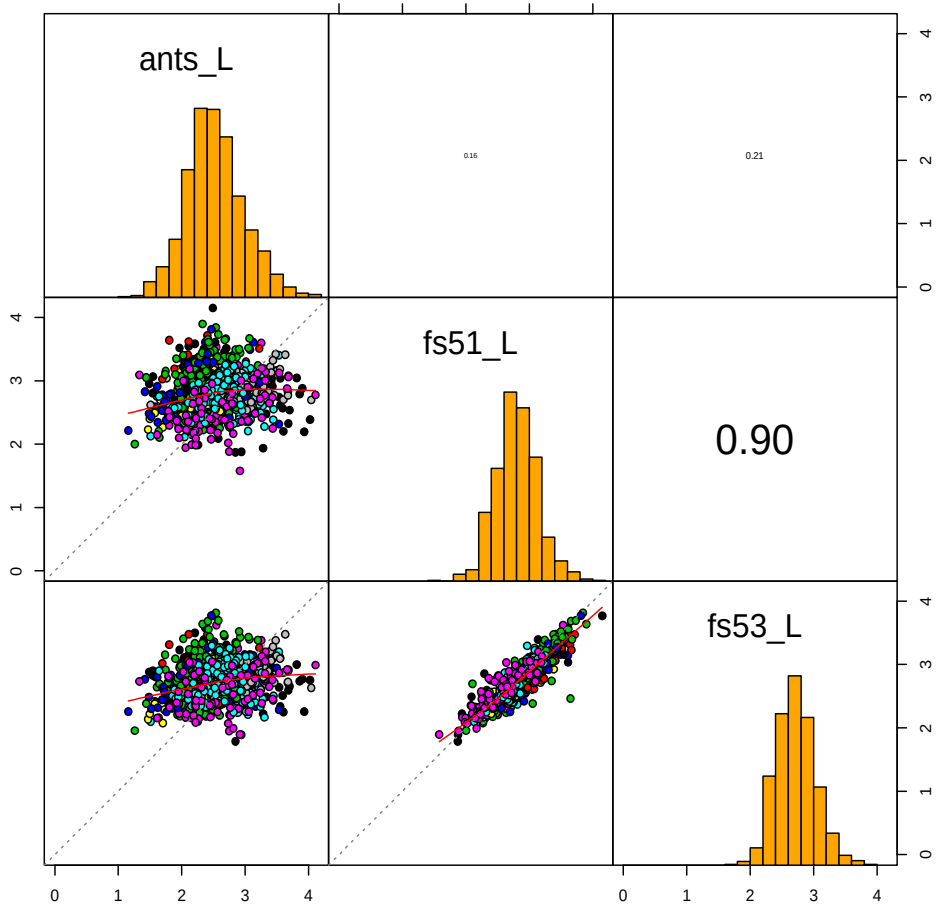
Same FS – different OS



ANTS – FS5.1- FS5.3



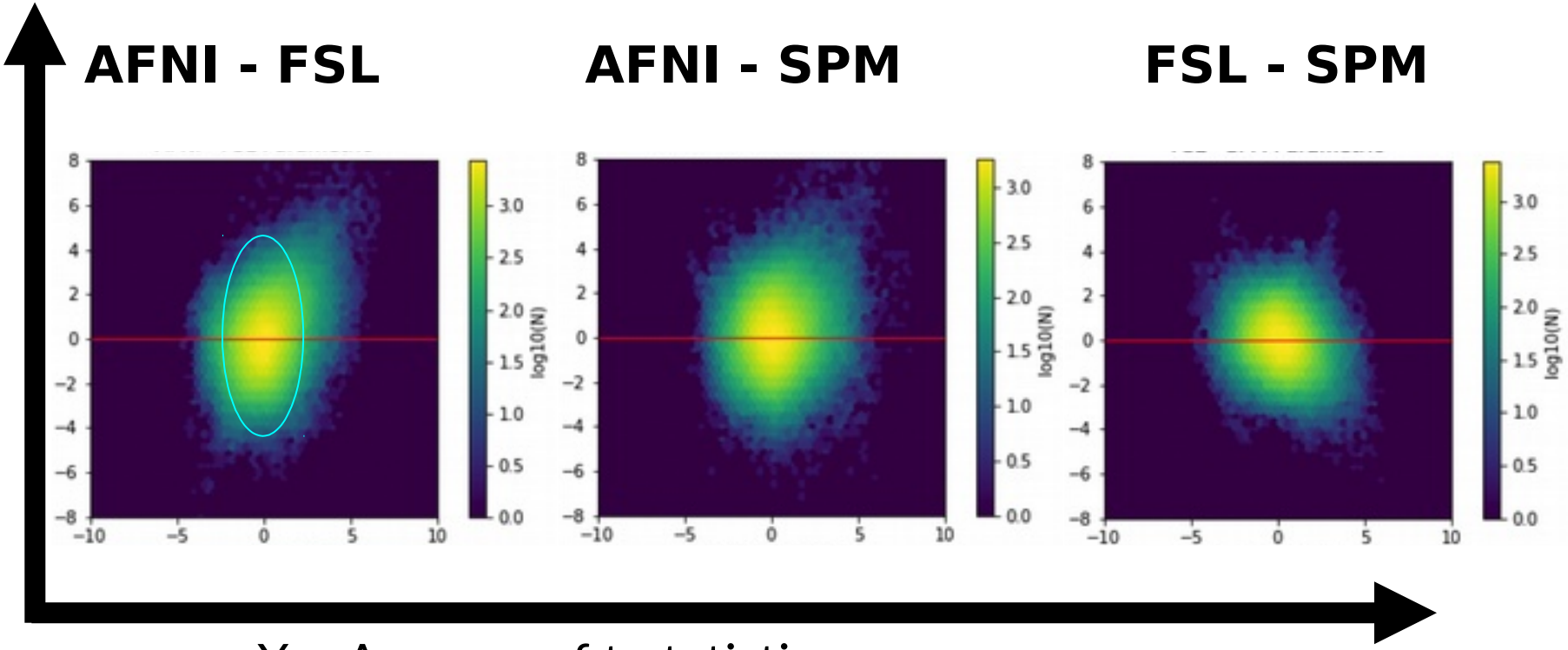
Size of the left caudal anterior Cingulate



Dickie E et al., 2017

How close are results on the same dataset?

