Reproducibility in neuroimaging: Challenges and solutions

Russell Poldrack

Department of Psychology
Stanford University

Wellcome Images

Science in crisis (?)

298 | NATURE | VOL 485 | 17 MAY 2012

IN THE WAKE OF HIGH-PROFILE CONTROVERSIES, PSYCHOLOGISTS ARE FACING UP TO PROBLEMS WITH REPLICATION.

BY ED YONG

Rigorous replication effort succeeds for just two of five cancer papers

By **Jocelyn Kaiser** | Jan. 18, 2017 , 1:00 PM

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

29 MARCH 2012 | VOL 483 | NATURE | 531

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

SCIENCE sciencemag.org 28 AUGUST 2015 • VOL 349 ISSUE 6251

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

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
- The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Neuroimaging: a perfect storm for irreproducibility

Altered Brain Activity in Unipolar Depression Revisited

Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD[1,2]; Edna C. Cieslik, PhD[1,2]; Ilinca Serbanescu, MSc[1]; et al

JAMA Psychiatry. 2017;74(1):47-55.

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.

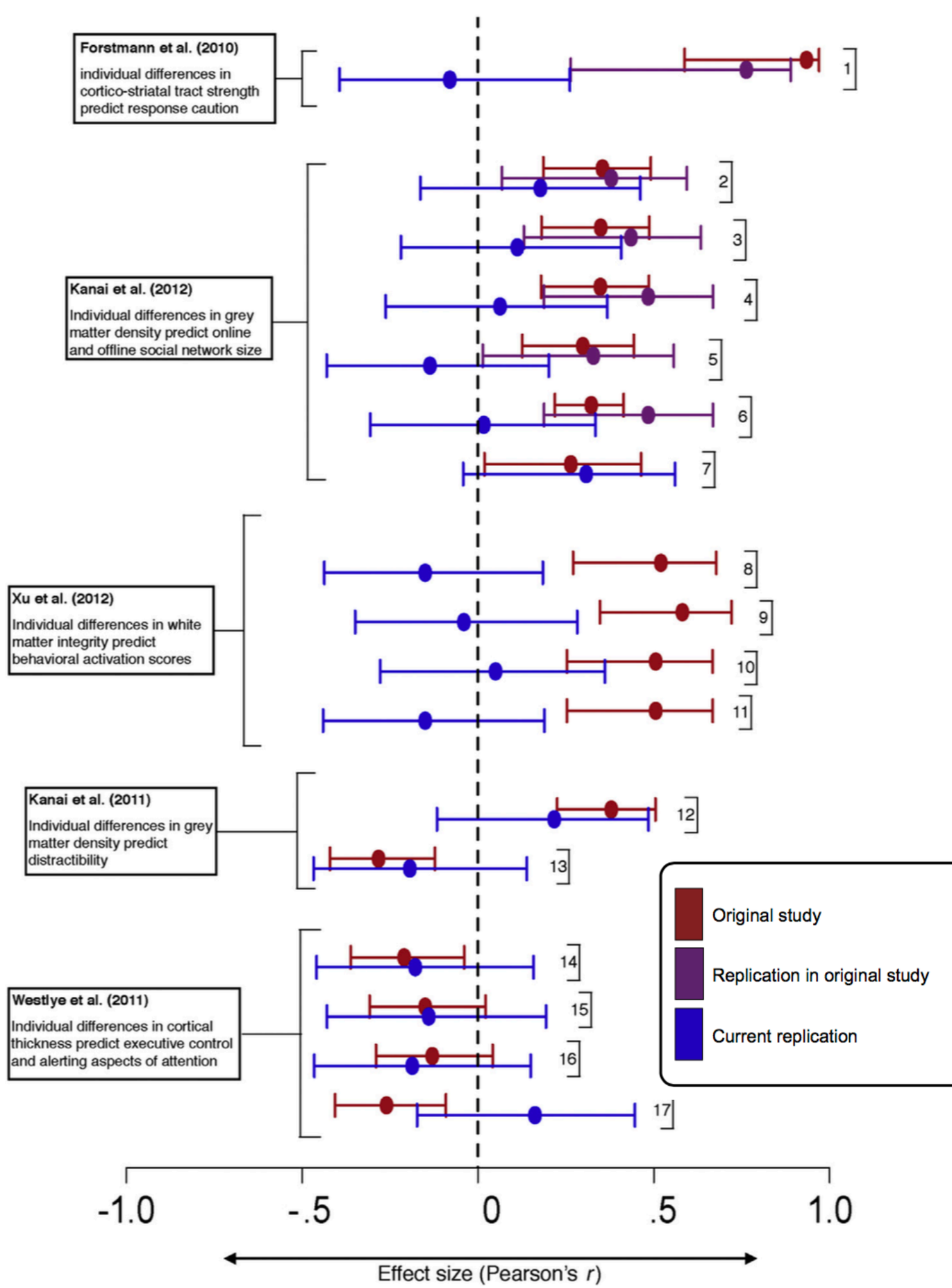
Irreproducible correlations

A purely confirmatory replication study of structural brain-behavior correlations

Wouter Boekel [a,], Eric-Jan Wagenmakers [a], Luam Belay [a], Josine Verhagen [a], Scott Brown [b] and Birte U. Forstmann [a]*

CORTEX 66 (2015) 115–133

Here, we attempt to replicate five structural brain-behavior correlation studies comprising a total of 17 effects. To prevent the impact of QRPs we employed a preregistered, purely confirmatory replication approach. For all but one of the 17 findings under scrutiny, confirmatory Bayesian hypothesis tests indicated evidence in favor of the null hypothesis ranging from anecdotal (Bayes factor < 3) to strong (Bayes factor > 10). In several studies, effect size estimates were substantially lower than in the original studies.



Stanford Center For
Reproducible Neuroscience

Making neuroscience open and reproducible

WHO WE ARE

Reproducibility matters

Neuroscience research is the basis for critical decisions about health and society. Our first goal as researchers is to ensure that the results of our research will stand the test of time.

Enabling better research

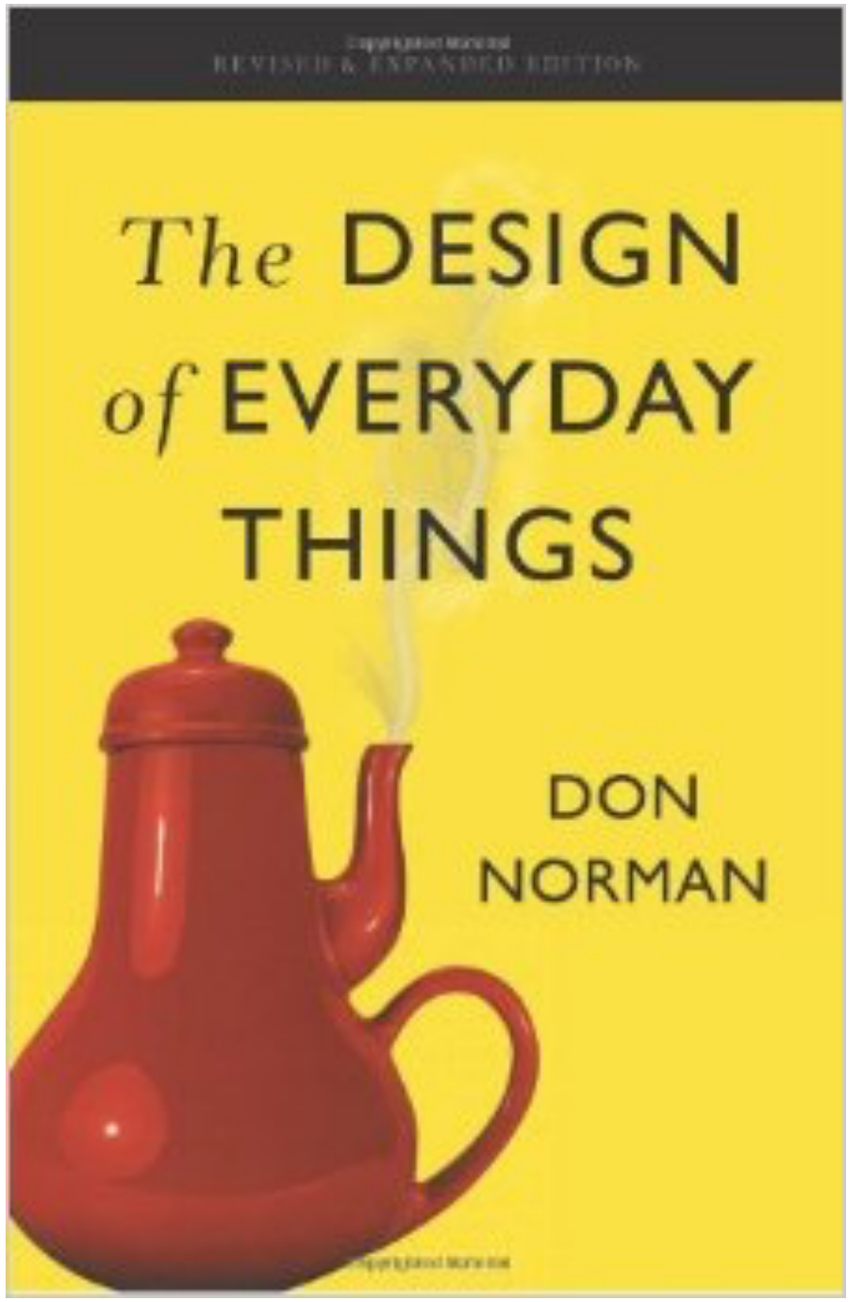
We are expanding the OpenfMRI project into a free and open platform that will enable the analysis and sharing of neuroimaging data, harnessing the power of high-performance computing to improve the quality of research.

From data to discovery

Our platform will provide neuroimaging researchers with leading-edge tools to analyze and share large datasets, with a focus on quantifying the reproducibility of the results.

http://reproducibility.stanford.edu

Designing a more reproducible scientific enterprise



FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

Cognitive biases in scientific reasoning

- "The first principle is that you must not fool yourself and you are the easiest person to fool"
  - R. Feynman

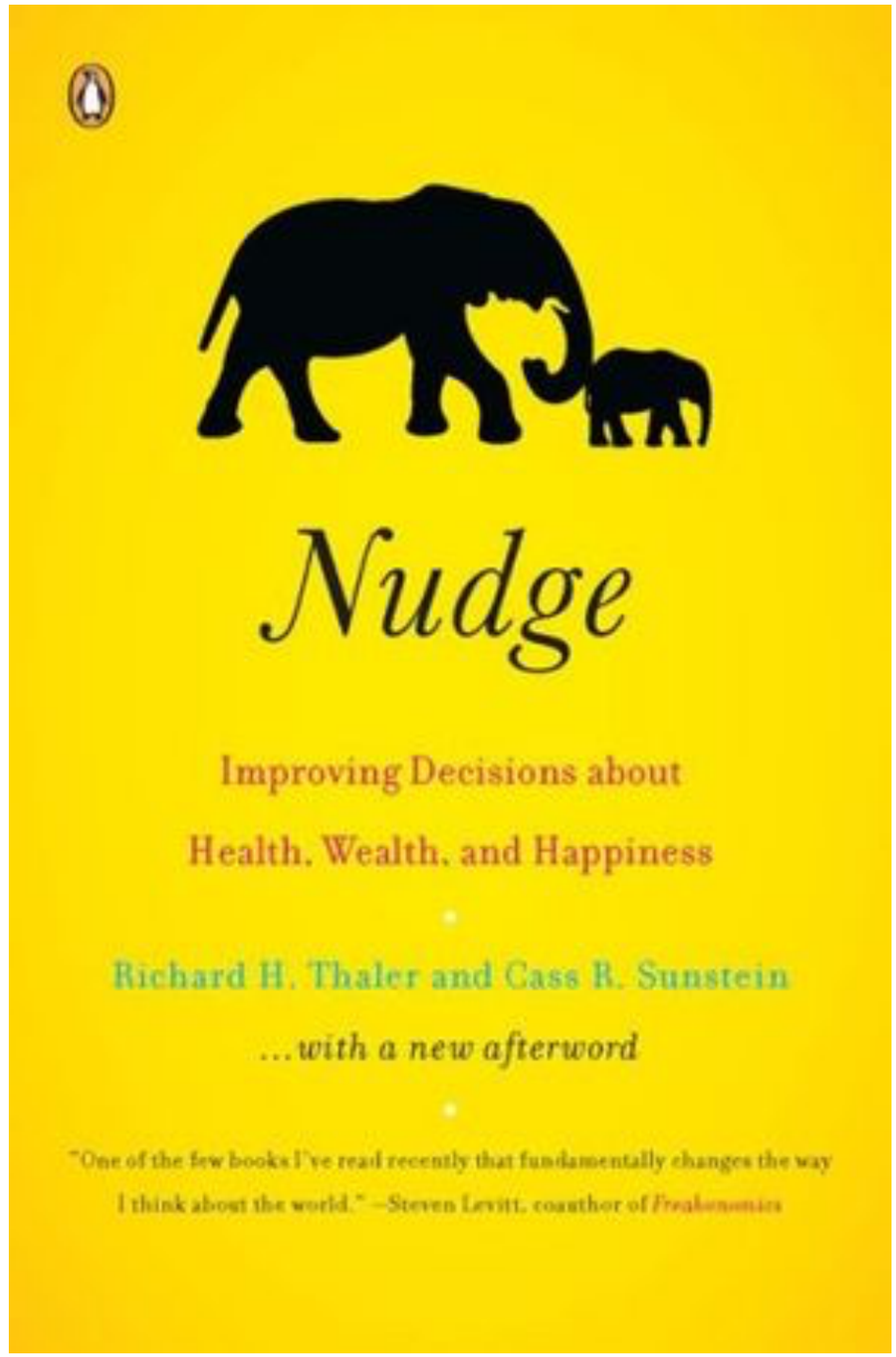
- We pay more attention to information that confirms our hypotheses or biases versus those that disconfirm them

- We fail to consider alternative hypotheses that could explain the data

- We fail to consider base rates

Improving the choice architecture of science

- Choice architecture
  - particular set of features that drive people toward or away from particular choices
- Nudges
  - Improving incentives
  - Using the power of defaults
  - Providing feedback
  - Expecting and prevent errors