Y = Diff. of t-statistics

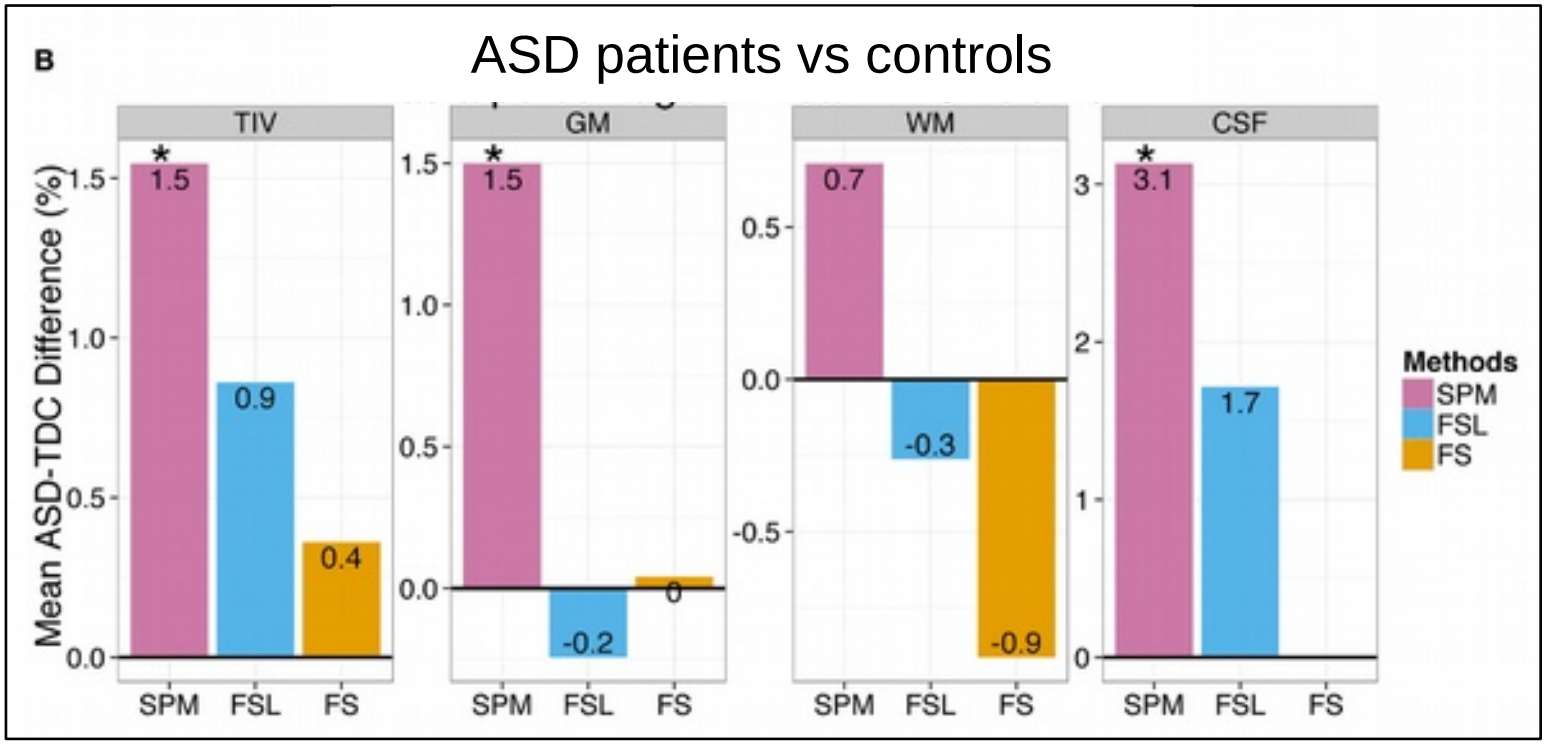


▷ Plots similar to expected variation if **independent** was fed into each package

Alex Bowring, Camille Maumet, Thomas Nichols

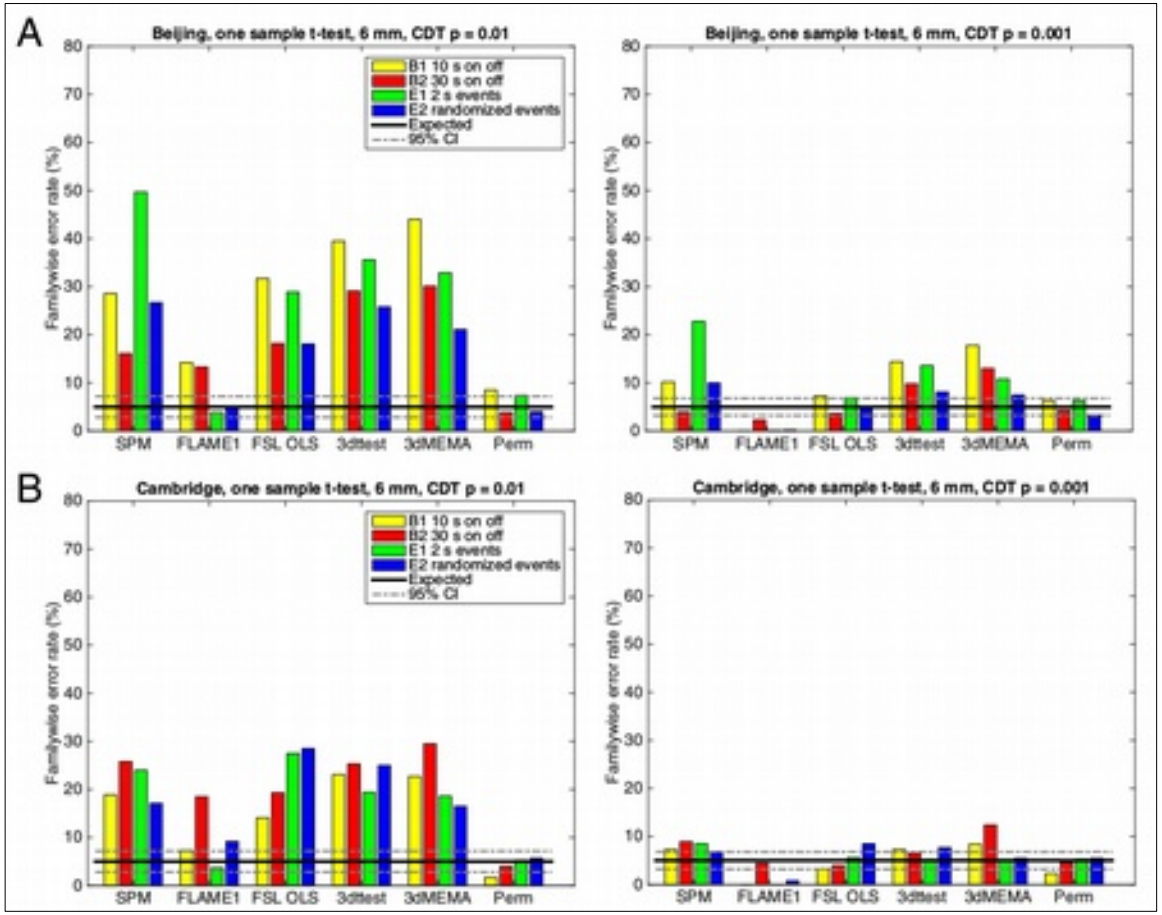
Software, version, OS

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

“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 ?

Cause: bugs in data

- 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

Three causes

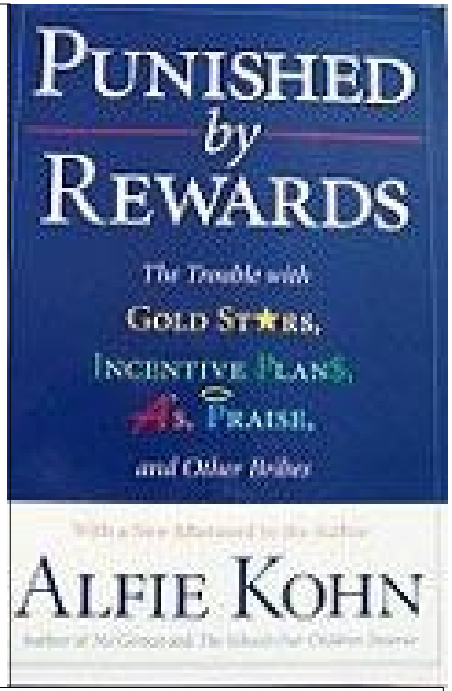
1. Poor statistical procedures

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

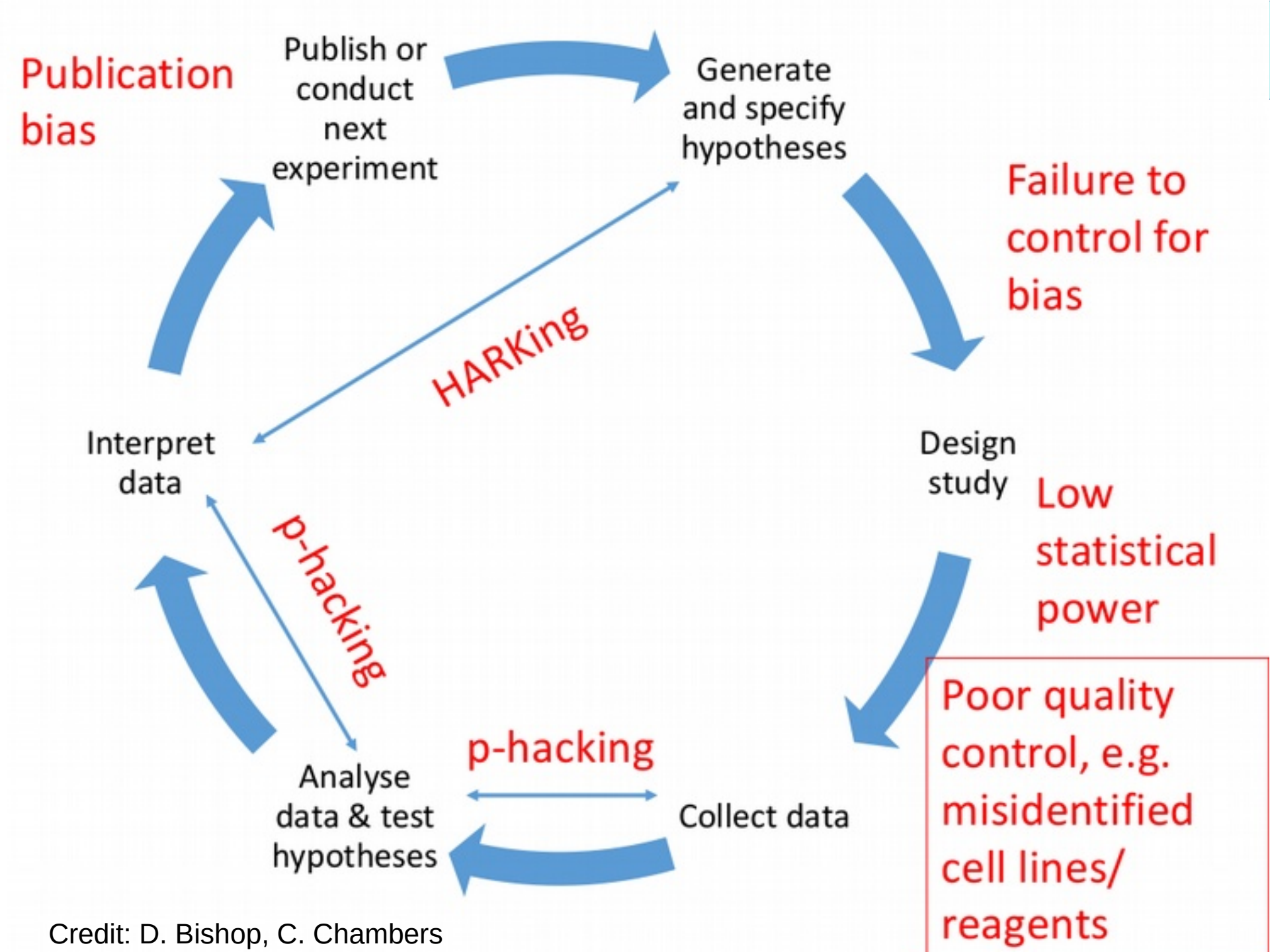
rsos.royalsocietypublishing.org

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



Credit: D. Bishop, C. Chambers