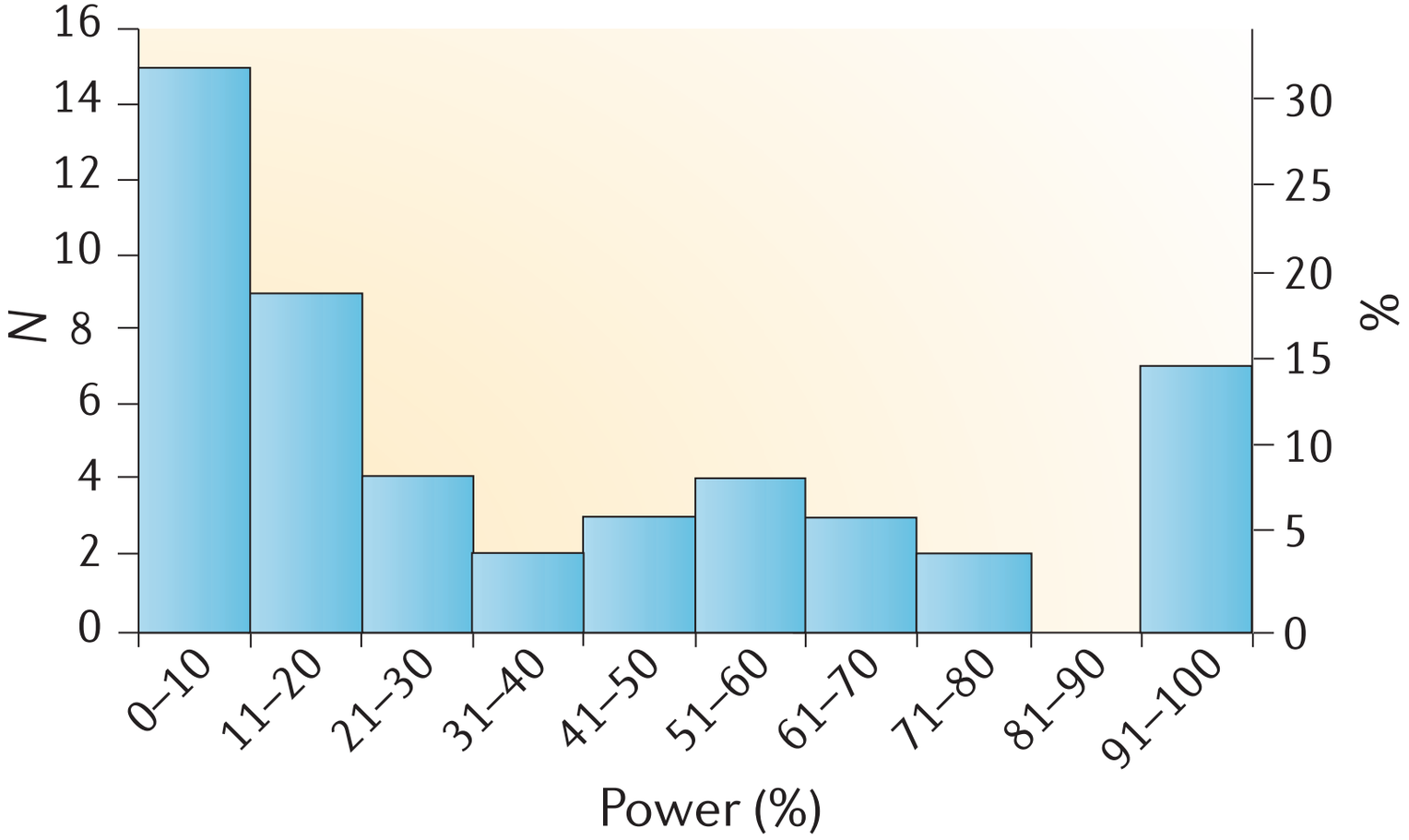


Threats to reproducibility: Low power

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

NATURE REVIEWS | NEUROSCIENCE VOLUME 14 | MAY 2013 | 365

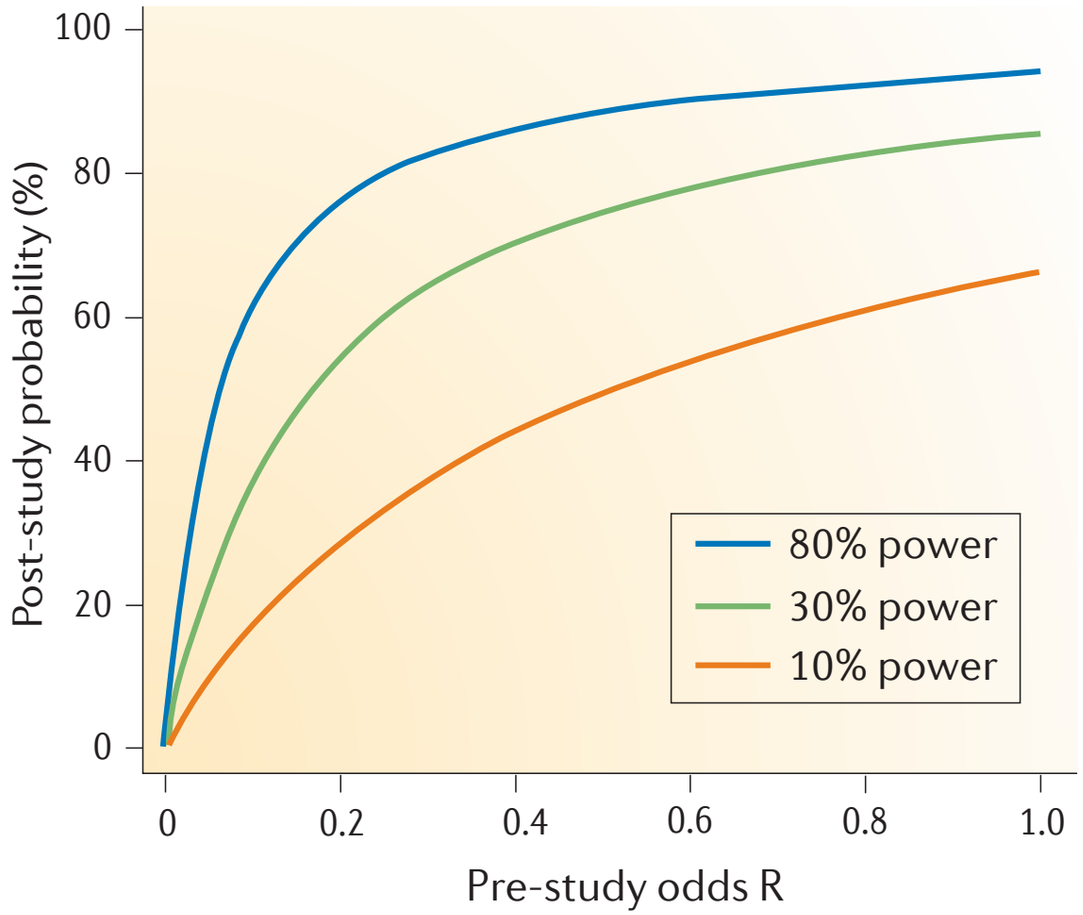


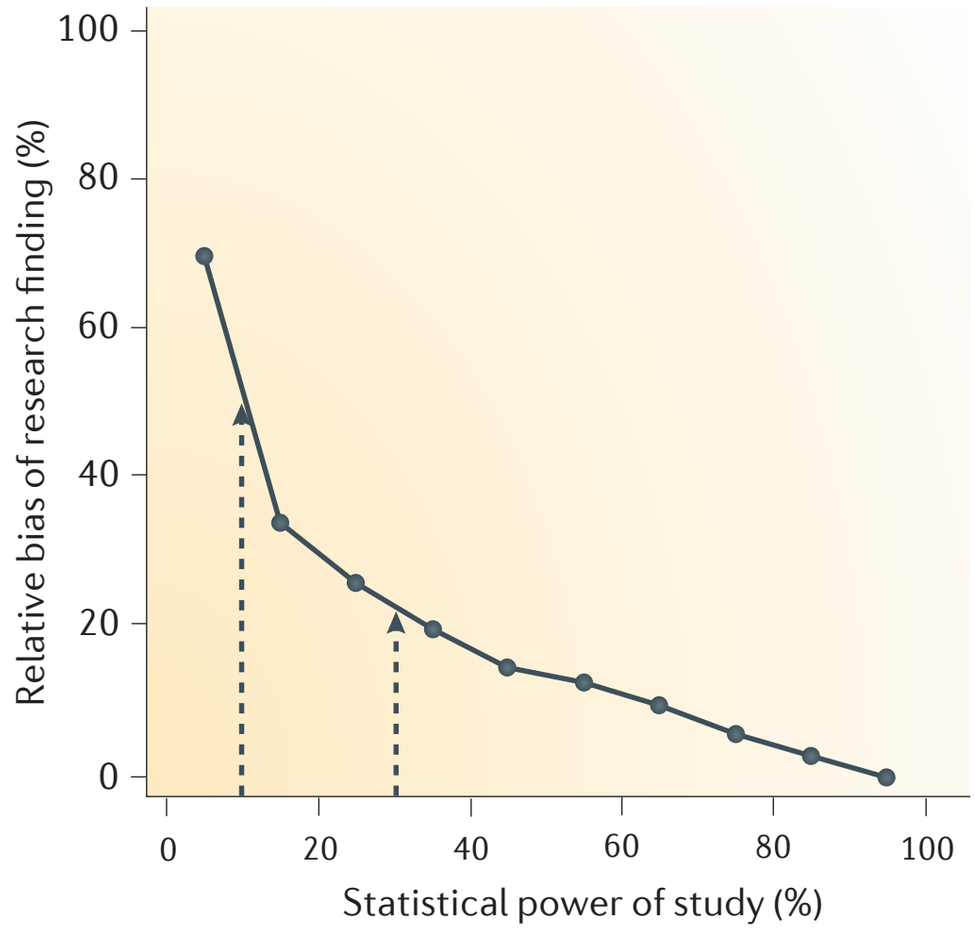
Low power -> unreliable science

Positive Predictive Value (PPV): The probability that a positive result is true

Winner's Curse: overestimation of effect sizes for significant results

$$PPV = ([1 - \beta] \times R) / ([1 - \beta] \times R + \alpha)$$





Button et al., 2013

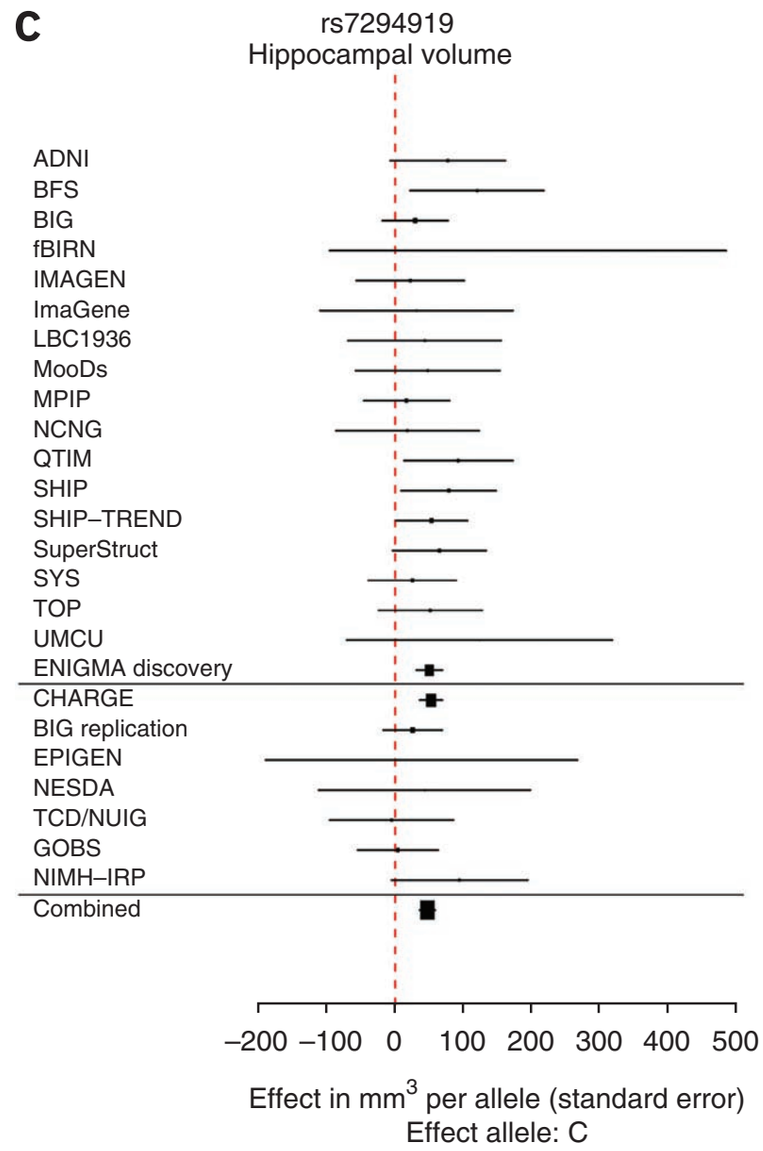
Underpowered science + publication bias: candidate genes

Identification of common variants associated with human hippocampal and intracranial volumes

Jason Stein et al. for the Enigma Consortium

VOLUME 44 | NUMBER 5 | MAY 2012 NATURE GENETICS

In general, previously identified polymorphisms associated with hippocampal volume showed little association in our meta-analysis (*BDNF*, *TOMM40*, *CLU*, *PICALM*, *ZNF804A*, *COMT*, *DISC1*, *NRG1*, *DTNBP1*), nor did SNPs previously associated with schizophrenia or bipolar disorder

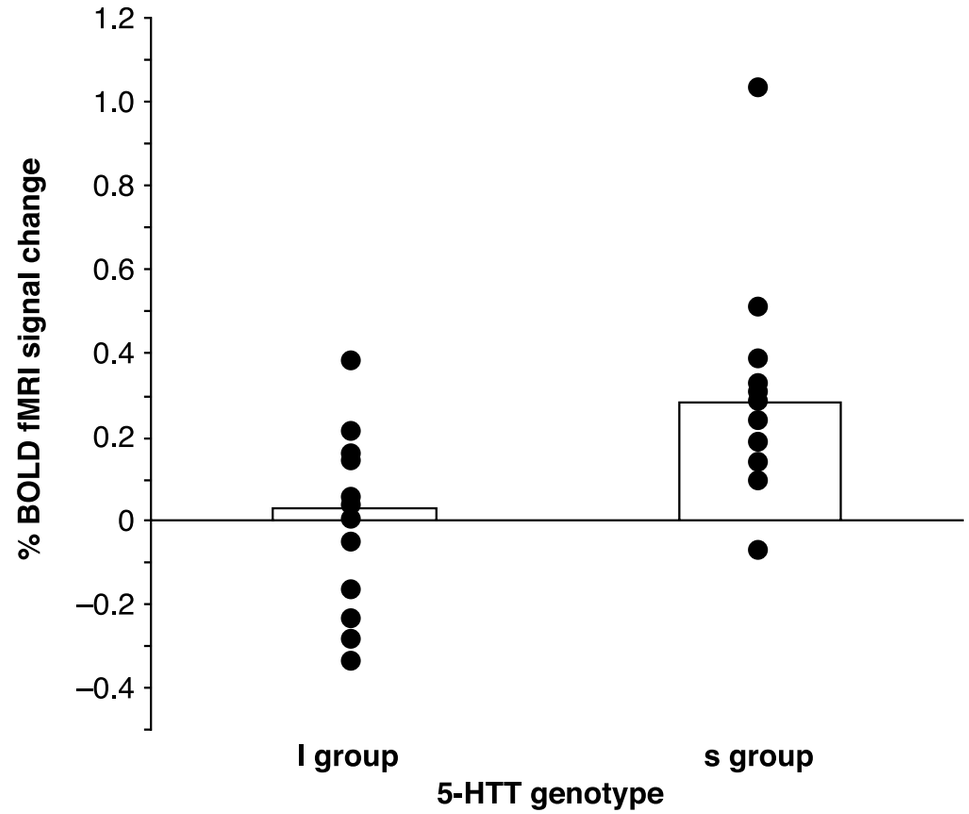


Effect size inflation due to low power

- "One important conclusion that emerges from both GWAS and sequencing studies is that there are no common variants of large effect. … for quantitative phenotypes, the expected effect sizes are less than 0.5% of phenotypic variance" (Flint & Munafo, 2013)

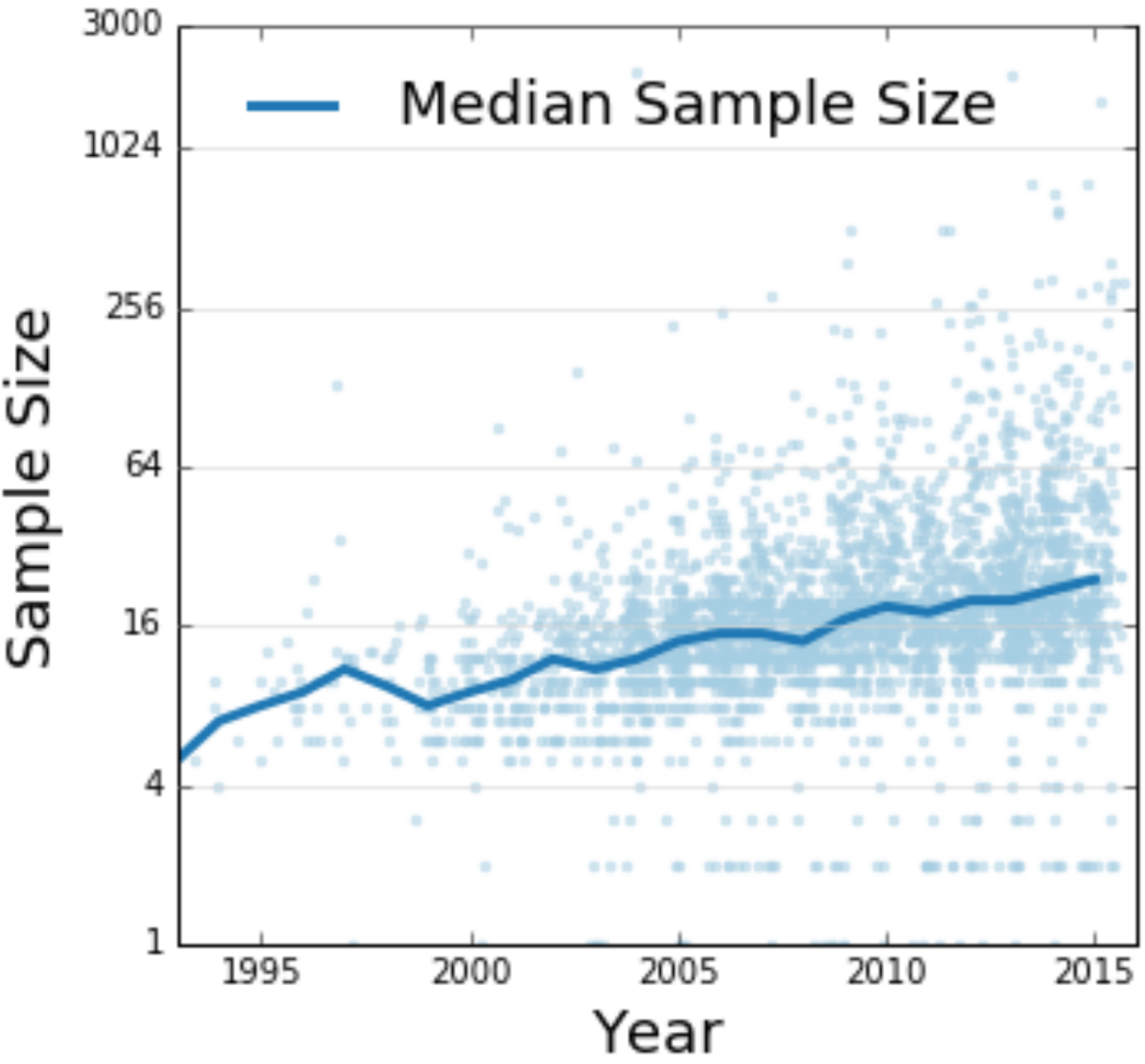
Serotonin Transporter Genetic Variation and the Response of the Human Amygdala

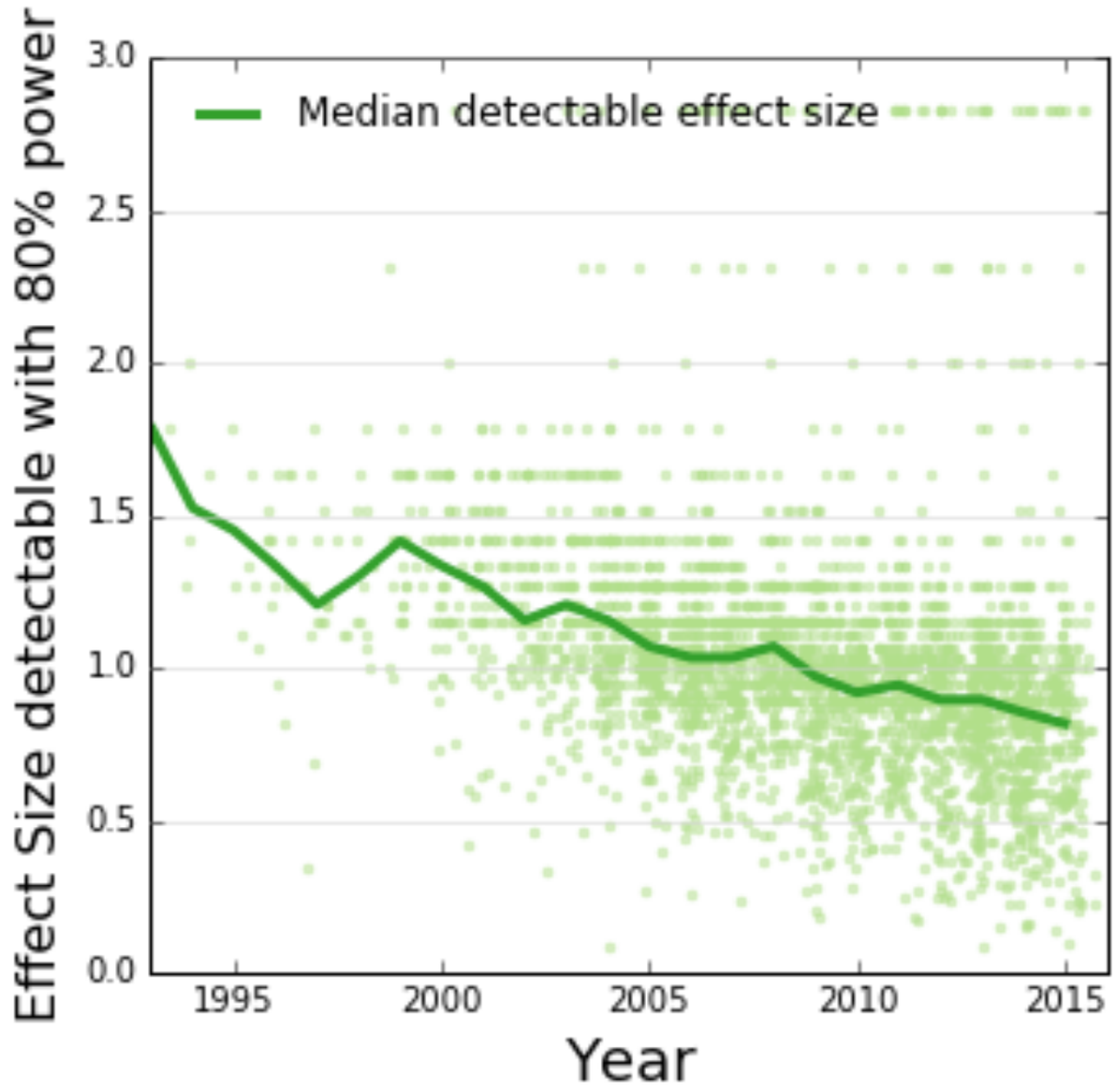
19 JULY 2002 VOL 297 SCIENCE www.sciencemag.org



"To obtain the degree of significance reported in the 2002 paper, the locus must explain about 28% of phenotypic variance"

Sample size and power in fMRI studies





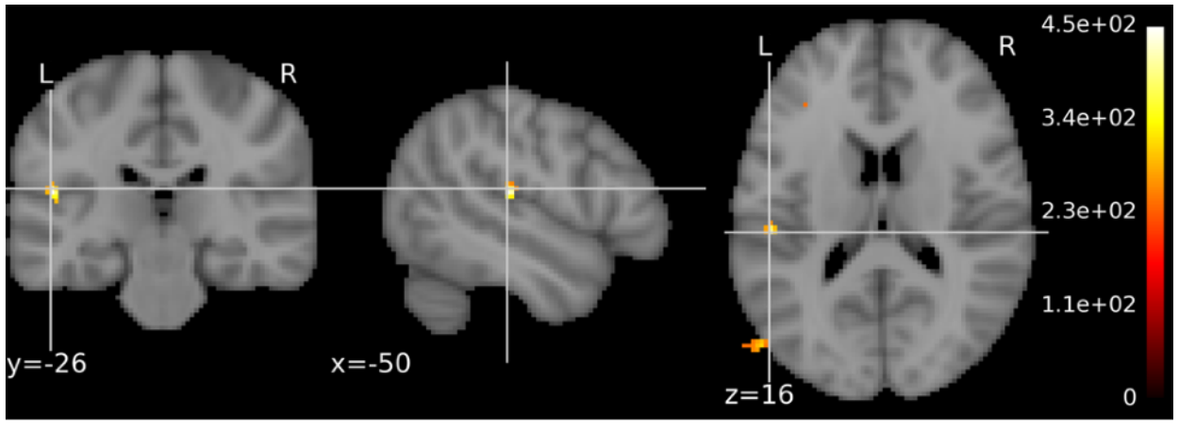
- Median study in 2015 was powered to find a single 200 voxel activation with d~0.75
- Is that a plausible effect size for fMRI?