Most methods not available

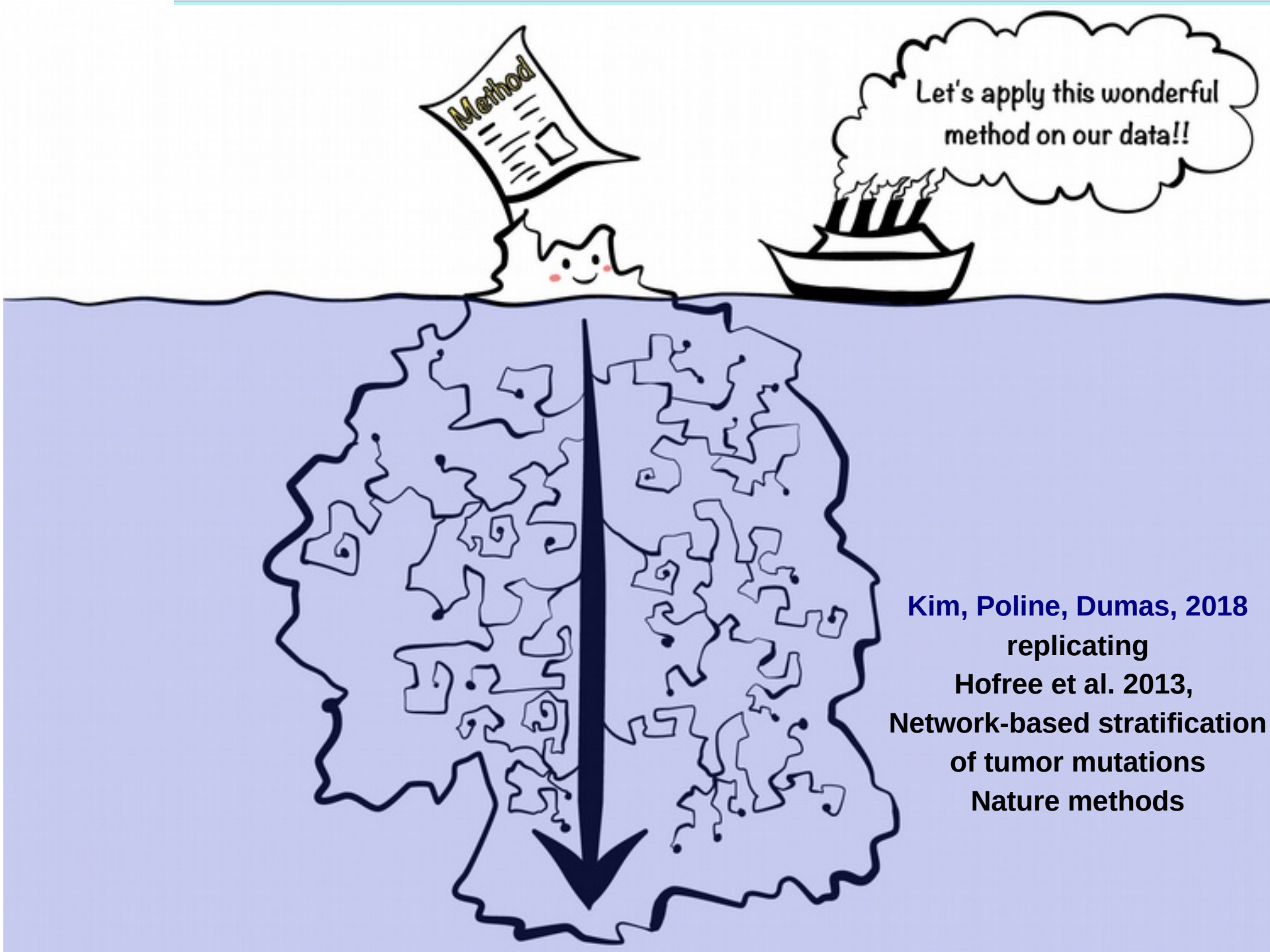
JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

Ioannidis, 2011: 9% of authors studied made data available

Generally, data and code not available at the time of publication, insufficient information in the publication for verification of results.

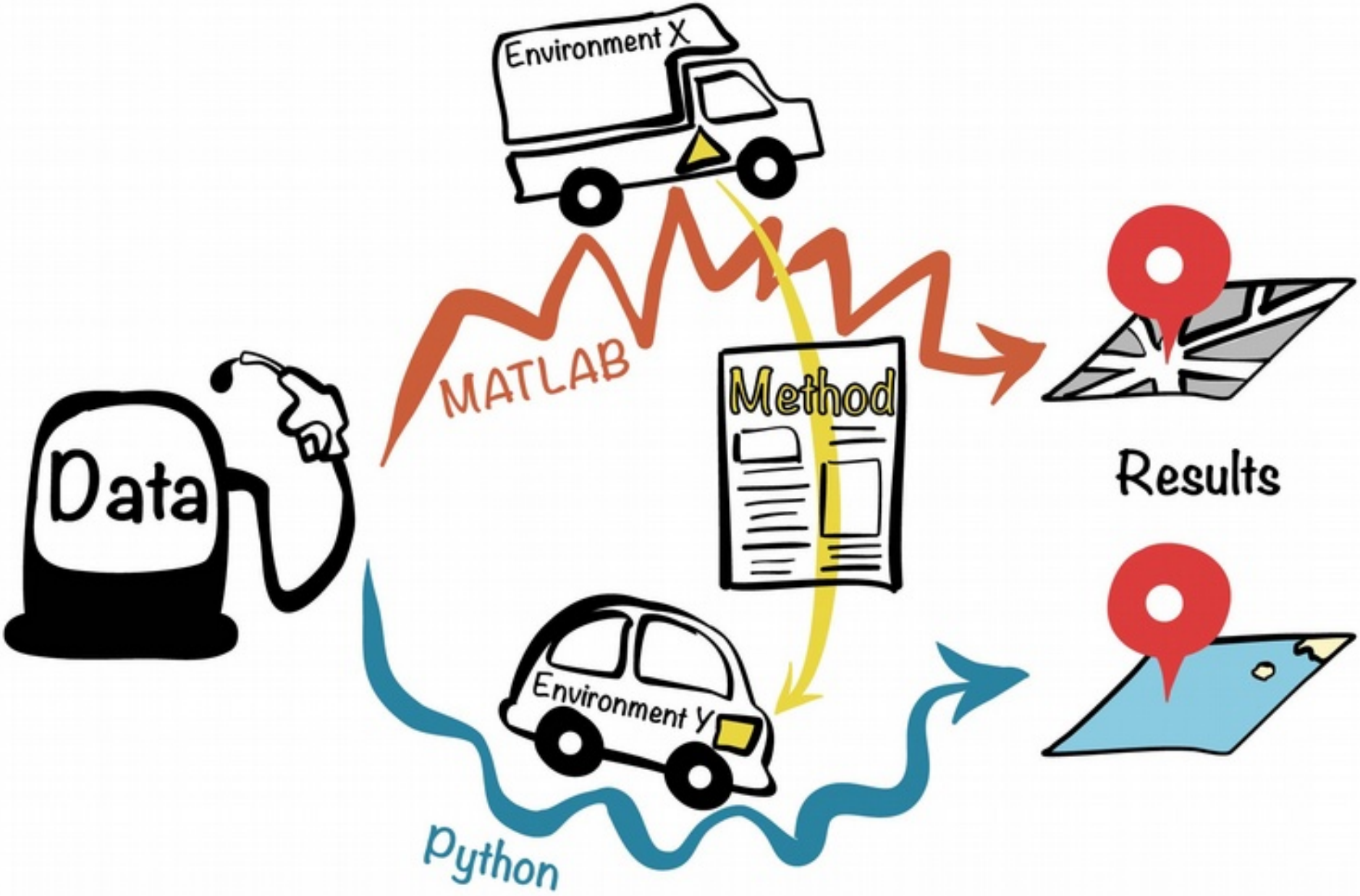
Data availability:

- “renting” data for authorship ?
- Indi – Adni – XXXX



Kim, Poline, Dumas, 2018
replicating
Hofree et al. 2013,
Network-based stratification
of tumor mutations
Nature methods

- Data and code available
- 6-8 months of a bioinformatics Masters student
- Actual help from 2 senior researchers
- Results partly reproduced
    - Matlab mex file compiled with specific library
    - Research code
        - Parameters fixed
        - Code not run
        - Code hard to read
    - Python package re-written from scratch


Environment X
MATLAB
Method
Data
Results
Environment Y
Python

Part I: Reproducibility: case studies

Part II : Etiology of Irreproducibility

Part III : Some therapeutic proposals

Reproducibility and reliability of biomedical research: improving research practice

Symposium report, October 2015

The Academy of Medical Sciences

BBSRC bioscience for the future

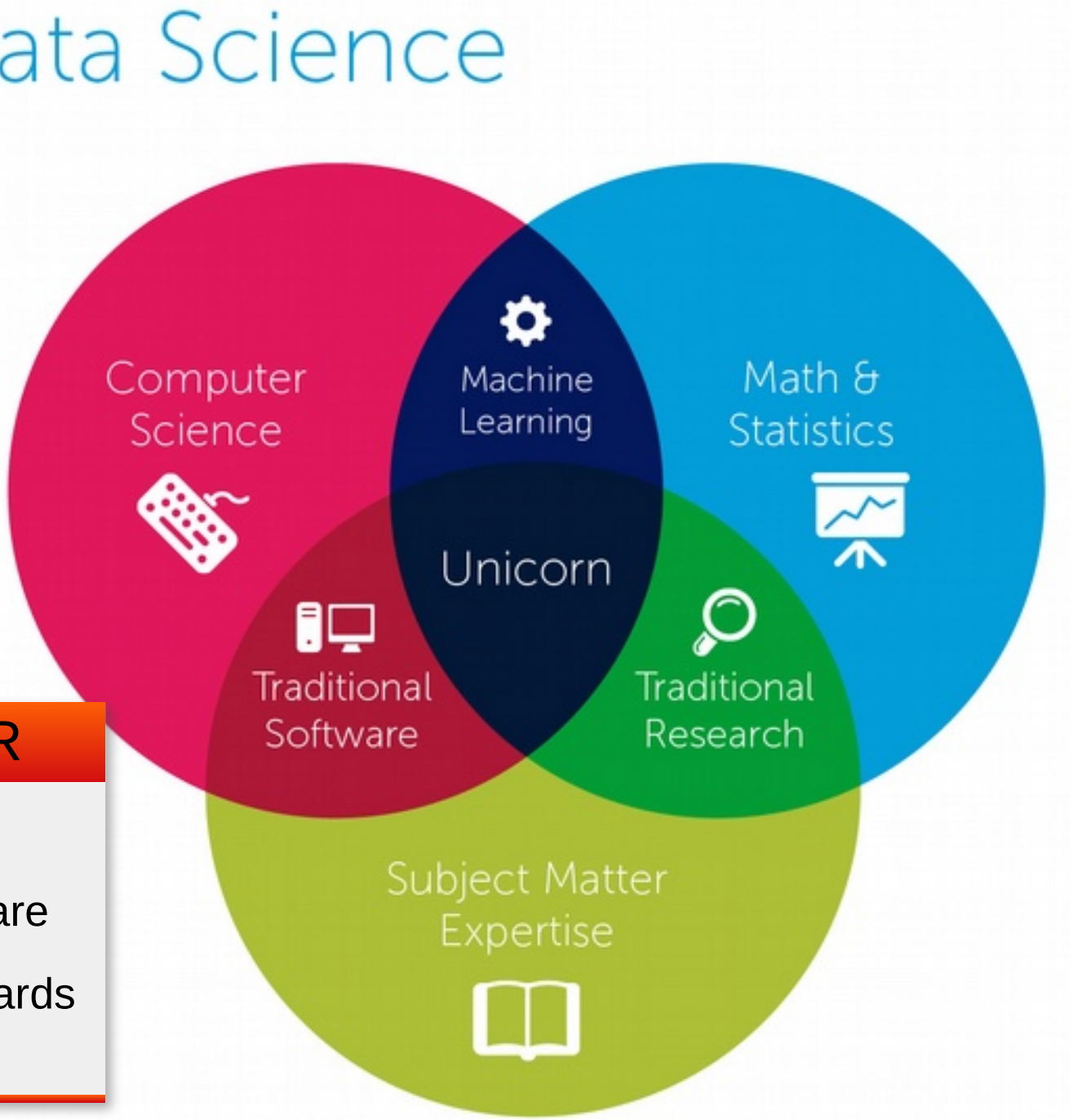
MRC Medical Research Council

wellcome trust

What can we do ?

- Improve Training
- Develop better tools – make these tools that could change the culture
- Change the incentives