Thanks to Sean David and Tal Yarkoni for sample size data

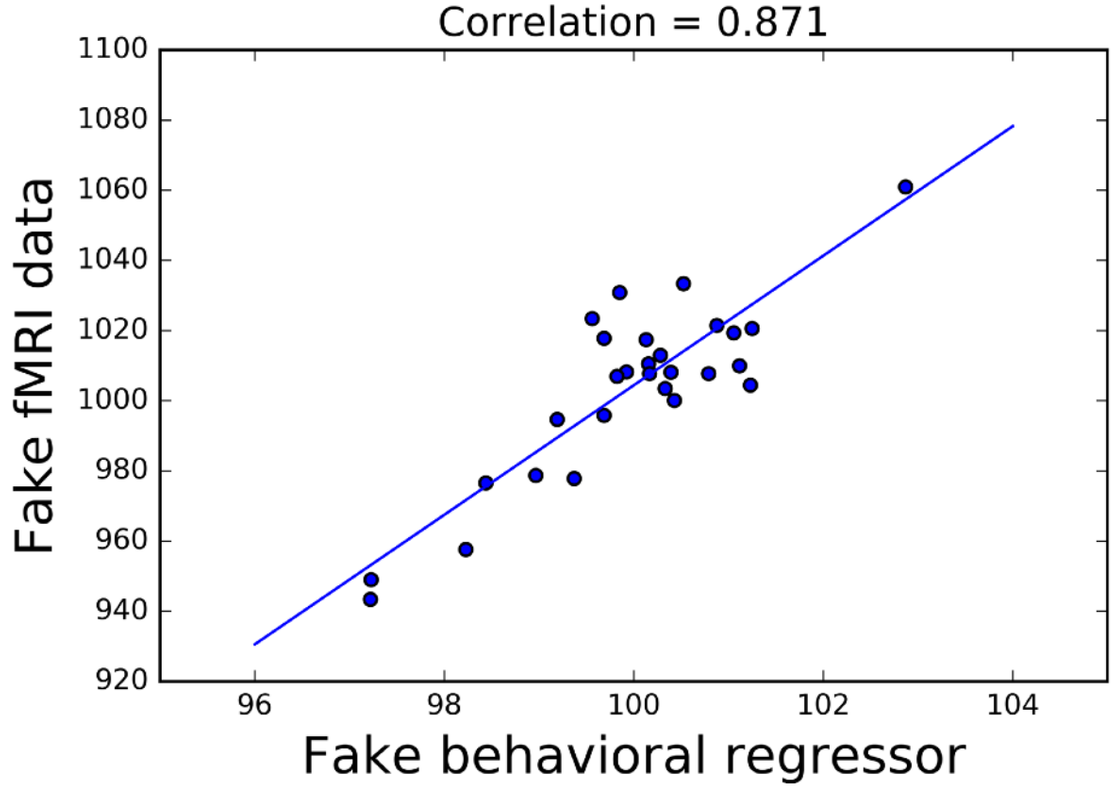
Poldrack et al, 2016, NRN

Circularity inflates effect size estimates

Correlation between random simulated behavioral variable and activation across 28 subjects



~220,000 voxels
$p < 0.001$
10 voxels cluster threshold



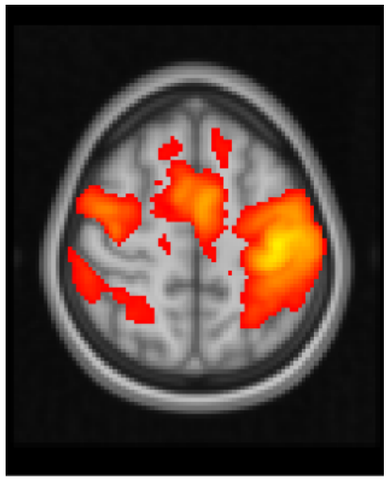
Cohen's d = 3.5

https://github.com/poldracklab/ScanningTheHorizon

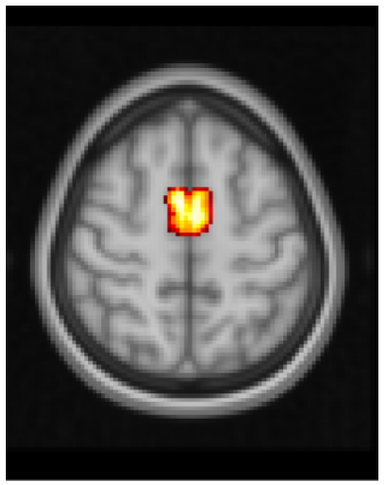
Estimating realistic effect sizes

HCP motor task data

group analysis



group map



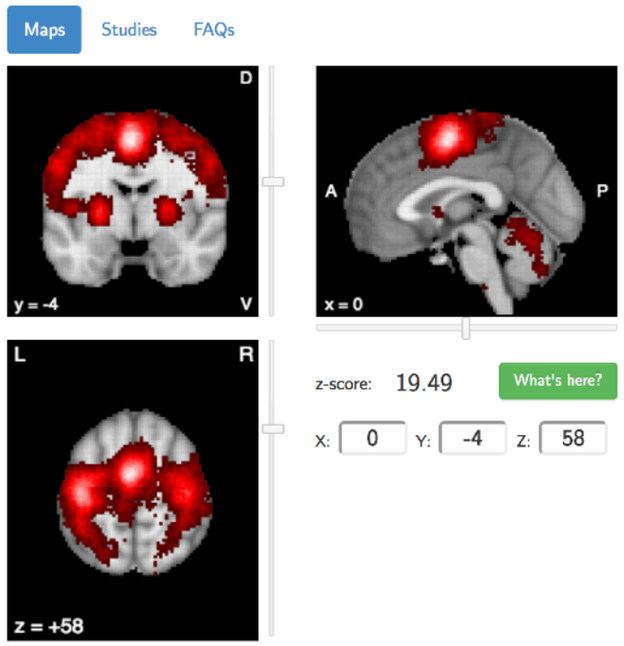
Regions of Interest

Primary Motor; Premotor; Striatum; Retinotopic Visual Areas

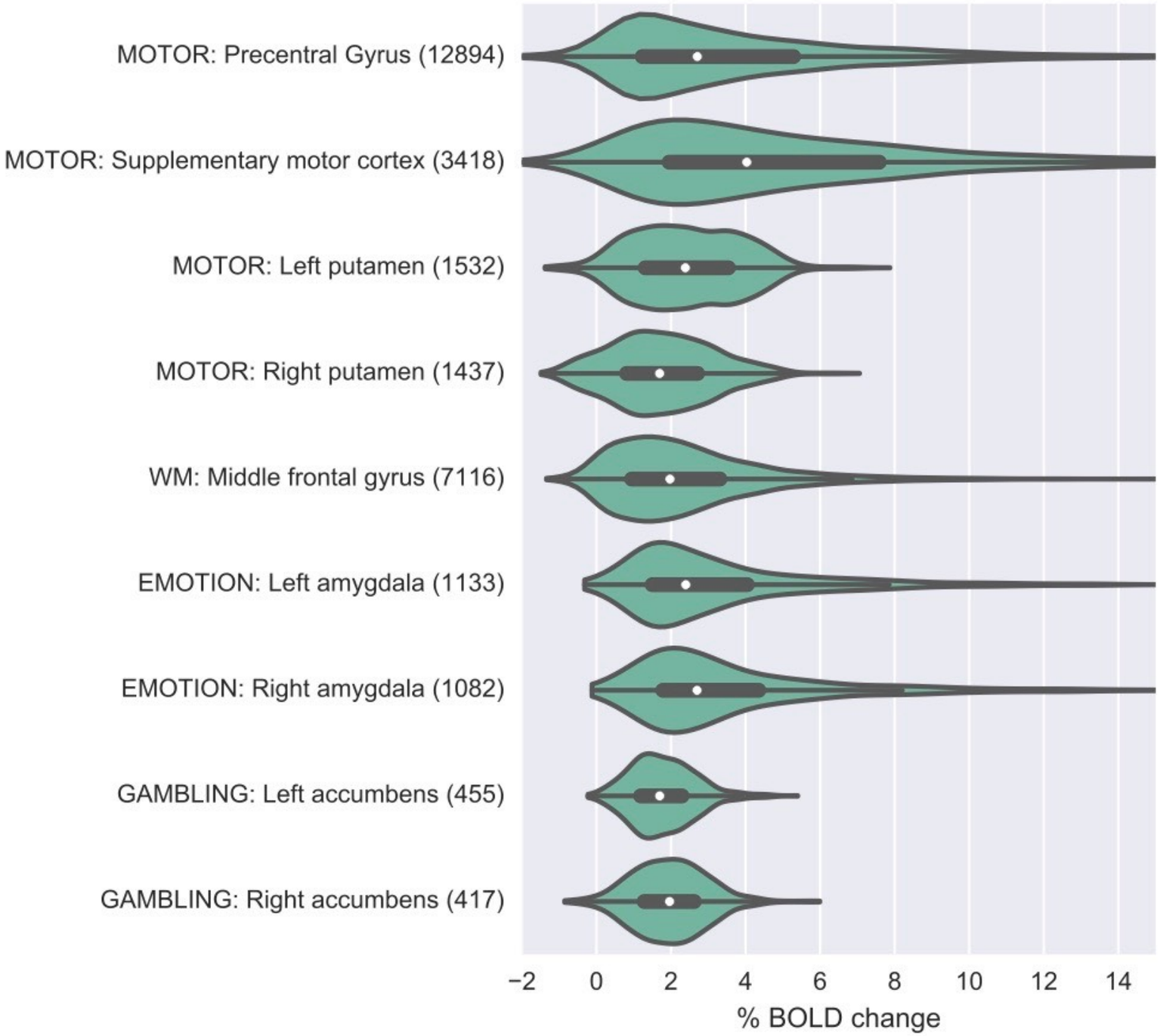
unbiased ES estimate

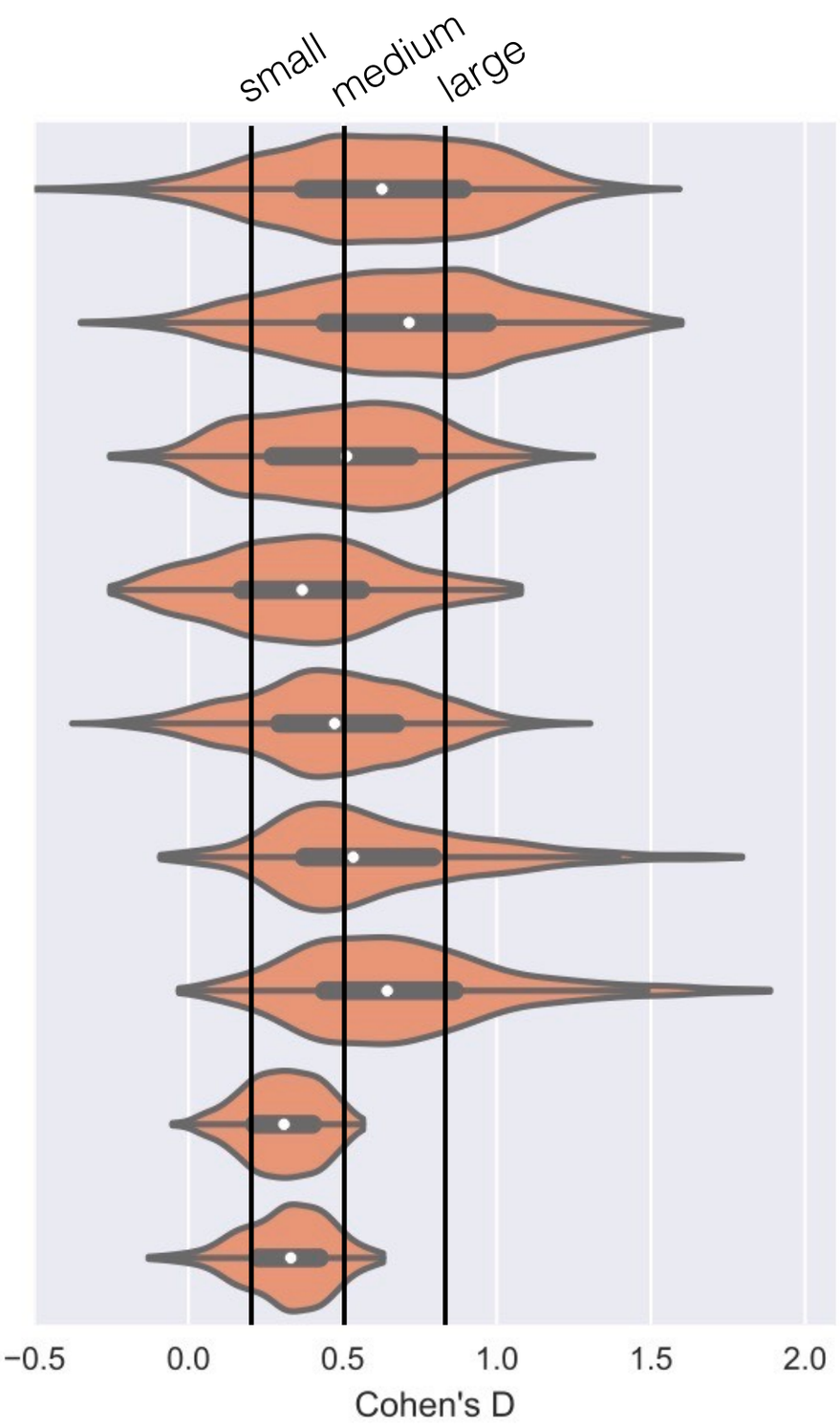
motor

An automated meta-analysis of 2081 studies



What are realistic effect sizes for fMRI?





Estimated from HCP task data
using combined anatomical + neurosynth ROIs

Poldrack et al, 2016, NRN

- "My result isn't significant, so I need to add more subjects…"

Sample size flexibility

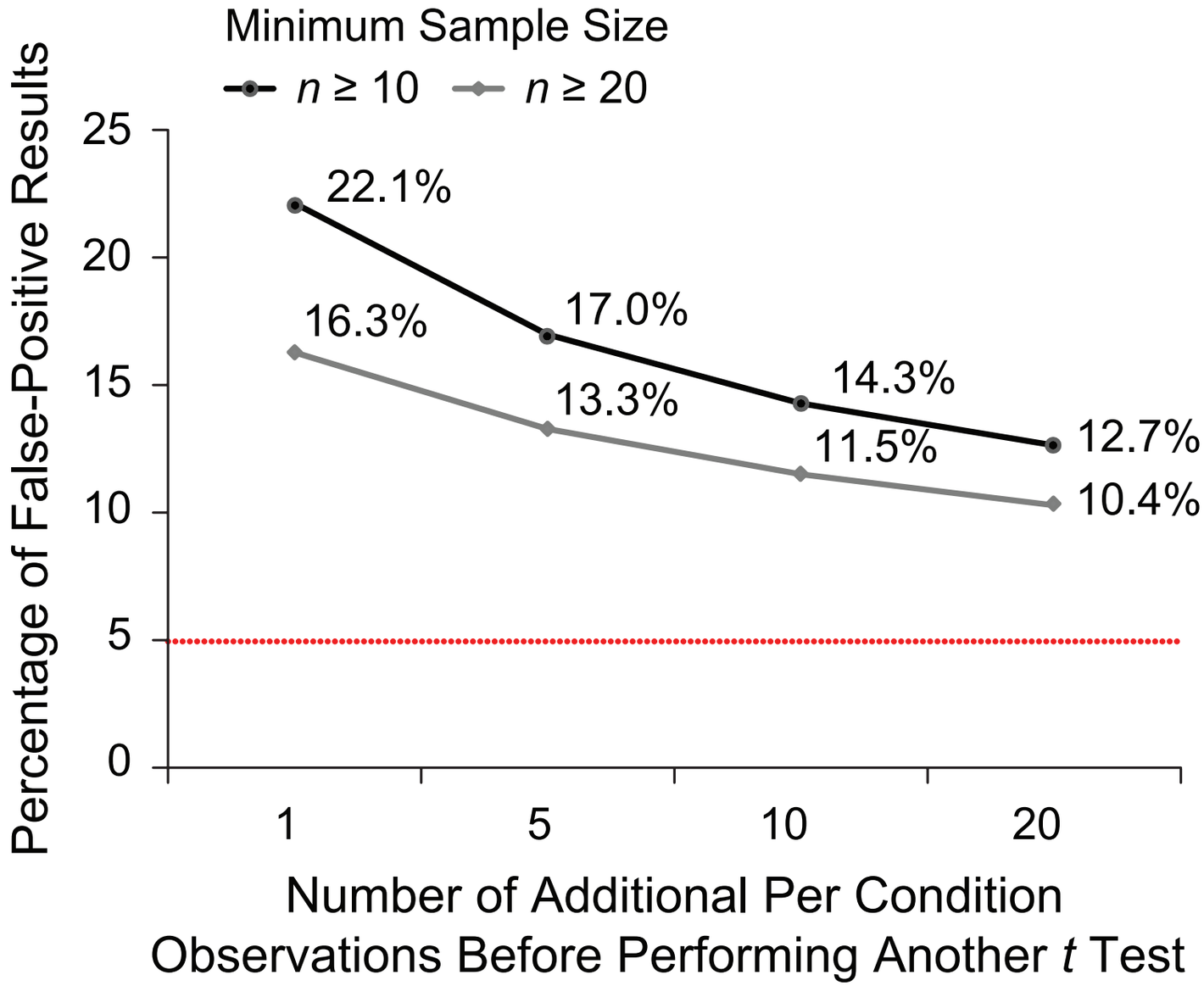


Fig. 1. Likelihood of obtaining a false-positive result when data collection ends upon obtaining significance ($p \leq .05$, highlighted by the dotted line). The figure depicts likelihoods for two minimum sample sizes, as a function of the frequency with which significance tests are performed.

-Simmons et al., 2011, Psychological Science

Improvement: always predetermine sample size

neuropowertools.org

NeuroPower

1. Load data | 2. Estimate model | 3. Power

Select your statistical parametric map for a certain contrast (T or Z) in nifti format (.nii, NOT .nii.gz).

Bladeren... spmT_0001.nii

Upload complete

Are the values Z- or T-values?

T

What is your peakforming threshold?

units = p-value

0.01

How many subjects?

18

Is the study a one- or two-sample test?

One-sample

How do you want the smoothness to be defined?

- (•) Estimate from the data
- ( ) Manual input

If manually: what is the FWHM in mm? (eg. '[8,8,8]')

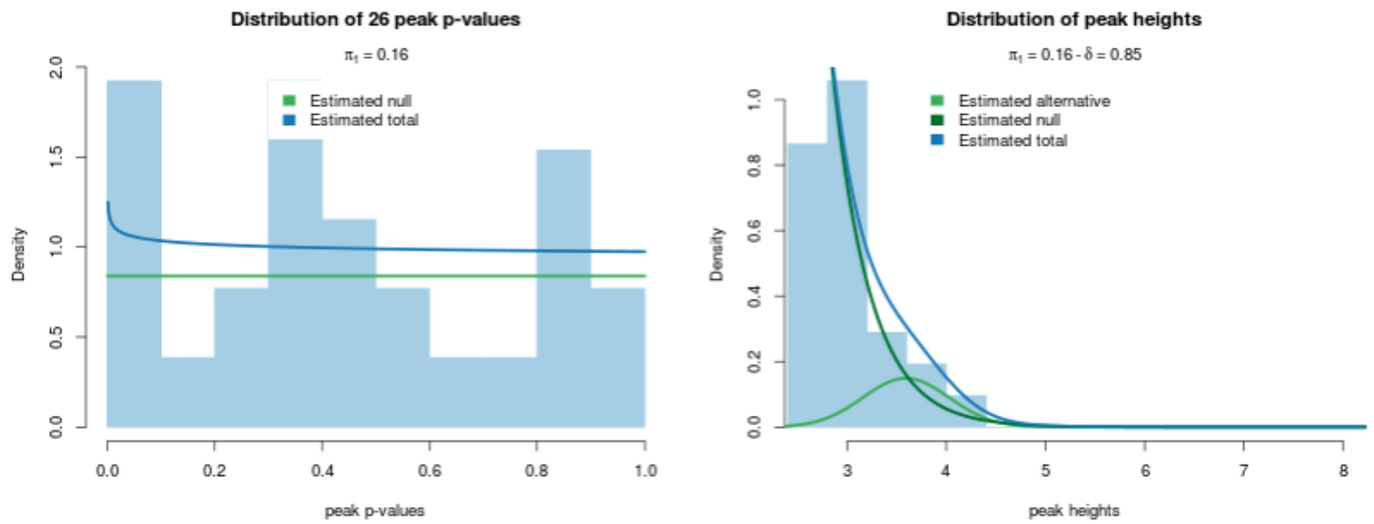
[8,8,8]

If manually: What is the voxelsize? (eg. '[2,2,2.3]')

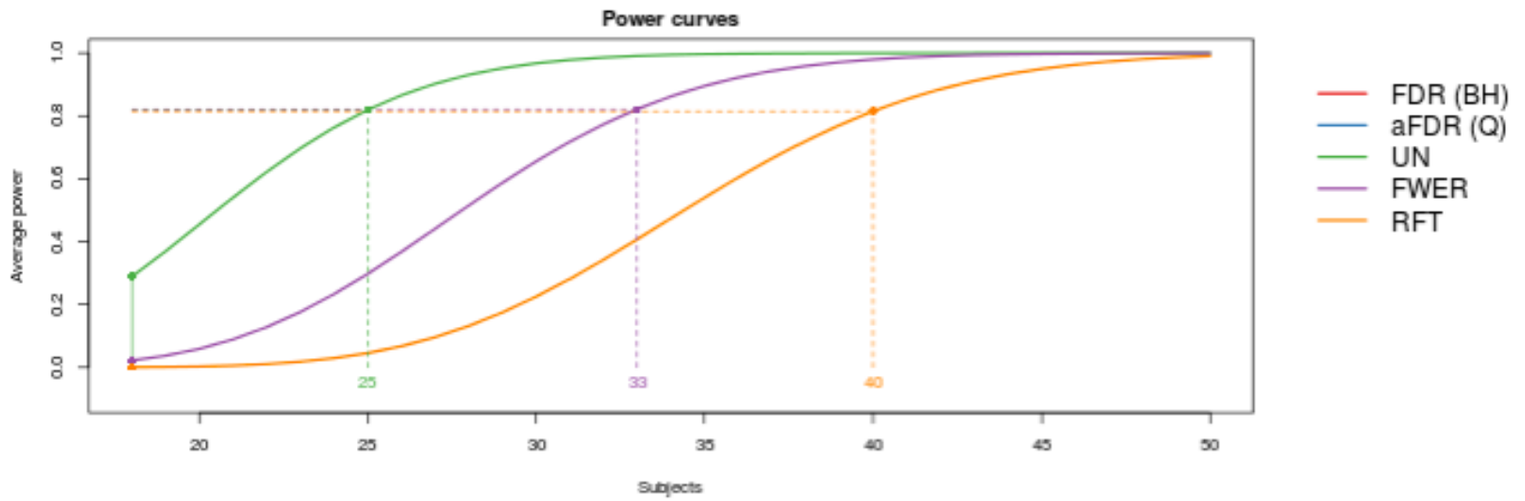
[3.9,3.9.4]

Extract peaks

Welcome | Peaks | Model fit | Post-hoc | Power



Welcome | Peaks | Model fit | Post-hoc | Power



To obtain a power level of 0.8 with RFT control at level 0.05 , the minimal sample size is 40 .

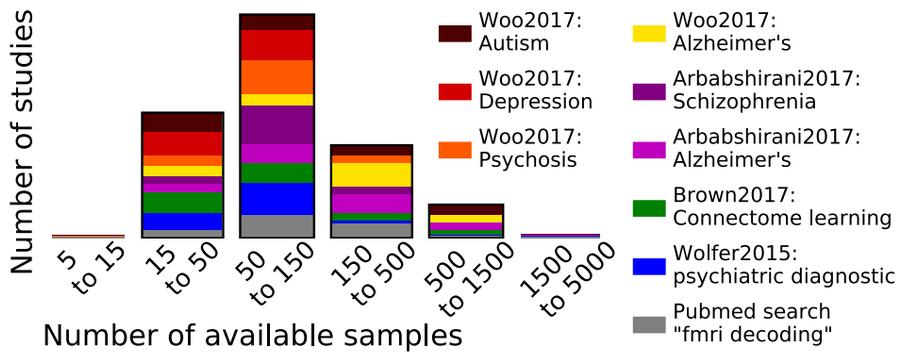
Joke Durnez

Improvement: Always pre-register study plans

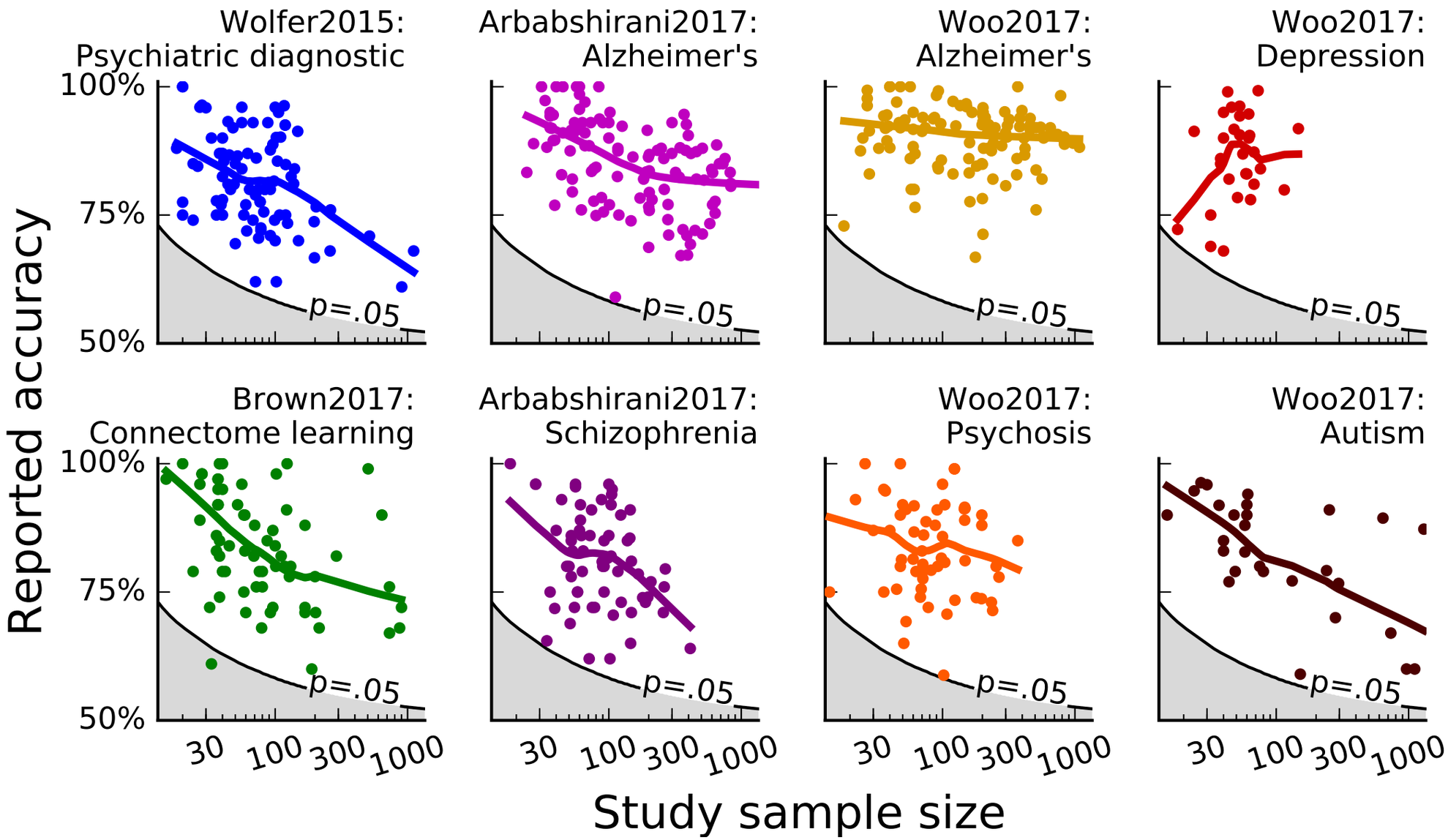
- Register sample size and analysis plan up front
- This does not prevent exploratory analysis
  - But planned and exploratory analyses should be clearly delineated in the paper

http://www.russpoldrack.org/2016/09/why-preregistration-no-longer-makes-me.html

Sample size and machine learning analyses



Varoquaux, 2017



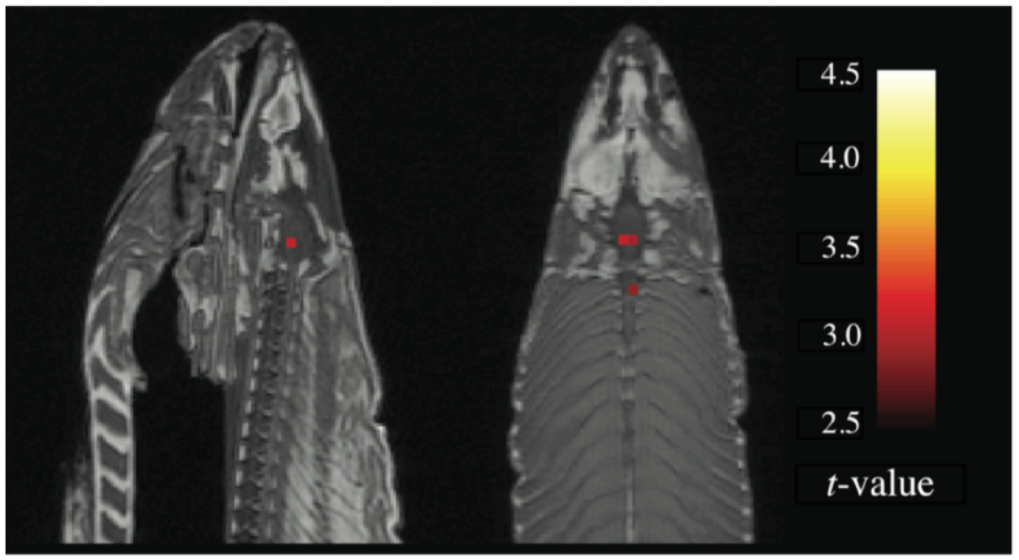
Threats to reproducibilty: high dimensionality

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

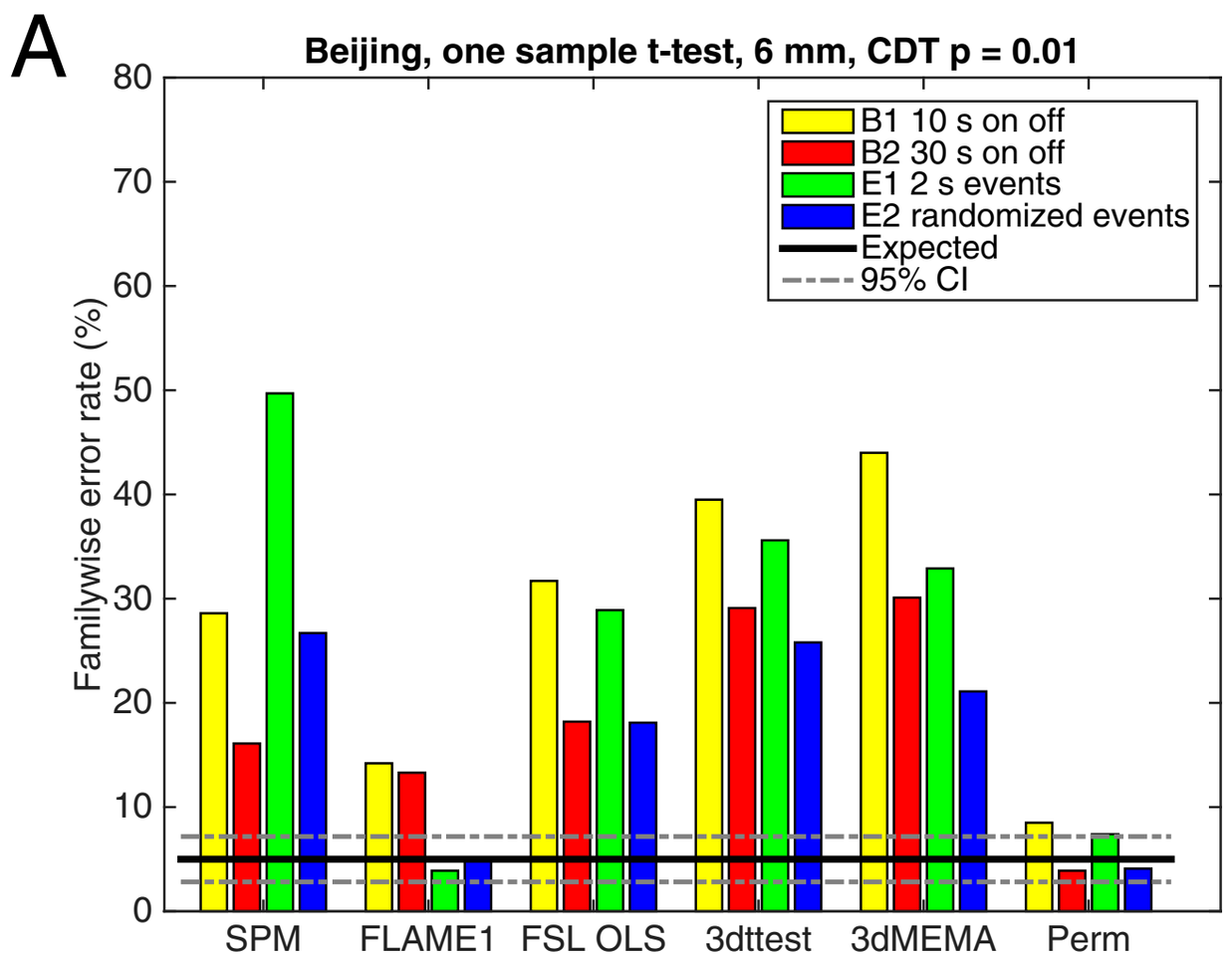
Improvement: Use nonparametric corrections

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

PNAS | **July 12, 2016** | vol. 113 | no. 28

- Common cluster-based methods perform badly at low cluster-forming thresholds
- Nonparametric methods are preferred



Eklund et al., 2016

Threats to reproducibility: Methodological flexibility

- Using standard FSL analysis options
  - 69,120 possible analysis workflows

Poldrack et al, 2017, NRN

Processing step	Reason	Options	Number of plausible options
Motion correction	Correct for head motion during scanning	Interpolation [linear vs. sinc] Reference volume [single vs. mean]	4
Slice timing correction	Correct for differences in acquisition timing of different slices	No/before motion correction/after motion correction	3
Field map correction	Correct for distortion due to magnetic susceptibility	Yes/No	2
Spatial smoothing	Increase SNR for larger activations and ensure assumptions of Gaussian random field theory	FWHM [4/6/8 mm]	3
Spatial normalization	Warp individual brain to match a group template	Method [linear/nonlinear]	2
High pass filter	Remove low-frequency nuisance signals from data	Frequency cutoff [100, 120]	2
Head motion regressors	Remove remaining signals due to head motion via statistical model	Yes/No If Yes: 6/12/24 parameters or single timepoint “scrubbing” regressors	5
Hemodynamic response	Account for delayed nature of hemodynamic response to neuronal activity	Basis function [single-gamma, double-gamma] Derivatives [none/shift/dispersion]	6
Temporal autocorrelation model	Model for the temporal autocorrelation inherent in fMRI signals.	Yes/no	2
Multiple comparison correction	Correct for large number of comparisons across the brain	Voxel-based GRF, Cluster-based GRF, FDR, nonparameteric	4
Total possible workflows			**69,120**

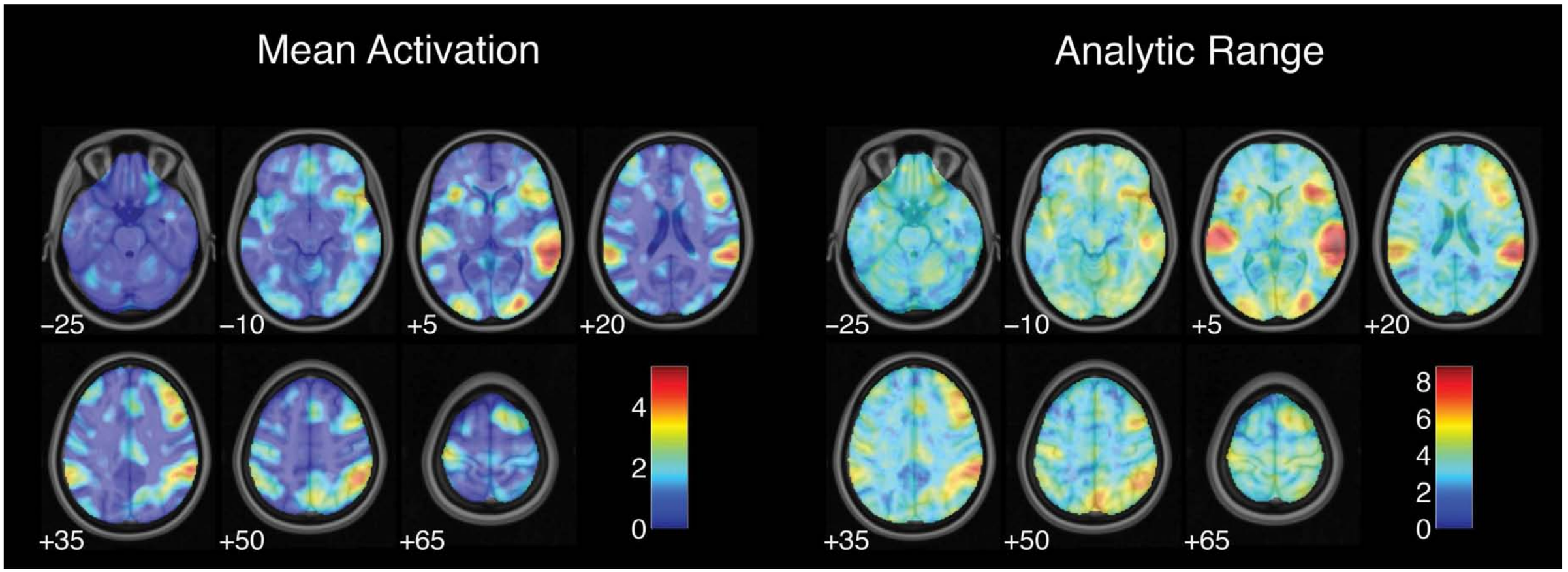
frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 October 2012
doi: 10.3389/fnins.2012.00149

On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments

Joshua Carp*

6,912 pipelines



P-hacking: Anything can become significant

Study 2: musical contrast and chronological rejuvenation

…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), $F(1, 17) = 4.92$, $p = .040$.