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



FAIR

Data

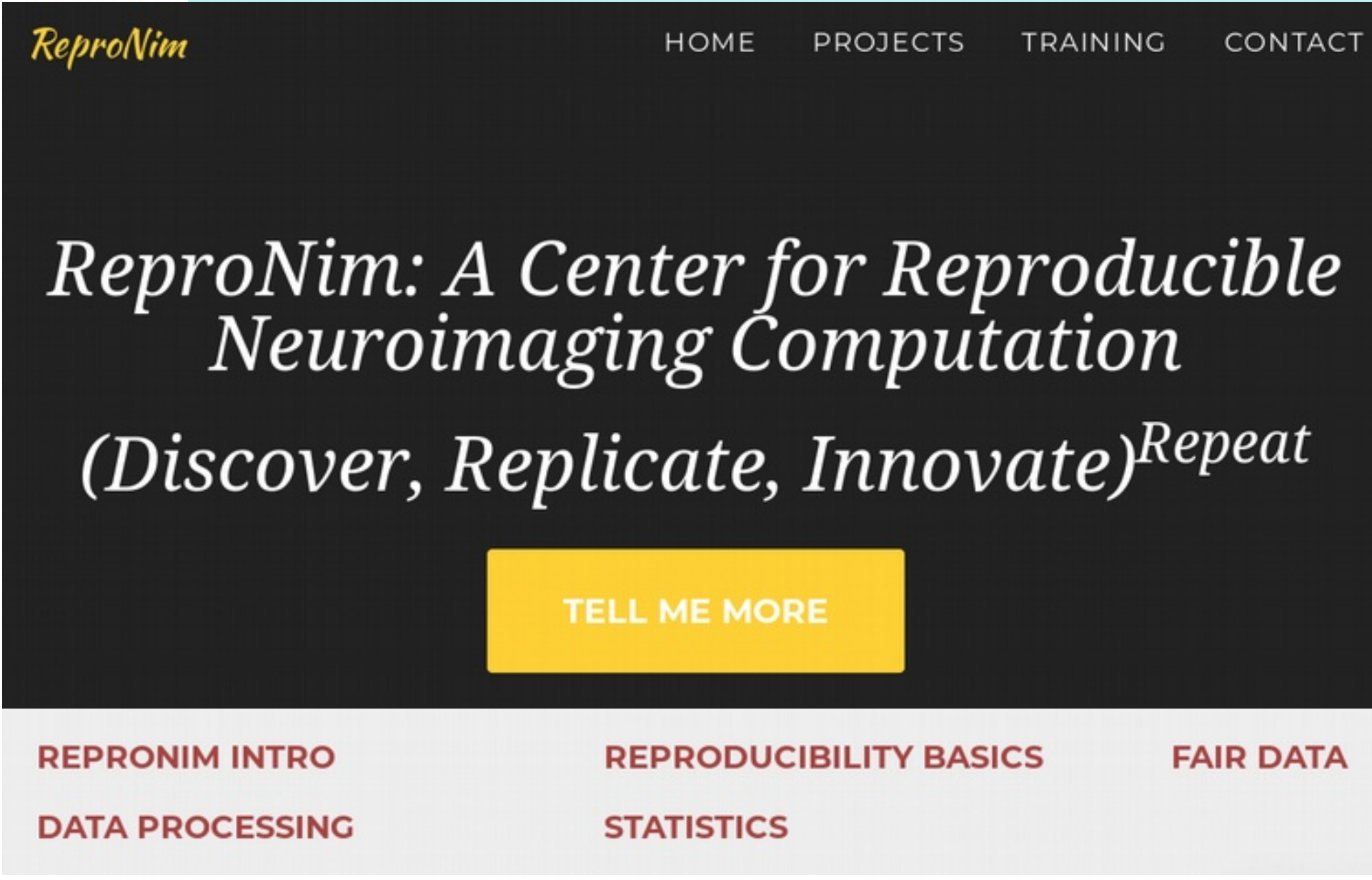
Software

Standards reuse

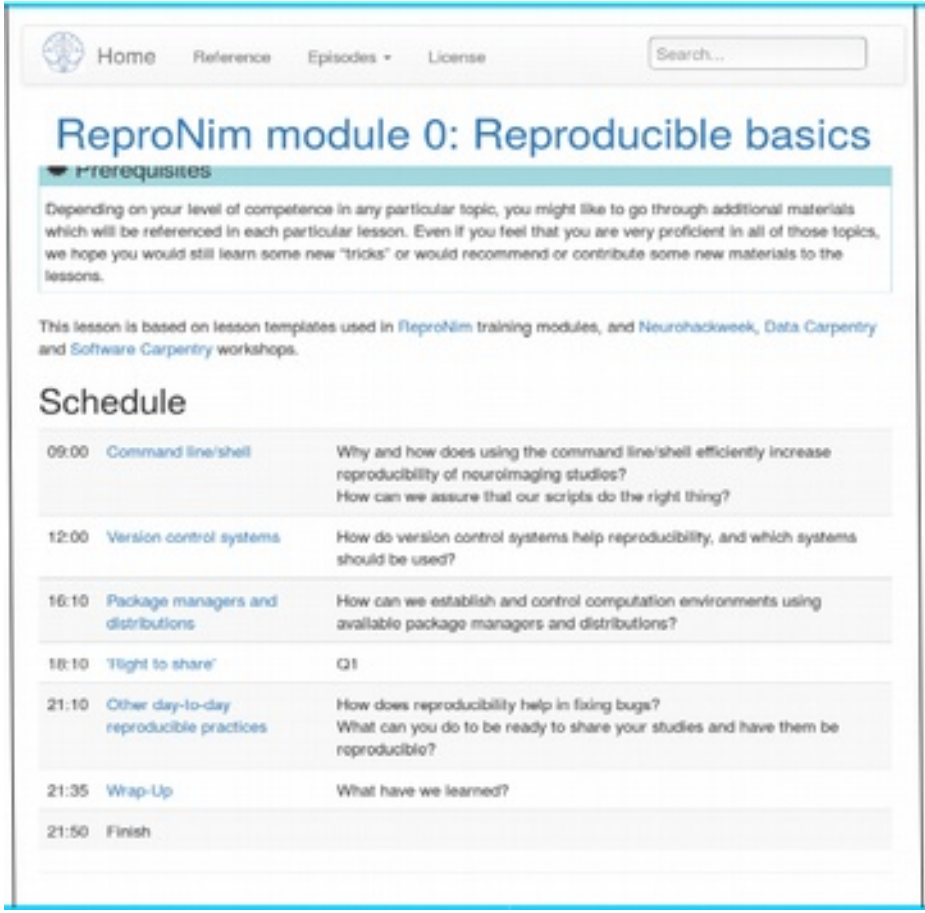
Research

Science and Technology

Training: NIH P41 ReproNim



Training



Home Reference Episodes License Search...