-Simmons et al., 2011, Psychological Science

P-hacking: Anything can become significant

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
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%

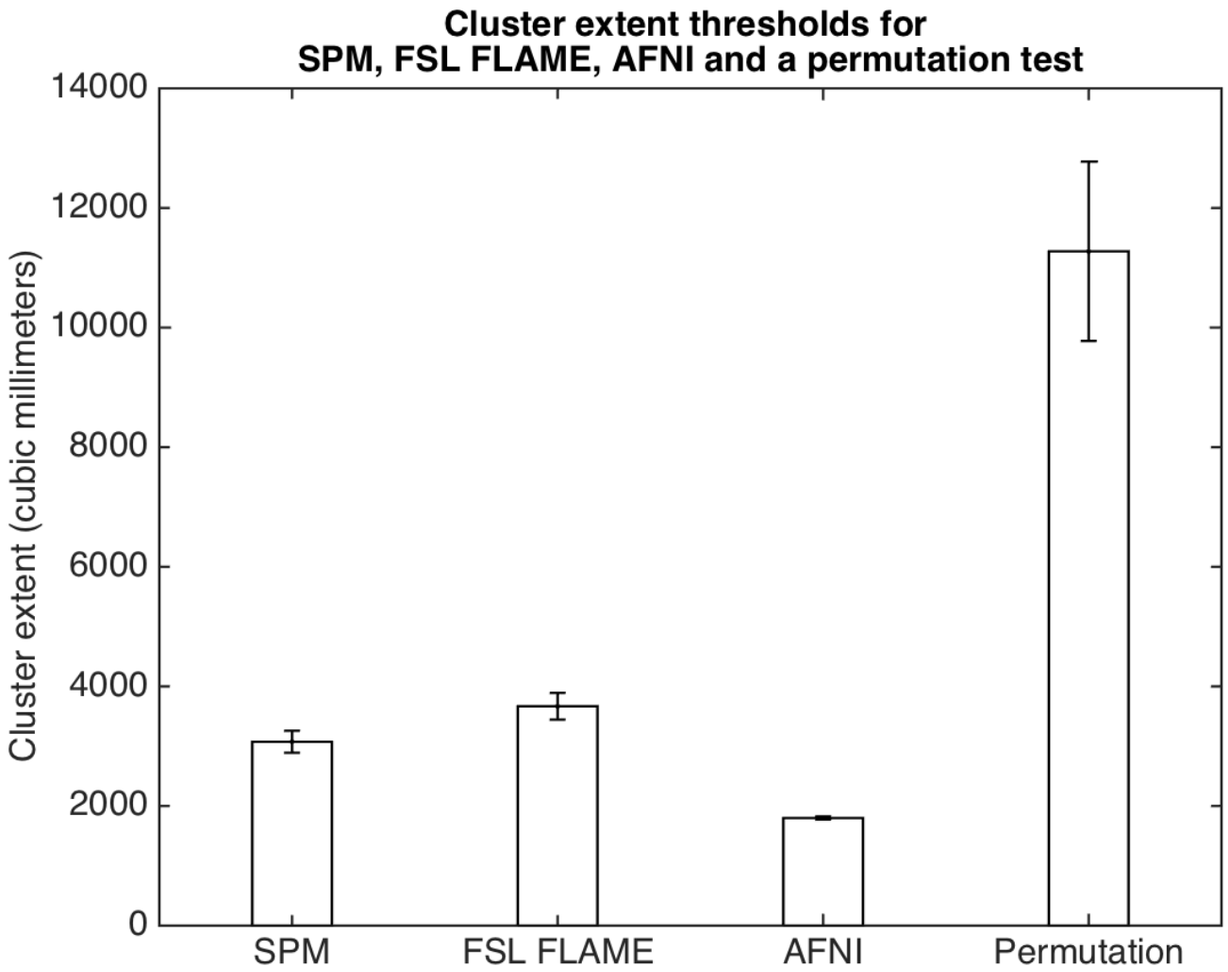
-Simmons et al., 2011, Psychological Science

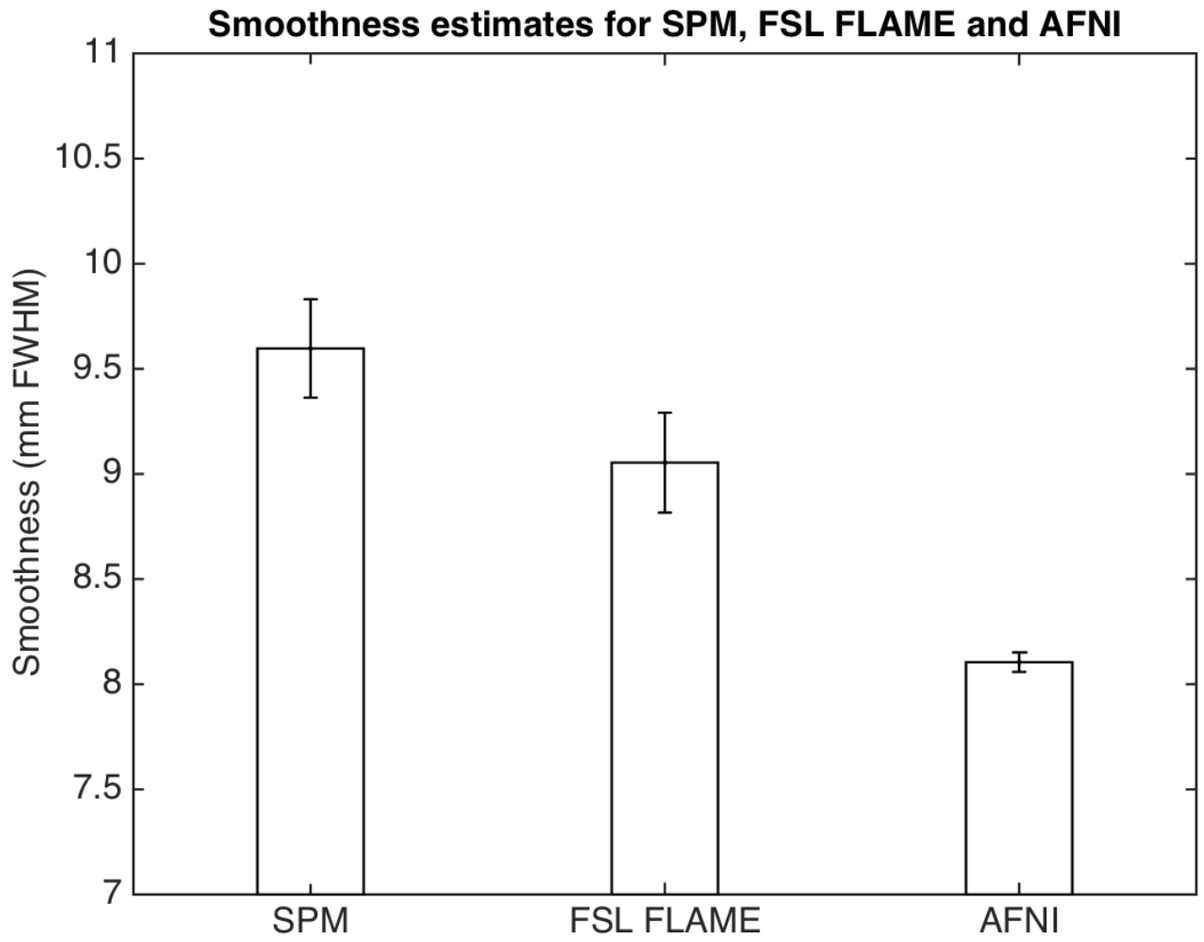
Multiple comparison correction

- Assessed latest 100 papers matching query for fMRI activation studies
  - 65 reported whole-brain activation data
  - Good news
    - only 3 papers reported uncorrected results
  - Bad news
    - 11% of papers analyzed data using SPM/FSL but then corrected for multiple comparisons using AFNI's alphasim/3dclustsim
  - Why is this a problem?

Poldrack et al, 2017, NRN

P-hacking multiple comparison correction?





AFNI AlphaSim gave more significant activation

- Eklund et al., 2016, PNAS

PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 146–157. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12639

How to get statistically significant effects in any ERP experiment (and why you shouldn't)

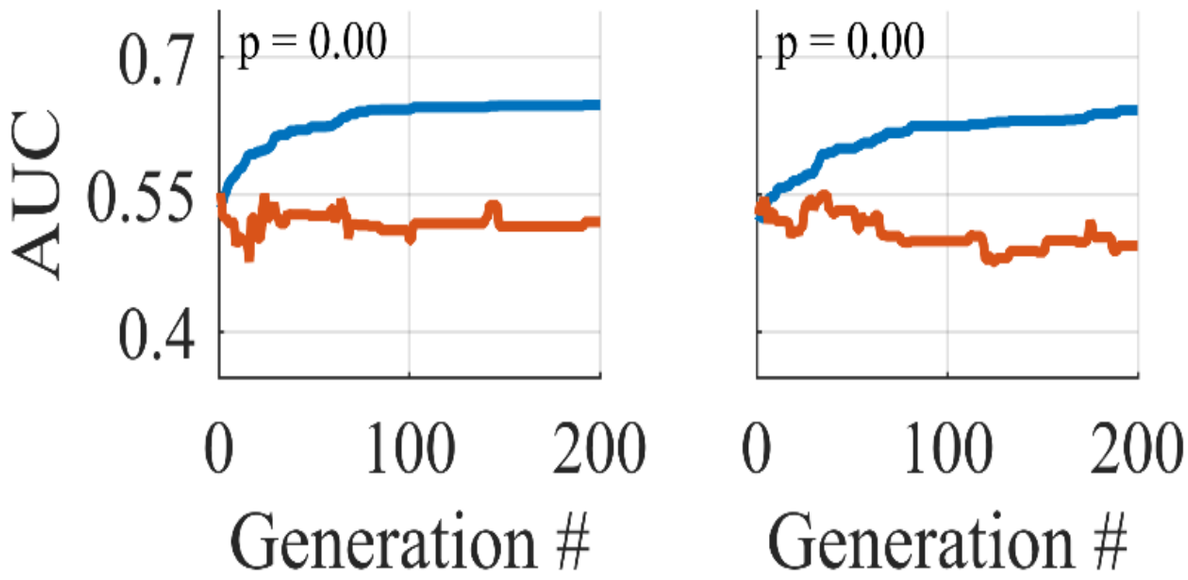
STEVEN J. LUCK[a,b] AND NICHOLAS GASPELIN[a]

The purpose of this paper is to demonstrate how common and seemingly innocuous methods for quantifying and analyzing ERP effects can lead to very high rates of significant but bogus effects, with the likelihood of obtaining at least one such bogus effect exceeding 50% in many experiments.

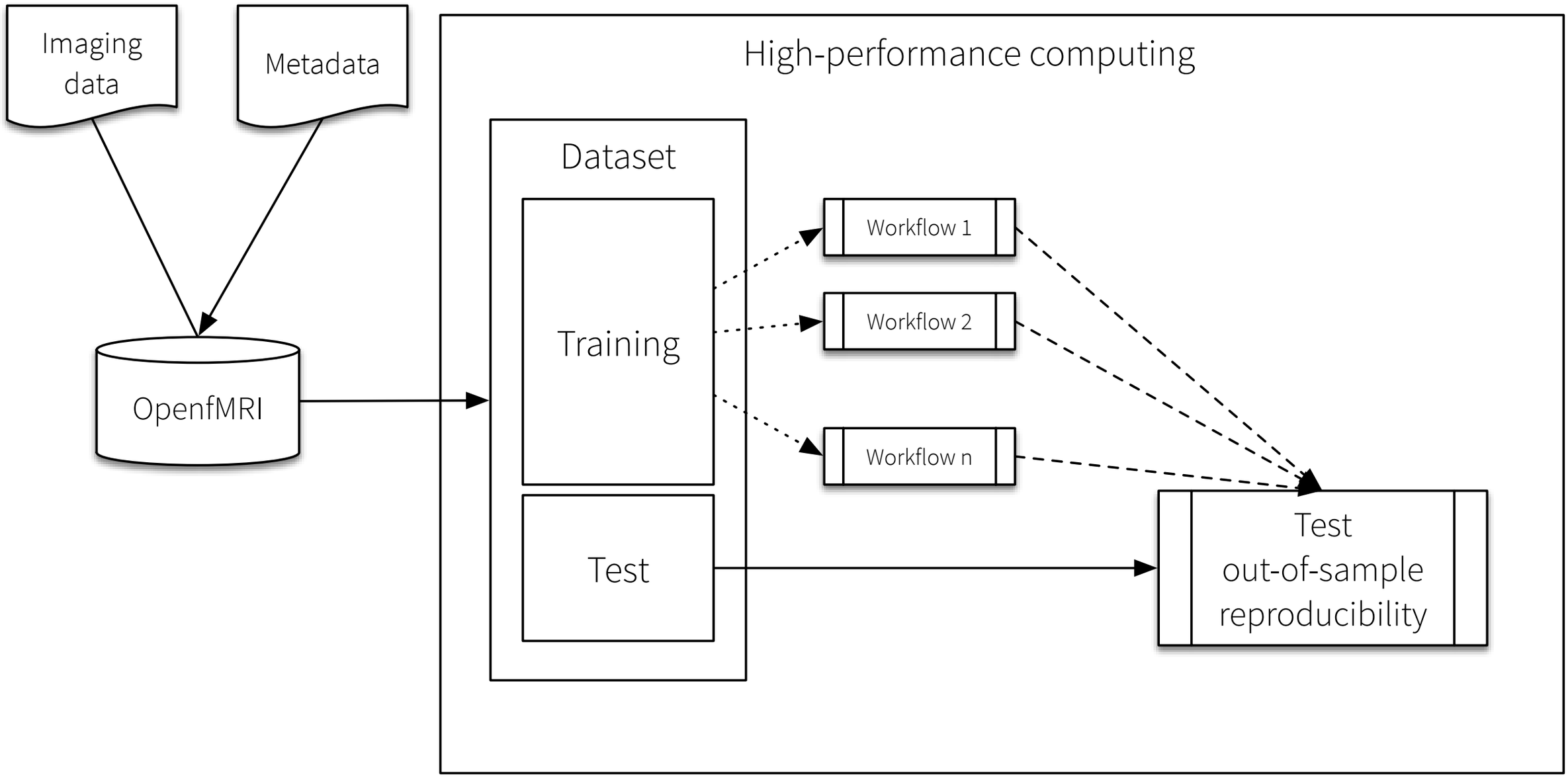
I TRIED A BUNCH OF THINGS: THE DANGERS OF UNEXPECTED OVERFITTING IN CLASSIFICATION

MICHAEL SKOCIK[1], JOHN COLLINS[2], CHLOE CALLAHAN-FLINTOFT[3],HOWARD BOWMAN[4], AND BRAD WYBLE[3]

In this article, we use Support Vector Machine (SVM) classifiers, and genetic algorithms to demonstrate the ease by which overfitting can occur, despite the use of cross validation. We demonstrate that comparable and non-generalizable results can be obtained on informative and non-informative (i.e. random) data by iteratively modifying hyperparameters seemingly innocuous ways.



Improvement: Quantifying "vibration of effects"

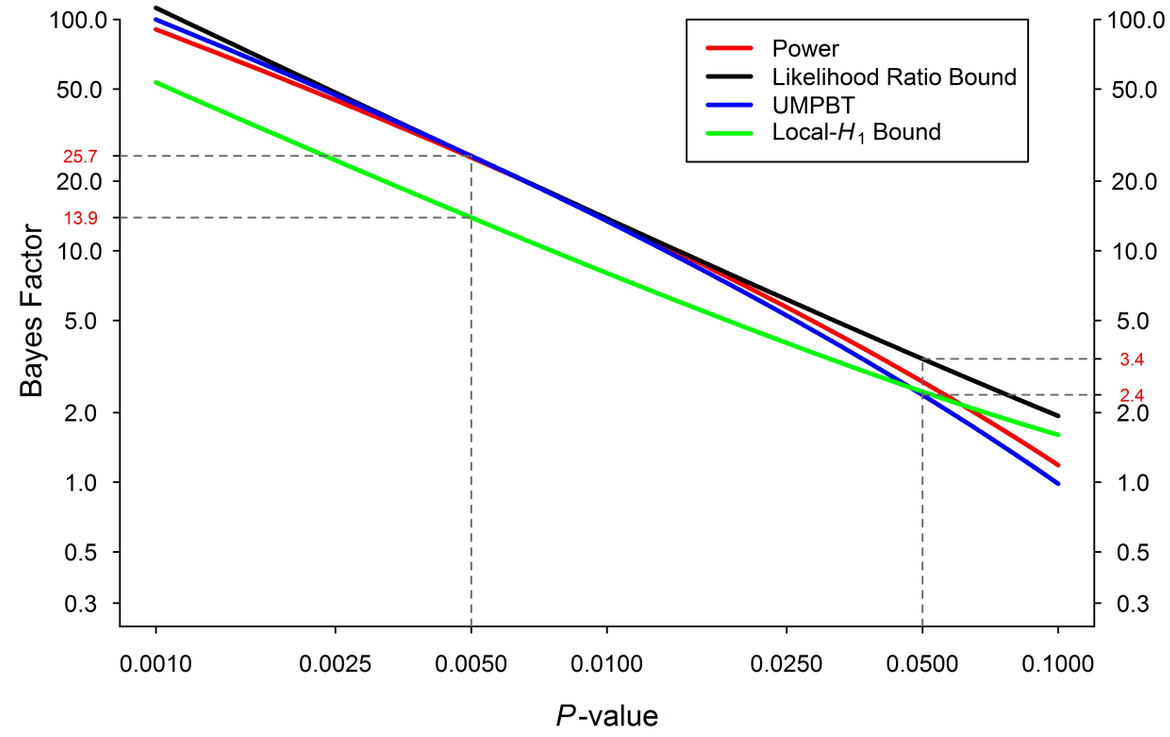


Focusing on finding generalizable results, rather than hitting the $p<0.05$ jackpot

Improvement: Increased stringency

Title: Redefine Statistical Significance

Authors: Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7, 10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12, 13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], James Holland Jones[39,40], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P.A. Ioannidis[37], Minjeong Jeon[38], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don Moore[49], Stephen L. Morgan[50], Marcus Munafò[51, 52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71], Valen E. Johnson[72]*



One Sentence Summary: We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005

Yes, this will require larger sample sizes to maintain sufficient power!