ReproNim module 0: Reproducible basics

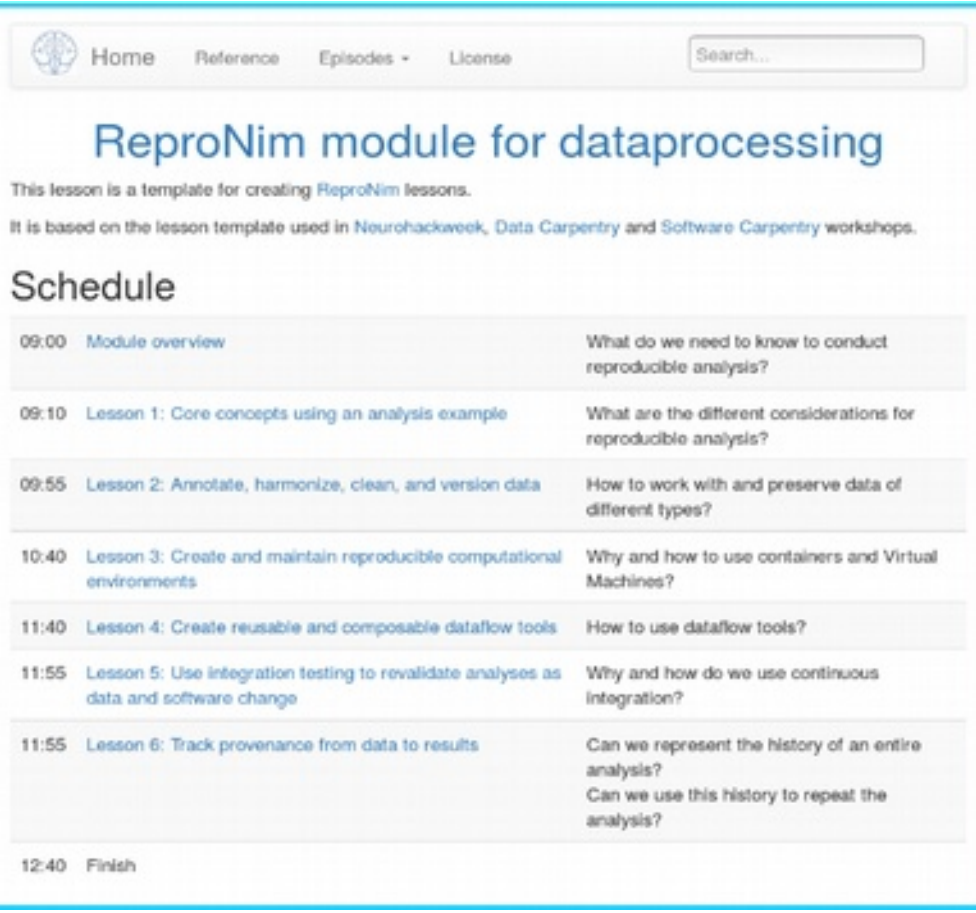
Prerequisites

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	

SCIENTIFIC DATA

OPEN

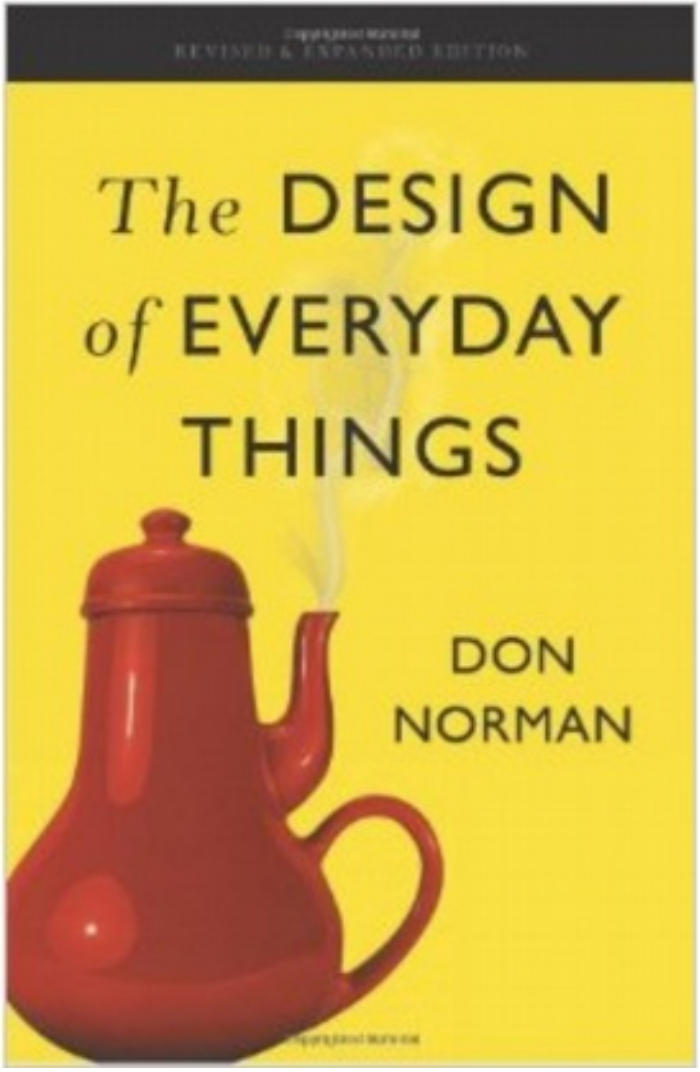
Comment: High-quality science requires high-quality open data infrastructure

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.

Building functional tools



Changing the publication model

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

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

Dropout

Srivastava, Nitish, et al. "Dropout: A simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014): 1929-1958.

- Srivastava's Master's(!) thesis.
- Training scheme that randomly masks neurons at every step.
- Usually gives a small performance boost.
- Mysterious.

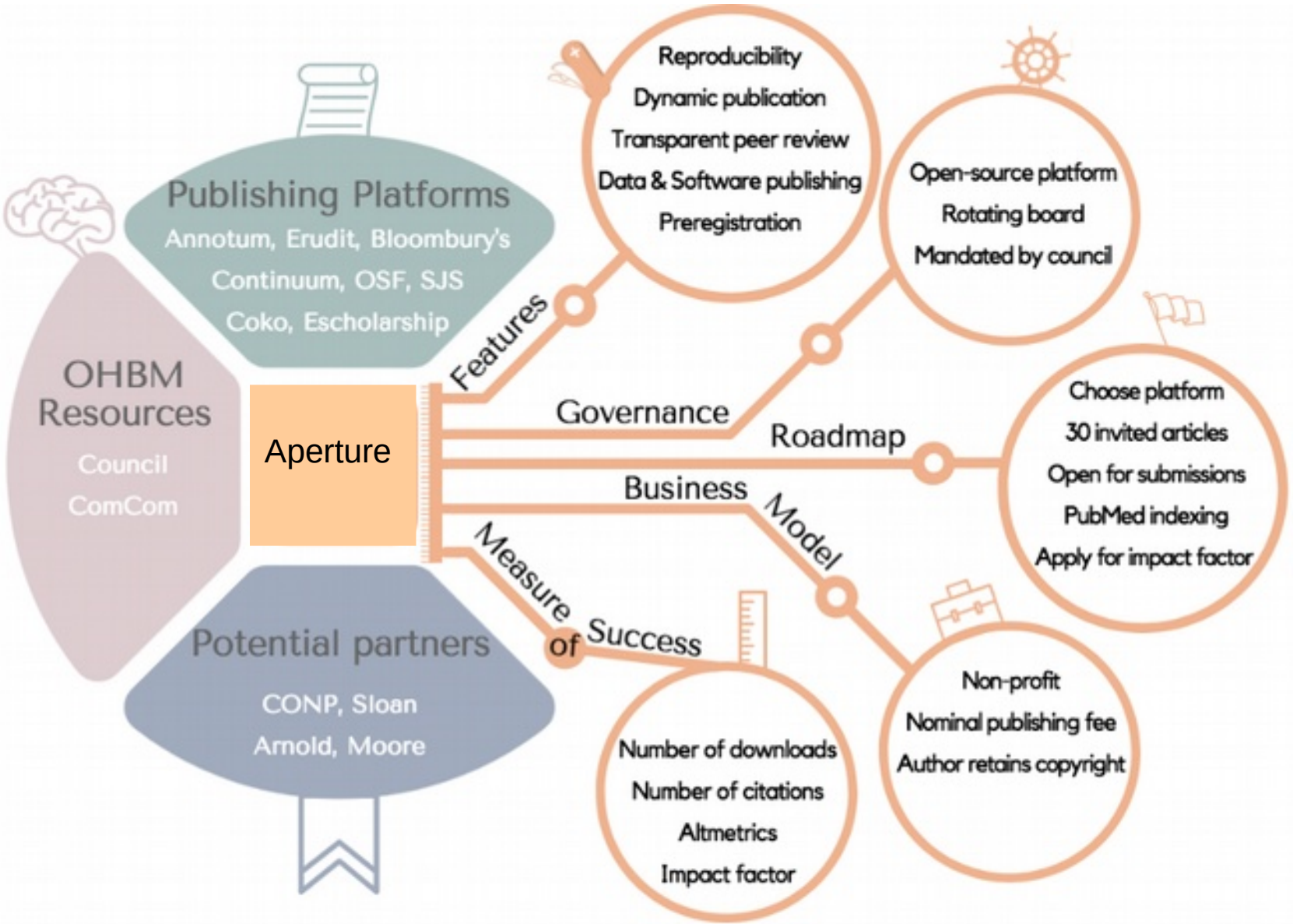
This paper was rejected from NIPS in 2012, and propagated solely as a preprint on arxiv.