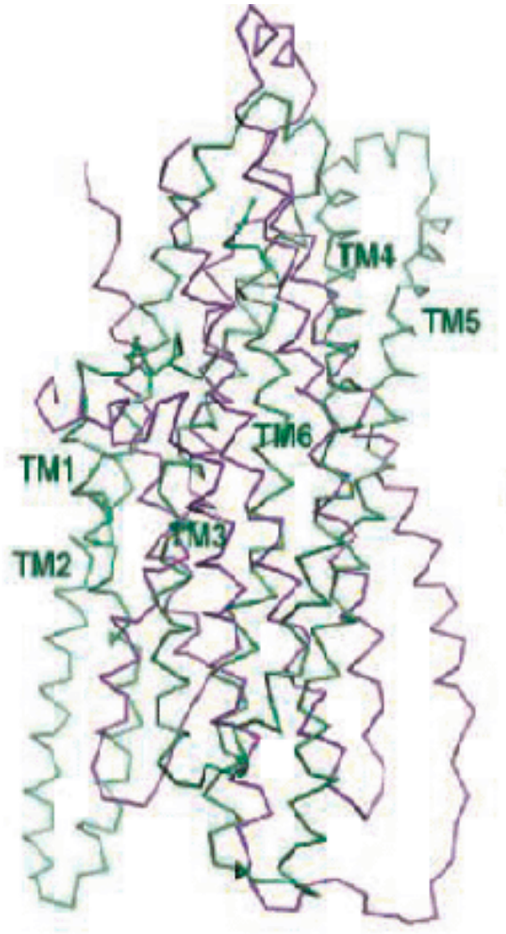
- How many of you have written computer code in the course of your research?

- How many of you have been trained in software engineering?

- How many of you have ever written a test for your code?

Threats to reproducibilty: software errors

Geoffrey Chang



Structure of MsbA from *E. coli*: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters

Geoffrey Chang* and Christopher B. Roth

Multidrug resistance (MDR) is a serious medical problem and presents a major challenge to the treatment of disease and the development of novel therapeutics. ABC transporters that are associated with multidrug resistance (MDR-ABC transporters) translocate hydrophobic drugs and lipids from the inner to the outer leaflet of the cell membrane. To better elucidate the structural basis for the "flip-flop" mechanism of substrate movement across the lipid bilayer, we have determined the structure of the lipid flippase MsbA from *Escherichia coli* by x-ray crystallography to a resolution of 4.5 angstroms. MsbA is organized as a homodimer with each subunit containing six transmembrane α-helices and a nucleotide-binding domain. The asymmetric distribution of charged residues lining a central chamber suggests a general mechanism for the translocation of substrate by MsbA and other MDR-ABC transporters. The structure of MsbA can serve as a model for the MDR-ABC transporters that confer multidrug resistance to cancer cells and infectious microorganisms.

Structure of the ABC Transporter MsbA in Complex with ADP·Vanadate and Lipopolysaccharide

Christopher L. Reyes and Geoffrey Chang*

Select members of the adenosine triphosphate (ATP)–binding cassette (ABC) transporter family couple ATP binding and hydrolysis to substrate efflux and confer multidrug resistance. We have determined the x-ray structure of MsbA in complex with magnesium, adenosine diphosphate, and inorganic vanadate ($Mg{\cdot}ADP{\cdot}V_i$) and the rough-chemotype lipopolysaccharide, Ra LPS. The structure supports a model involving a rigid-body torque of the two transmembrane domains during ATP hydrolysis and suggests a mechanism by which the nucleotide-binding domain communicates with the transmembrane domain. We propose a lipid "flip-flop" mechanism in which the sugar groups are sequestered in the chamber while the hydrophobic tails are dragged through the lipid bilayer.

X-ray Structure of the EmrE Multidrug Transporter in Complex with a Substrate

Owen Pornillos, Yen-Ju Chen, Andy P. Chen, Geoffrey Chang*

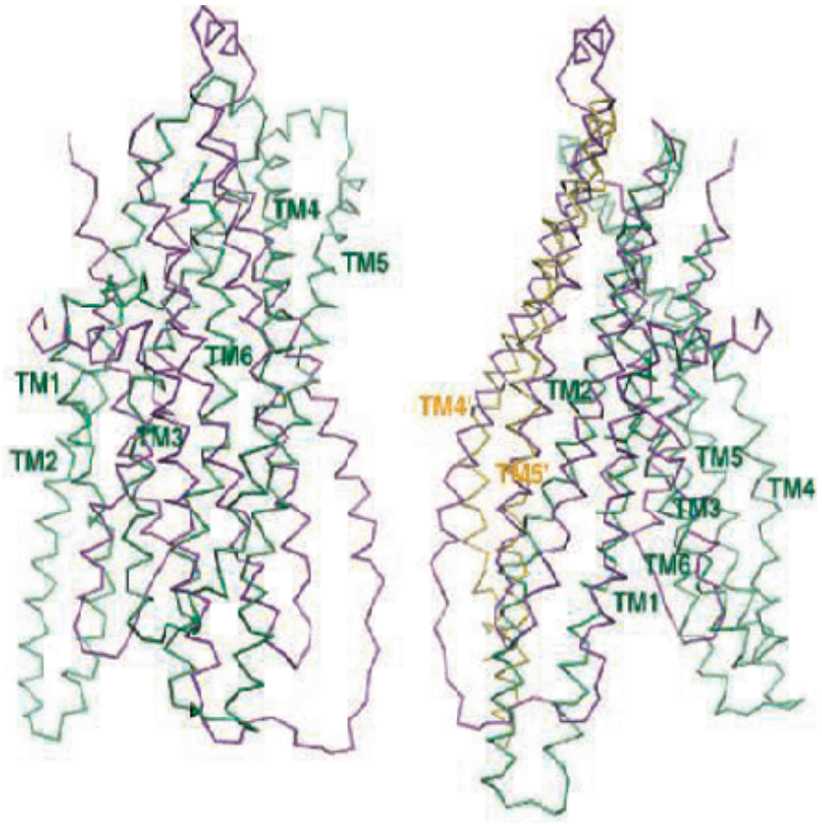
EmrE is a prototype of the Small Multidrug Resistance family of efflux transporters and actively expels positively charged hydrophobic drugs across the inner membrane of *Escherichia coli*. Here, we report the x-ray crystal structure, at 3.7 angstrom resolution, of one conformational state of the EmrE transporter in complex with a translocation substrate, tetraphenylphosphonium. Two EmrE polypeptides form a homodimeric transporter that binds substrate at the dimerization interface. The two subunits have opposite orientations in the membrane and adopt slightly different folds, forming an asymmetric antiparallel dimer. This unusual architecture likely confers unidirectionality to transport by creating an asymmetric substrate translocation pathway. On the basis of available structural data, we propose a model for the proton-dependent drug efflux mechanism of EmrE.

Retraction

WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP•vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.



Small errors can have big effects

```
# 23-class classification problem

skf=StratifiedKFold(labels,8)

if trainsvm:
    pred=N.zeros(len(labels))
    for train,test in skf:
        clf=LinearSVC()
        clf.fit(data[train],labels[train])          data[:,train]
        pred[test]=clf.predict(data[test])          data[:,test]
```

Results:
93% accuracy

Results:
53% accuracy

http://www.russpoldrack.org/2013/02/anatomy-of-coding-error.html

Bug-hacking

- Bugs that confirm our predictions are less likely to be uncovered than bugs that disconfirm them

Improvement: The principle of assumed error

- Whenever you find a seemingly good result (e.g. one that fits your predictions), assume that it occurred due to an error in your code
- Helps protect from bug hacking

Improvement: Software testing and validation

- Smoke tests and unit tests may be useful but are not sufficient
- For complex analyses:
  - Parameter recovery: Generate data for which the true answer is known, and assess ability of code to recover the correct answer
  - Randomization: Generate data for which the null hypothesis of no relationship should be true on average, and ensure that the observed false positive rate is accurate (cf. Eklund et al., 2016, PNAS)

http://www.russpoldrack.org/2016/08/the-principle-of-assumed-error.html

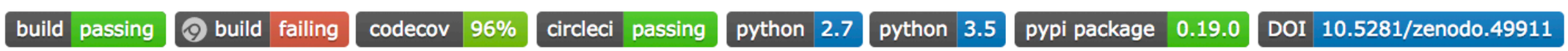
WWW.SERENAATKINS.COM

Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says **Nick Barnes**.

Improvement: Use established libraries when possible

- Avoid the NIH ("not invented here") effect
  - rejecting existing solutions in favor of home-grown ones
  - "I need to write a new DICOM to Nifti converter"
- Prefer libraries that use good software engineering practices



scikit-learn

scikit-learn is a Python module for machine learning built on top of SciPy and distributed under the 3-Clause BSD license.

Study reporting and transparency

- In 22 of the 65 papers we analyzed for multiple comparison procedures, it was impossible to identify precisely which correction technique was used
- beyond generic terms such as “cluster based correction”

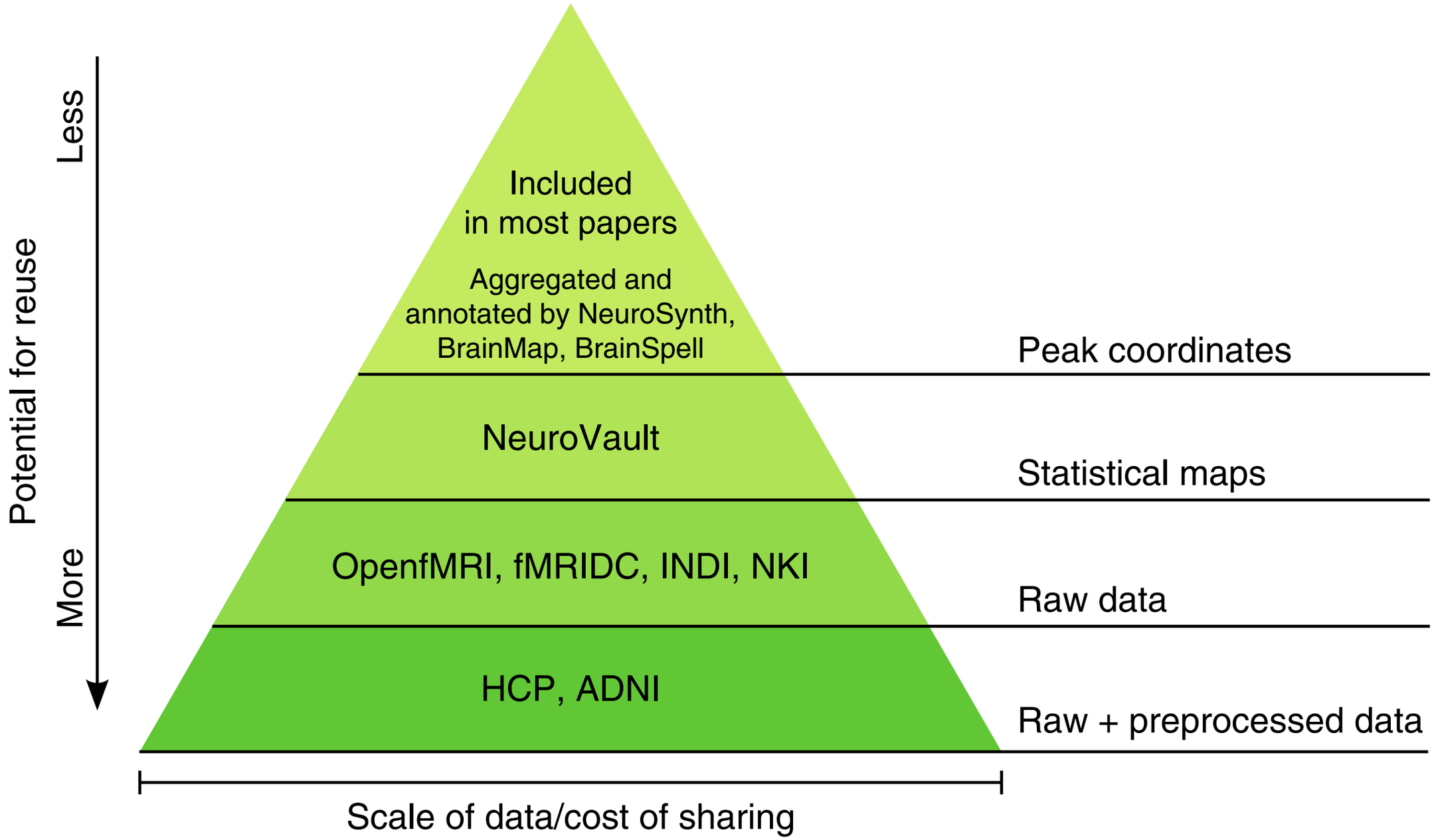
Best practices in data analysis and sharing in neuroimaging using MRI

Thomas E Nichols[1], Samir Das[2,3], Simon B Eickhoff[4,5], Alan C Evans[2,3], Tristan Glatard[2,6], Michael Hanke[7,8], Nikolaus Kriegeskorte[9], Michael P Milham[10,11], Russell A Poldrack[12], Jean-Baptiste Poline[13], Erika Proal[14], Bertrand Thirion[15], David C Van Essen[16], Tonya White[17] & B T Thomas Yeo[18]

NATURE NEUROSCIENCE VOLUME 20 | NUMBER 3 | MARCH 2017

www.humanbrainmapping.org/cobidas/

Improvement: Data Sharing

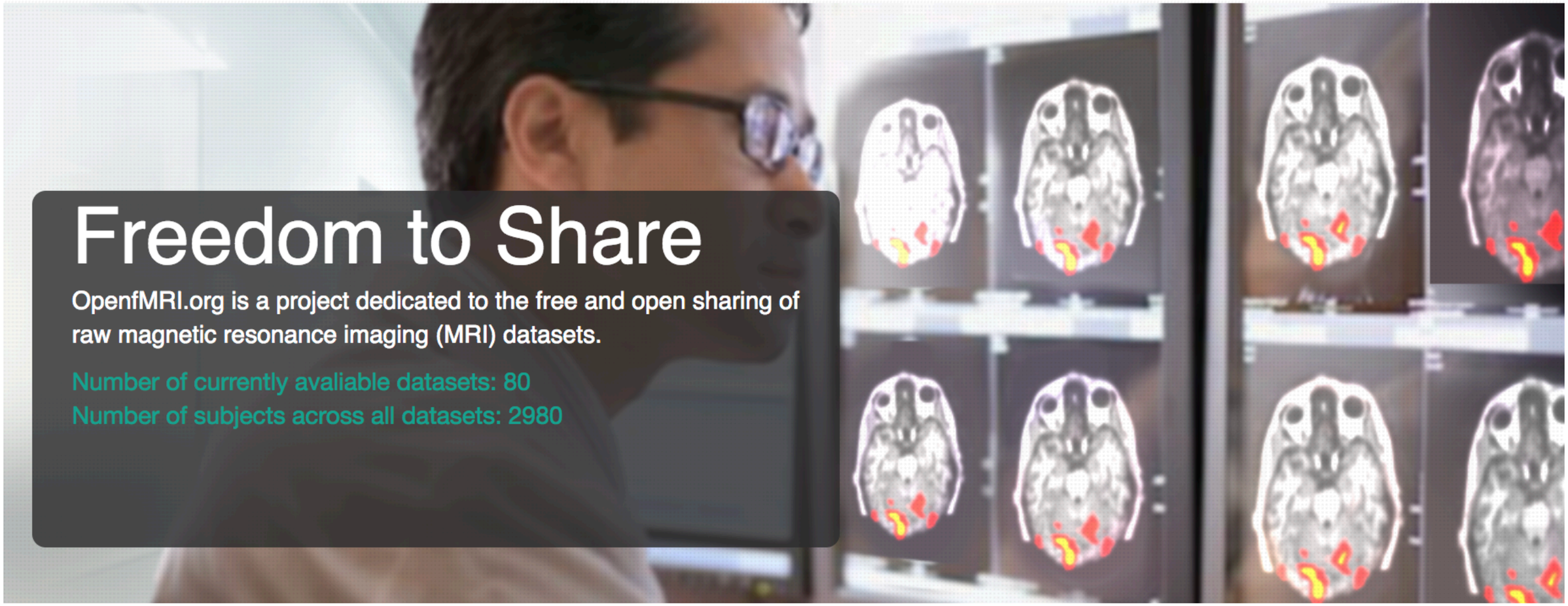


Poldrack & Gorgolewski, 2014

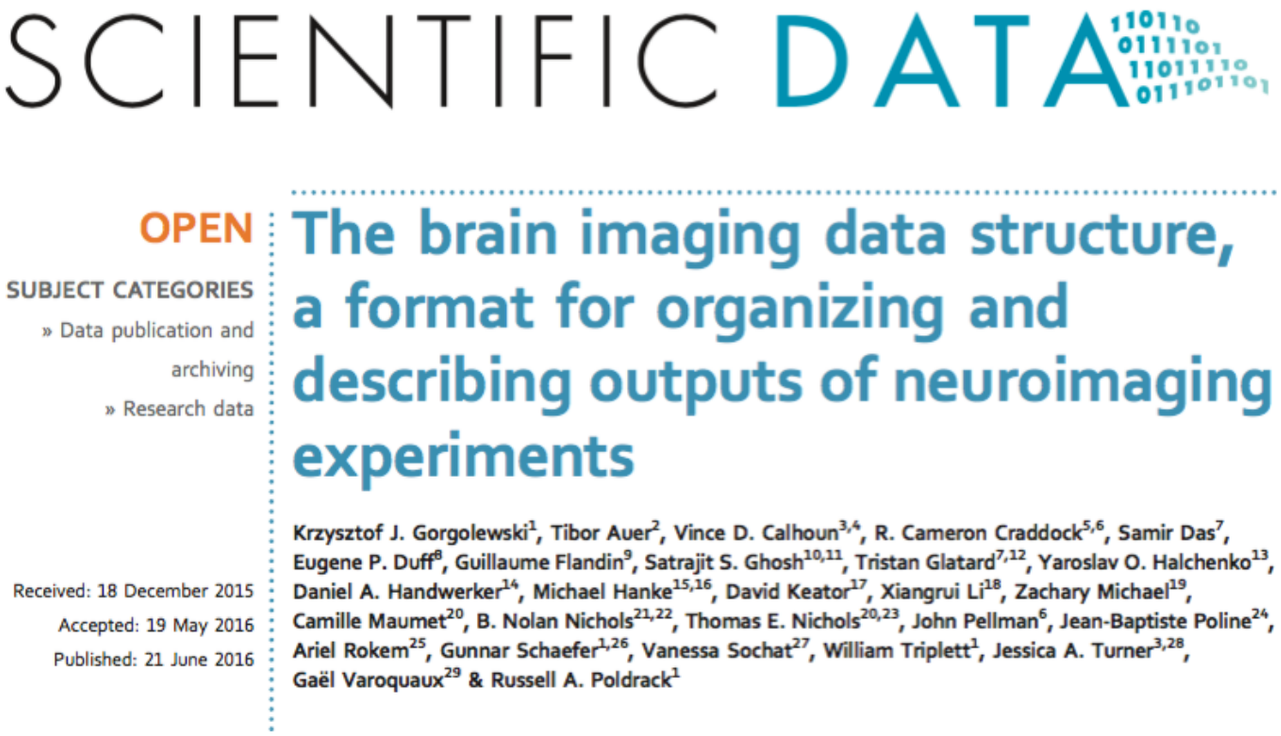
OpenfMRI: Sharing complete raw datasets

OpenfMRI

View Datasets FAQs Submit a new Dataset Login



Brain Imaging Data Structure (BIDS)



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Data publication and archiving
» Research data

The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

Krzysztof J. Gorgolewski[1], Tibor Auer[2], Vince D. Calhoun[3,4], R. Cameron Craddock[5,6], Samir Das[7], Eugene P. Duff[8], Guillaume Flandin[9], Satrajit S. Ghosh[10,11], Tristan Glatard[7,12], Yaroslav O. Halchenko[13], Daniel A. Handwerker[14], Michael Hanke[15,16], David Keator[17], Xiangrui Li[18], Zachary Michael[19], Camille Maumet[20], B. Nolan Nichols[21,22], Thomas E. Nichols[20,23], John Pellman[6], Jean-Baptiste Poline[24], Ariel Rokem[25], Gunnar Schaefer[1,26], Vanessa Sochat[27], William Triplett[1], Jessica A. Turner[3,28], Gaël Varoquaux[29] & Russell A. Poldrack[1]

Received: 18 December 2015
Accepted: 19 May 2016
Published: 21 June 2016

- sub-control01
  - anat
    - sub-control01_T1w.nii.gz
    - sub-control01_T1w.json
    - sub-control01_T2w.nii.gz
    - sub-control01_T2w.json
  - func
    - sub-control01_task-nback_bold.nii.gz
    - sub-control01_task-nback_bold.json
    - sub-control01_task-nback_events.tsv
    - sub-control01_task-nback_physio.tsv.gz
    - sub-control01_task-nback_physio.json
    - sub-control01_task-nback_sbref.nii.gz
  - dwi
    - sub-control01_dwi.nii.gz

http://bids.neuroimaging.io

Neurovault: Sharing statistical maps

NeuroVault Collections FAQ Give feedback Log in

NEUROVAULT

A public repository of unthresholded statistical maps, parcellations, and atlases of the human brain

What is it?

A place where researchers can publicly store and share unthresholded statistical maps, parcellations, and atlases produced by MRI and PET studies.

Why use it?

- Interactive visualization
- A permanent URL
- Publicly shareable
- Improves meta-analyses

Get started and upload an image!

Supported by

incf Stanford University

Gorgolewski et al., 2015

Collections

A collection is a set of images grouped together for some sensible reason.

Most commonly, a collection contains all of the images from a single study.

Show 25 entries Search: retest

Name	Num. images	Description	Has DOI?
A test-retest fMRI dataset for motor, language and spatial attention functions	11	OpenfMRI ds000114	Yes
Brainpedia - A test-retest fMRI dataset for motor, language and spatial attention functions	11	OpenfMRI ds114	
Classification learning and stop-signal (1 year test-retest)	26	OpenfMRI ds000017 A group of eight subjects performed two tasks (selective stop-signal and probabilistic classification) on two different occasions separated by about one year. ds000017A reflects data from timepoint 1 and ds000017B reflects data from timepoint 2.	
Long-term test–retest reliability of functional MRI in a classification learning task	6	OpenfMRI ds000002	Yes
Validating the Why/How contrast for functional MRI studies of Theory of Mind	5	The ability to impute mental states to others, or Theory of Mind (ToM), has been the subject of hundreds of neuroimaging studies. Although reviews and meta-analyses of these studies have concluded that ToM recruits a coherent brain network, mounting evidence suggests that this network is an abstraction based on pooling data from numerous studies, most of which use different behavioral tasks to investigate ToM. Problematically, this means that no single behavioral task can be used to reliably measure ToM Network function as currently conceived. To make ToM Network function scientifically tractable, we need standardized tasks capable of reliably measuring specific aspects of its functioning. Here, our goal is to validate the Why/How Task for this purpose. Several prior studies have found that when compared to answering how-questions about another person's behavior, answering why-questions about that same behavior activates a network that is anatomically consistent with	Yes

A test-retest fMRI dataset for motor, language and spatial attention functions

Contributed by ChrisFiloGorgolewski

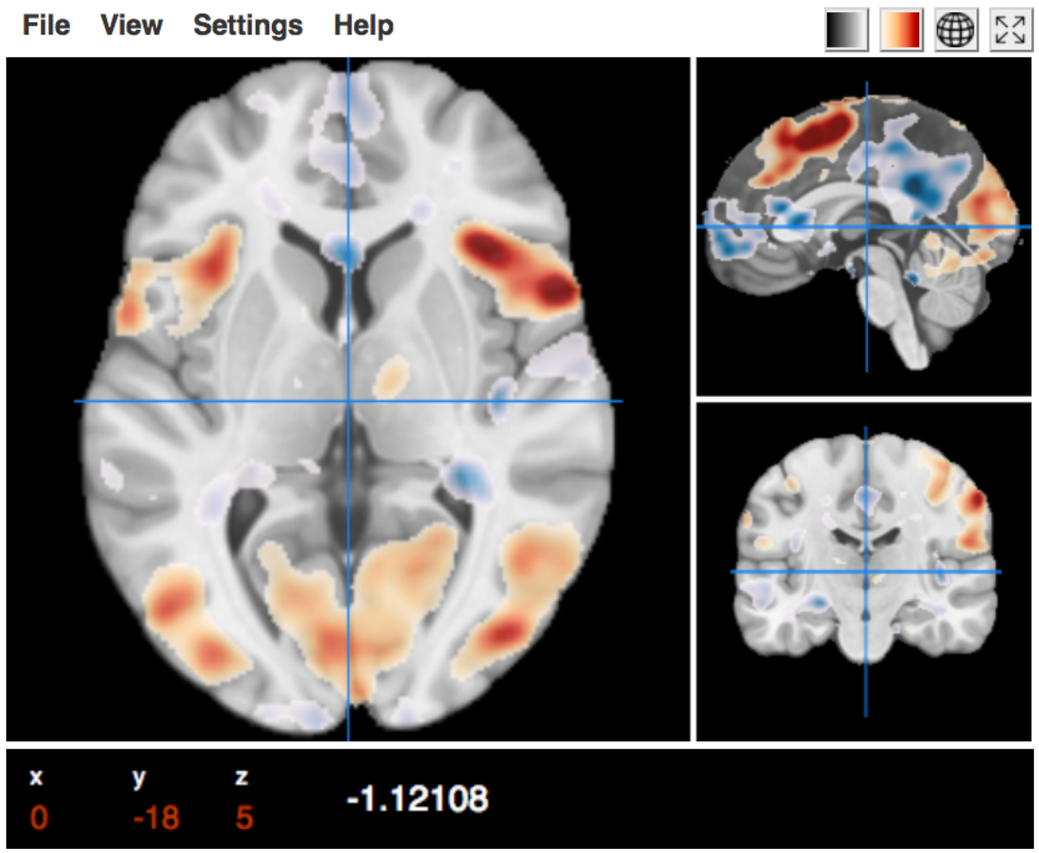
Krzysztof J Gorgolewski, Amos Storkey, Mark E Bastin, Ian R Whittle, Joanna M Wardlaw and Cyril R Pernet

Corresponding publication

Source data

3D View Download

File View Settings Help



Images Details

Show 7 entries Search:

View	ID	Name	Type
	299	landmark incorrect task	T map
	300	landmark correct task	T map
	301	motor lips	T map
	303	landmark no response task	T map
	304	overt verb generation	T map
	305	landmark no response control	T map
	307	motor finger	T map

Showing 1 to 7 of 11 entries First Previous Next Last

overt verb generation

Contributed by ChrisFiloGorgolewski on Jan. 21, 2016

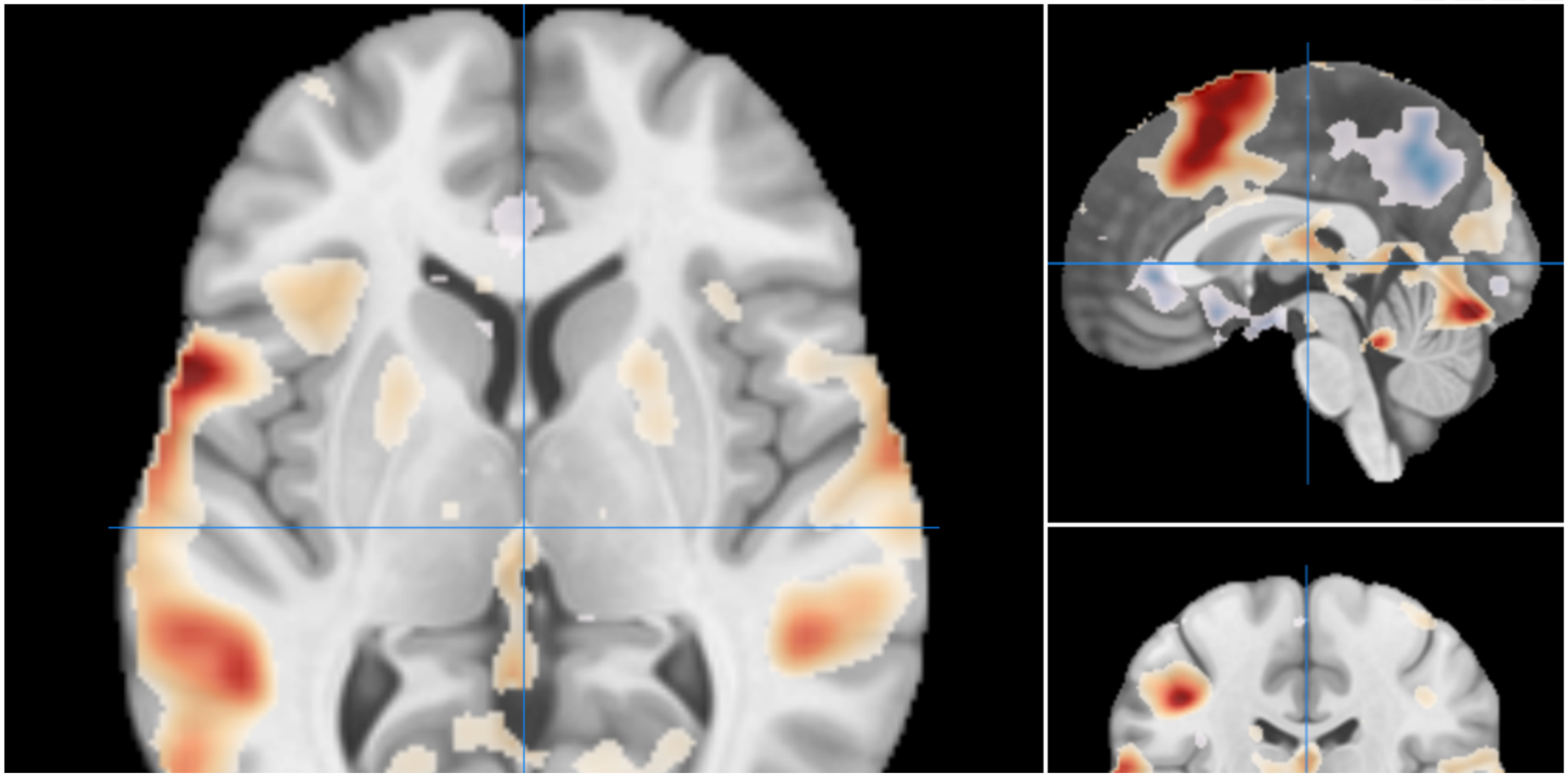
Collection: A test-retest fMRI dataset for motor, language and spatial attention functions

Description: FSL5.0

Task View | 3D View | Download | Analysis

Papaya viewer | Details | Embed

File View Settings Help



NeuroVault Collections FAQ Give feedback RussPoldrack

overt verb generation

Contributed by ChrisFiloGorgolewski on Jan. 21, 2016

Collection: A test-retest fMRI dataset for motor, language and spatial attention functions

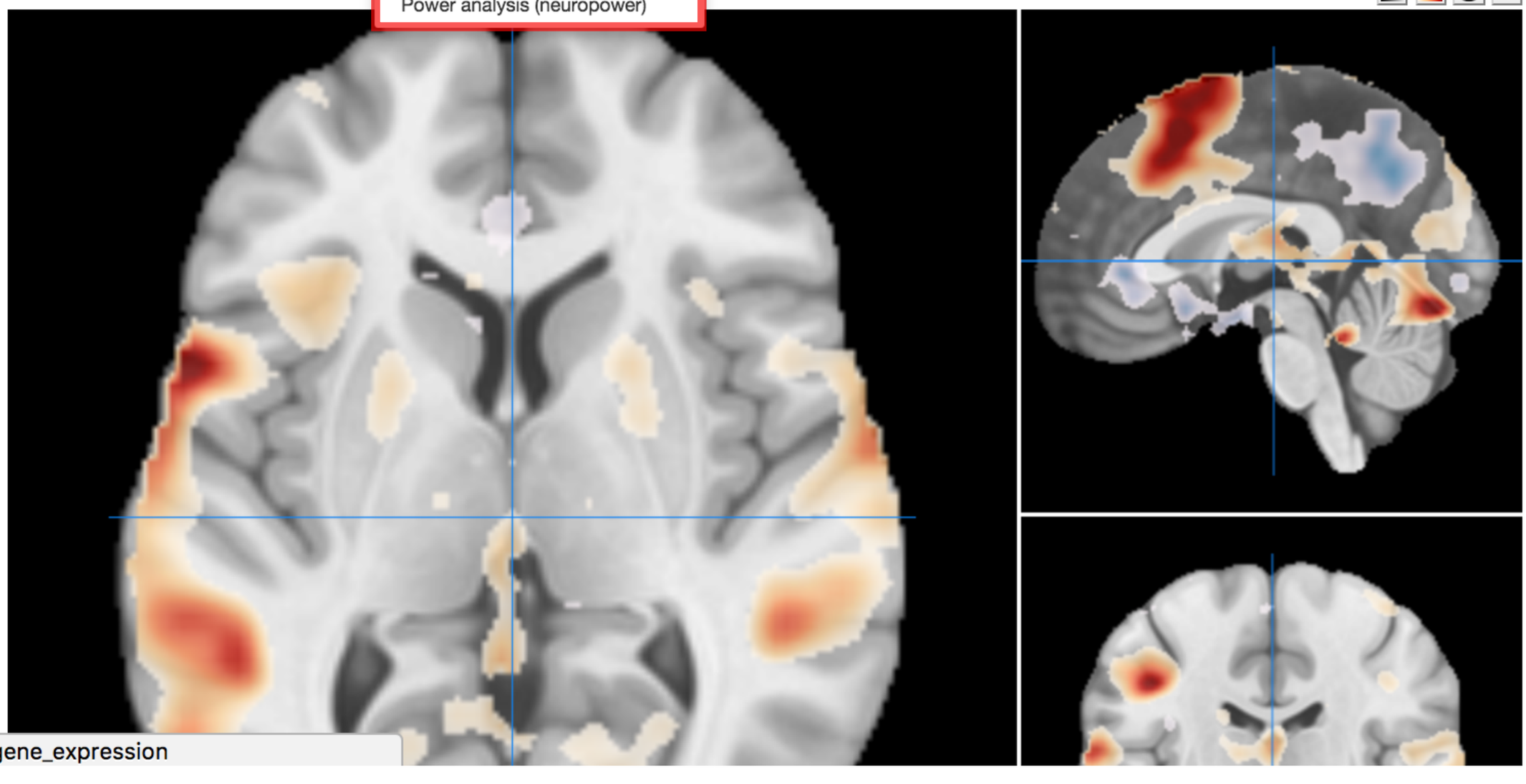
Description: FSL5.0

Task View | 3D View | Download | Analysis

- Cognitive decoding (neurosynth)
- Similar maps search
- Gene expression decoding
- Power analysis (neuropower)

Papaya viewer | Details | Embed

File View Settings Help



304/gene_expression

NeuroPowerTools NeuroPower NeuroDesign

OVERVIEW | START | PEAK TABLE | MODEL FIT | POWER CALCULATION | POWER TABLE | RESET

There is not enough power to estimate a threshold for BH.

Power

To see the power for a certain sample size or vice versa, please fill out either the minimal power or the sample size.

MCP*

Random Field Theory

Sample size

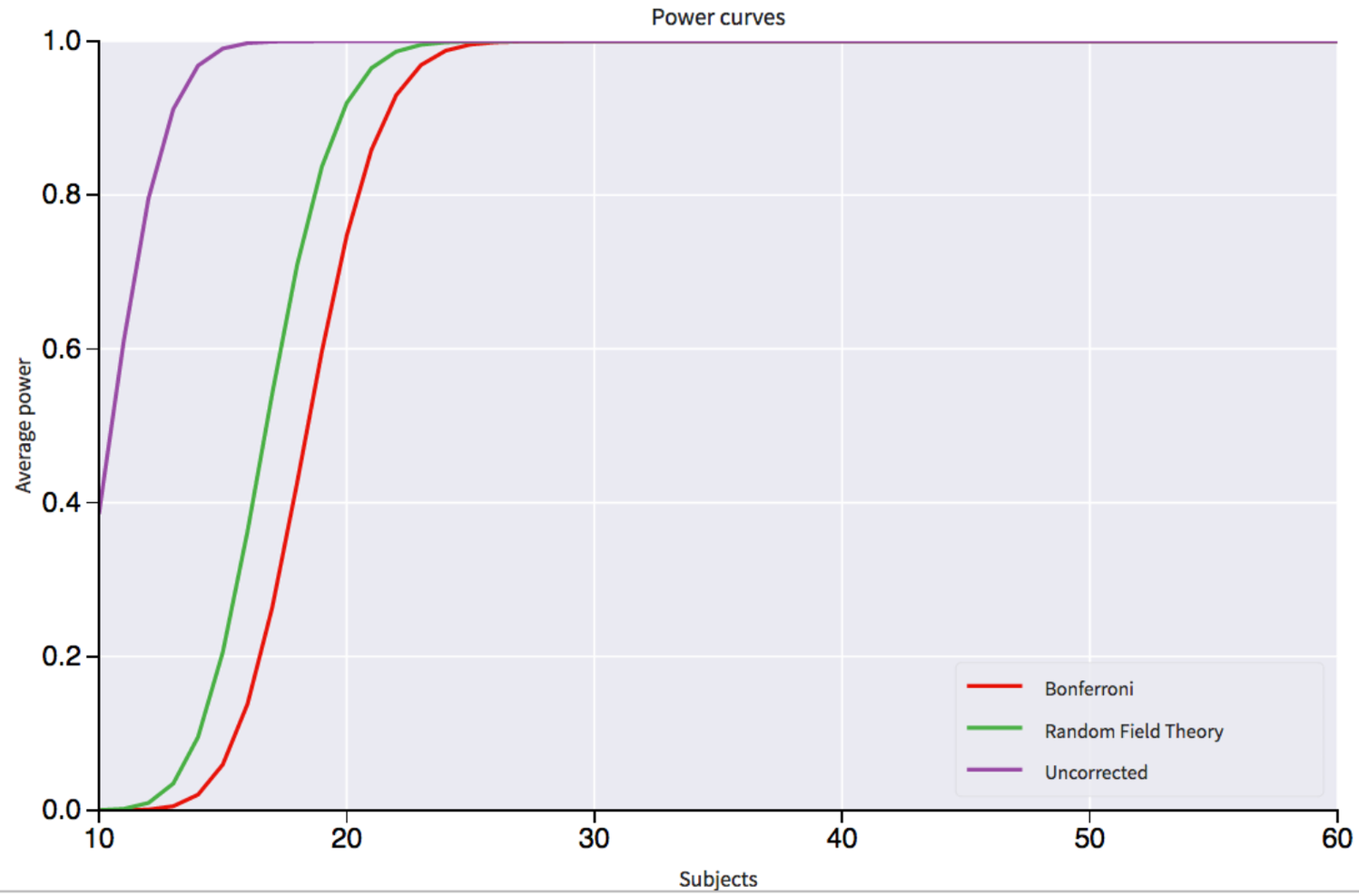
40

Power

0.0

Submit parameters

Hover over the lines to see detailed power predictions



NeuroVault Collections FAQ Give feedback RussPoldrack

overt verb generation

Contributed by ChrisFiloGorgolewski on Jan. 21, 2016

Collection: A test-retest fMRI dataset for motor, language and spatial attention functions