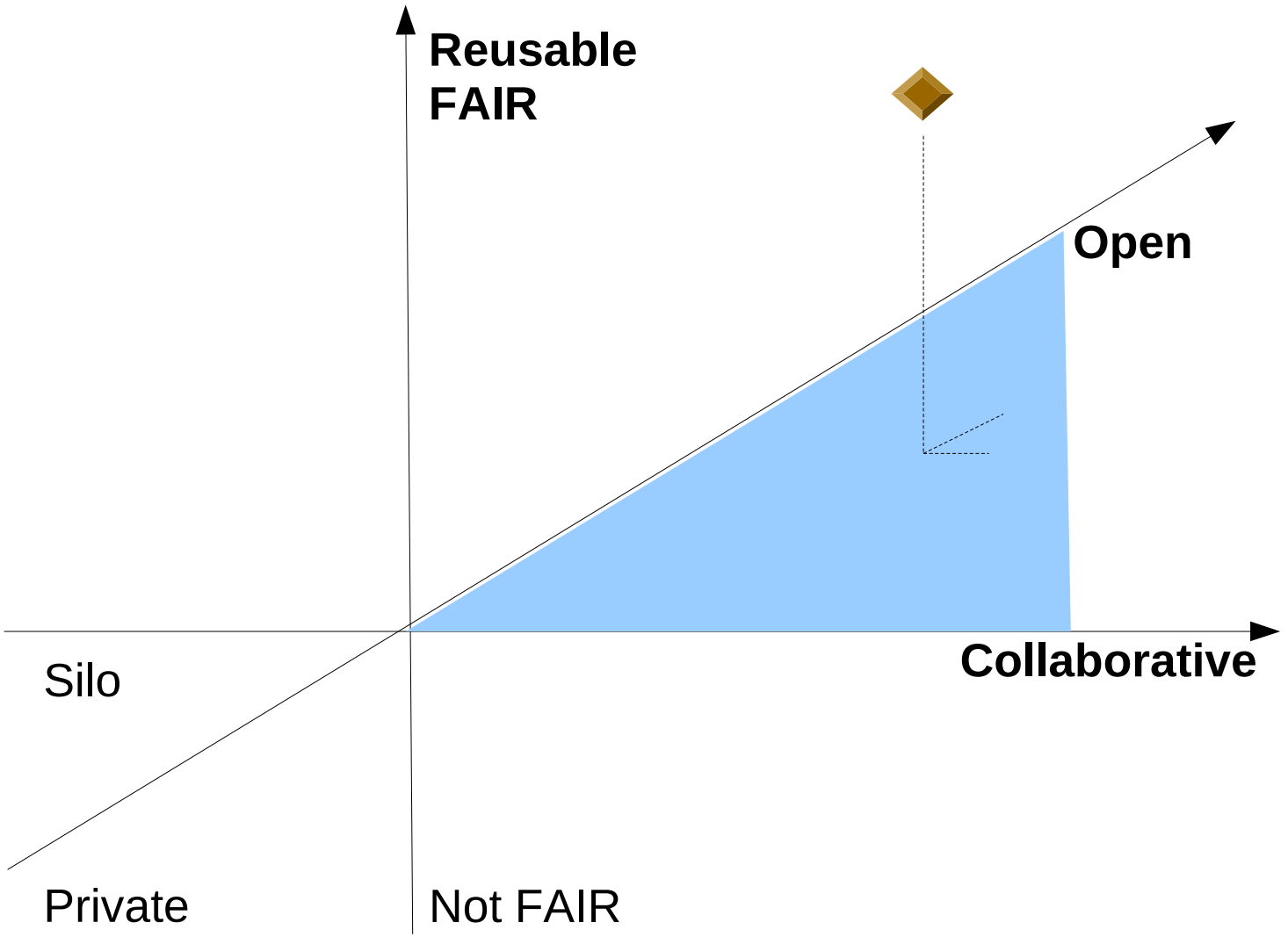
Cited by 13271 in 2019

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

OHBM Open Publishing Initiative




Reusable
FAIR
Open
Collaborative
Silo
Private
Not FAIR

Thank you

- My **McGill** colleagues: S. Brown, T. Glatard, G. Kiar, A. Evans, C. Greenwood, and others
- My **ReproNim** colleagues: D. Kennedy, D. Keator, S. Ghosh, M. Martone, J. Grethe, M. Hanke, Y. Halchenko
- My **Berkeley** colleagues: S. Van der Walt, M. Brett, J. Millman, Dan Lurie, M. D'Esposito, et al
- My **Pasteur** colleagues: G. Dumas, R. Toro, T. Bourgeron, , and others
- My **Paris** colleagues: B. Thirion, G. Varoquaux, V. Frouin, et al

Questions ?