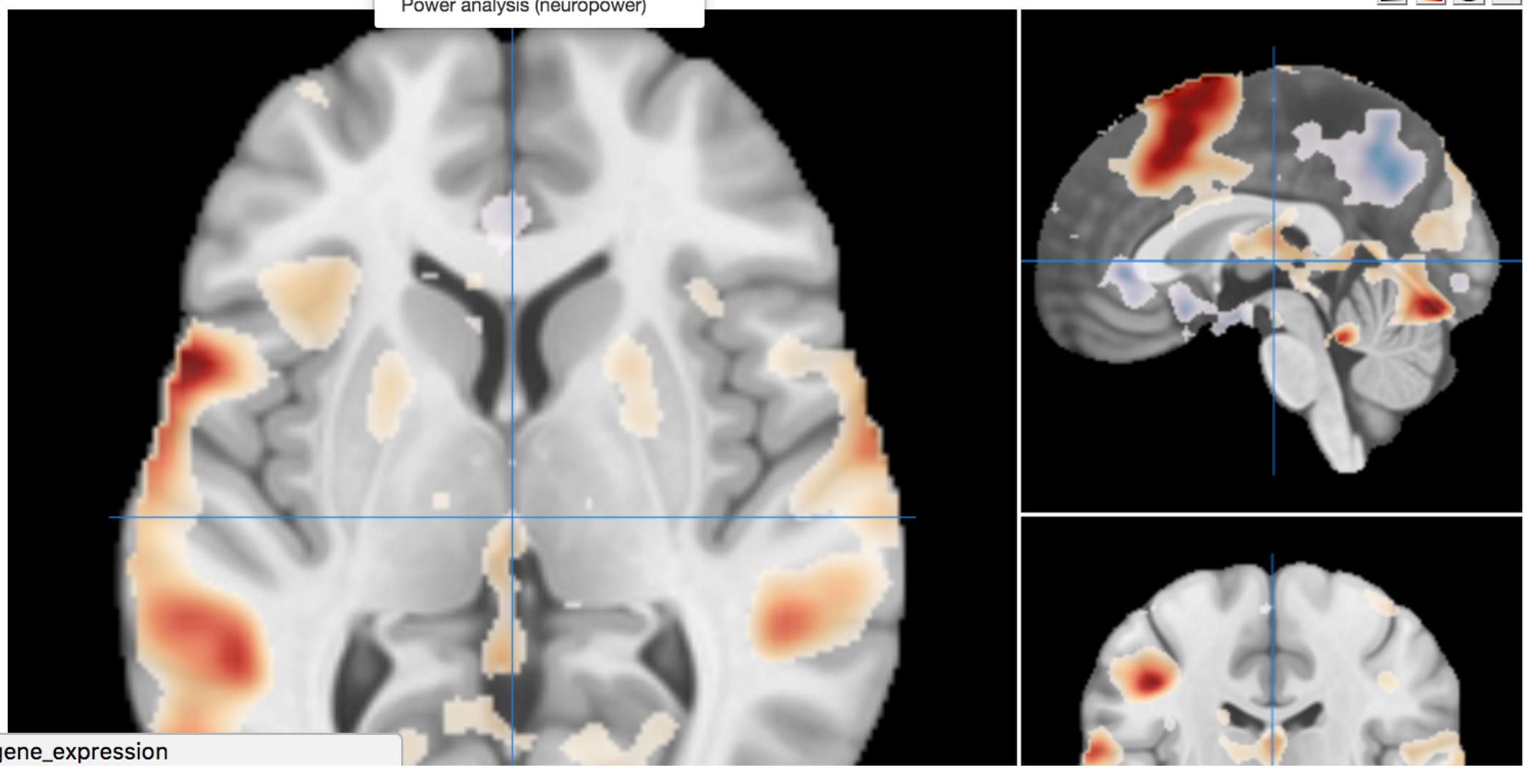
Description: FSL5.0

Task View | 3D View | Download | Analysis

- Cognitive decoding (neurosynth)
- Similar maps search
- Gene expression decoding
- Power analysis (neuropower)

Papaya viewer | Details | Embed

File View Settings Help



304/gene_expression

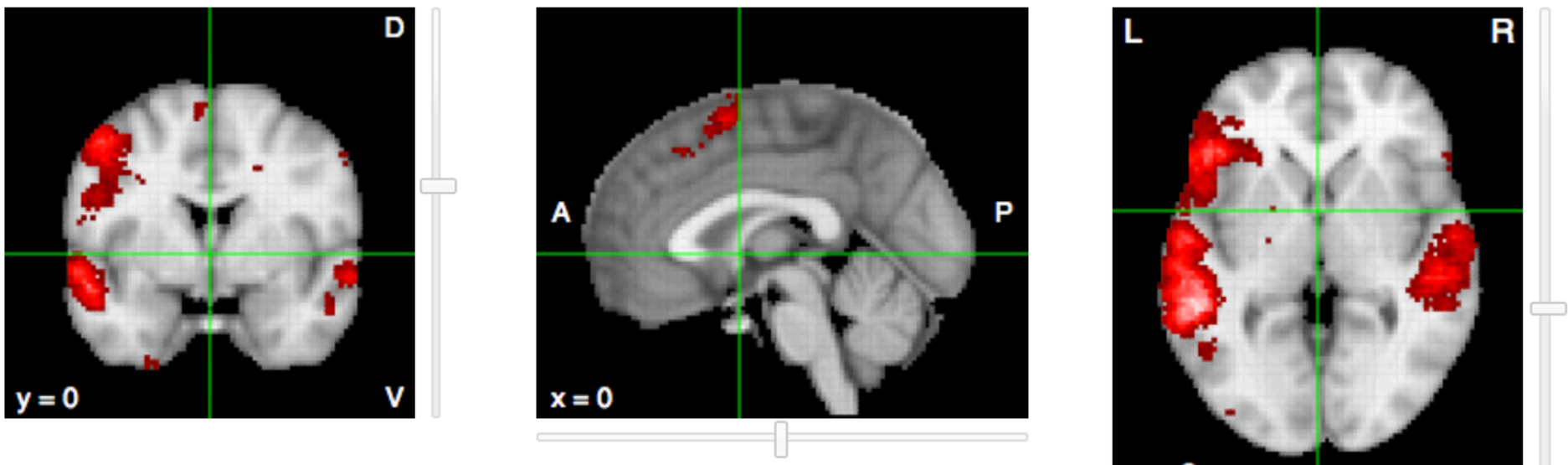
Neurosynth: Sharing coordinate-based data

Neurosynth.org (beta) Home Meta-analyses Studies Locations Decoder Code FAQs Sign in

neurosynth.org

Neurosynth is a platform for large-scale, automated synthesis of functional magnetic resonance imaging (fMRI) data.

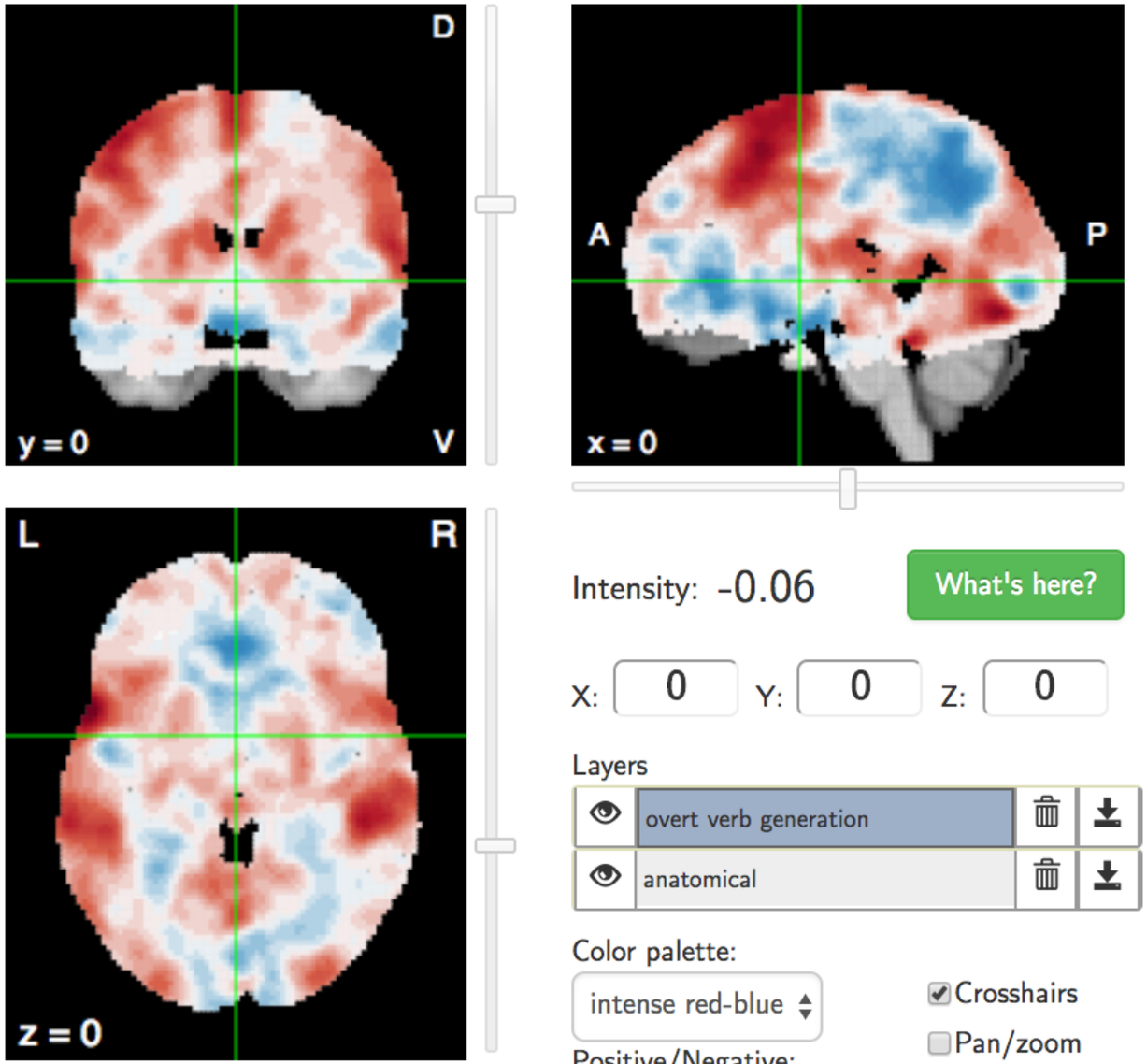
It takes thousands of published articles reporting the results of fMRI studies, chews on them for a bit, and then spits out images that look like this:



An automated meta-analysis of 885 studies of language

Decoding results for image overt verb generation

Map Plot



Intensity: -0.06 What's here?

X: 0 Y: 0 Z: 0

Layers

overt verb generation

anatomical

Color palette:

intense red-blue

Positive/Negative:

both

Crosshairs

Pan/zoom

Labels

Thresholds:

Term similarity

To compare the decoded image against a Neurosynth term, click on an arrow below.

Copy CSV XLS

Show 10 entries

Search:

analysis	corr.
phonological	0.32
language	0.306
speech production	0.29
production	0.288
word	0.28
words	0.266
reading	0.254
speech	0.254
frontal	0.24
inferior frontal	0.236

Showing 1 to 10 of 2,911 entries

Previous Next

OPEN ACCESS Freely available online

PLoS one

Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results

Jelte M. Wicherts*, Marjan Bakker, Dylan Molenaar

Psychology Department, Faculty of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, The Netherlands

Why people don't share data

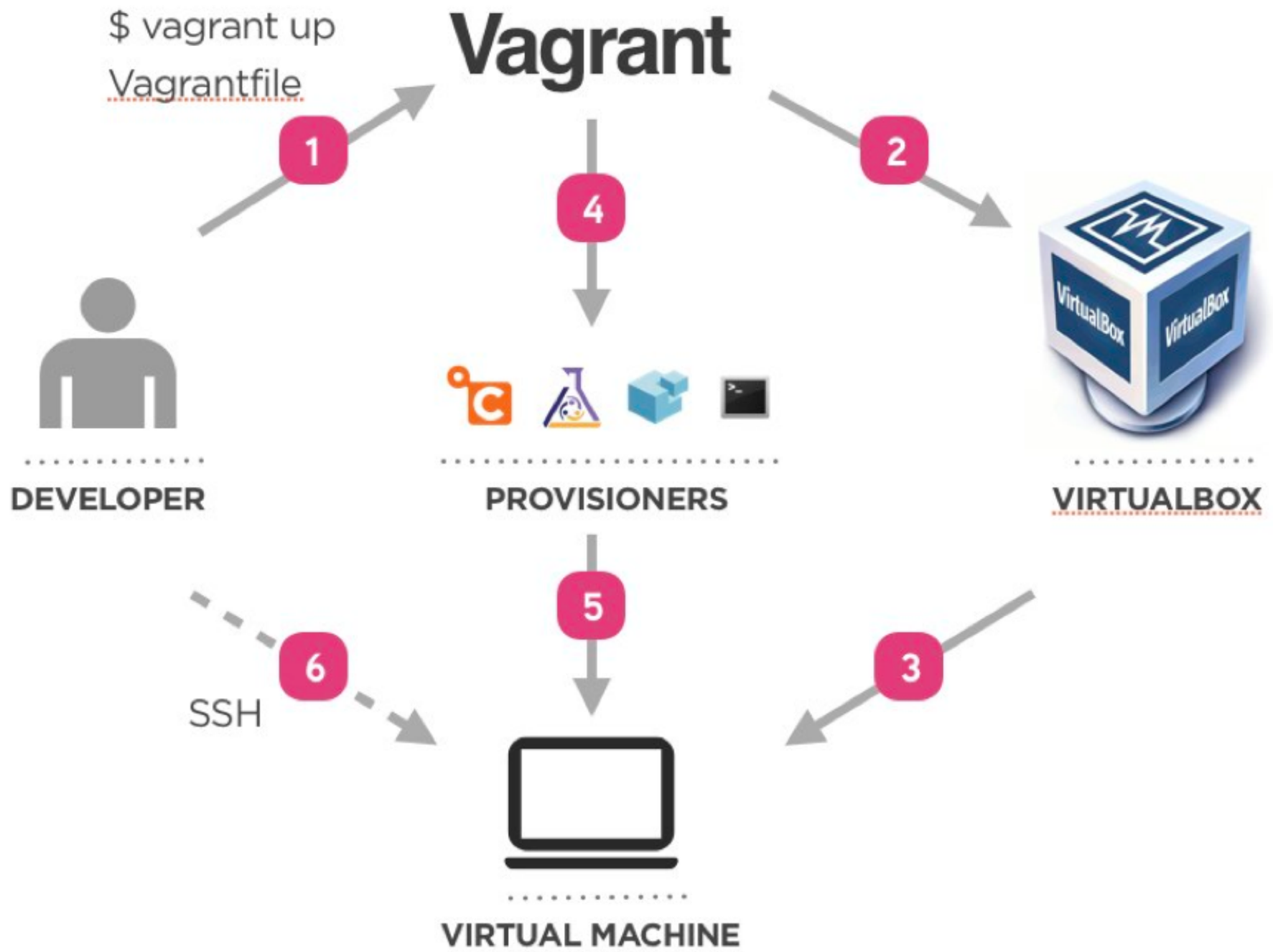
- Concern about being scooped
  - "The thing that matters the least is being scooped. The thing that matters the most is being ignored." - Gary King
- Concern about errors being discovered
  - We should *want* to know if we made an error
  - We need to be more forgiving of honest admission of errors
- Concern about the time and effort involved
  - Sharing of statistical maps is easy and fast
  - Sharing of full dataset is easier when you format using BIDS from the beginning

Improvement: Sharing of analysis platforms

- "an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." - Buckheit & Donoho, 1995
- The tale of myconnectome

Virtual machines as tools for reproducible science

VAGRANT



Google Cal | https://www | https://twitt | https://app. | Nature | Shared with | Handling Mi | poldrack/my | Russ

GitHub, Inc. [US] https://github.com/poldrack/myconnectome-vm

Apps | Google Calendar - W | SERA | Axess | PubMed | sklearn | nipype

This repository Search | Pull requests | Issues | Gist

poldrack / myconnectome-vm

Unwatch 2 | Star 0 | Fork 1

Virtual machine setup for MyConnectome data analysis — Edit

86 commits | **1** branch | **0** releases | **3** contributors

Branch: **master** | **myconnectome-vm** / +

Merge pull request #14 from vsoch/master ···

poldrack authored 23 days ago — latest commit 3b45da4ddb

LICENSE	Initial commit	2 months ago
README.md	Update README.md	24 days ago
Vagrantfile	removing supervisor controller from application - will be run with st…	23 days ago

README.md

MyConnectome-VM: A virtual machine to implement MyConnectome analyses.

The MyConnectome project is a project meant to investigate the relations between mind, brain, and body across an extended period of time in a single individual. One of the major goals of the project is to serve as a testbed for reproducible analysis practices. For this reason, we have released the data and as much code as possible for the processing and analyses.

Code
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Settings

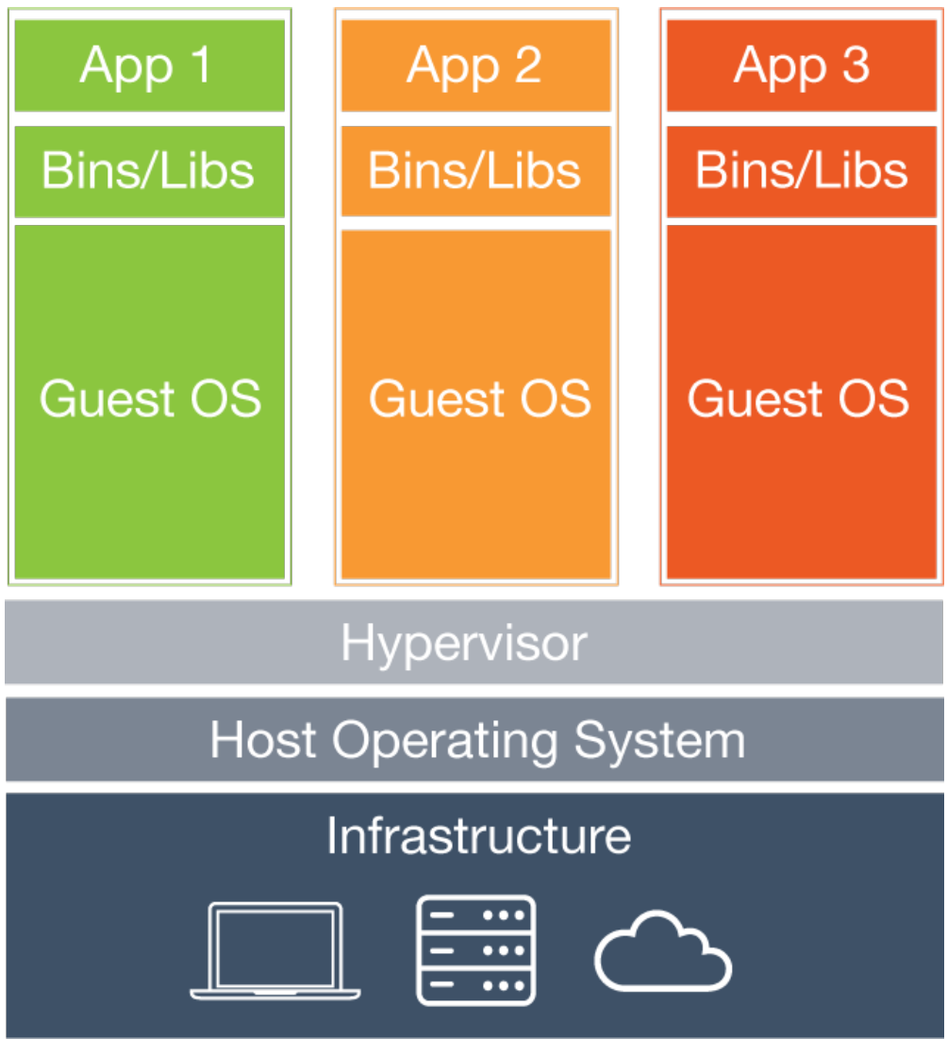
SSH clone URL
git@github.com:poldr

You can clone with HTTPS, SSH, or Subversion.

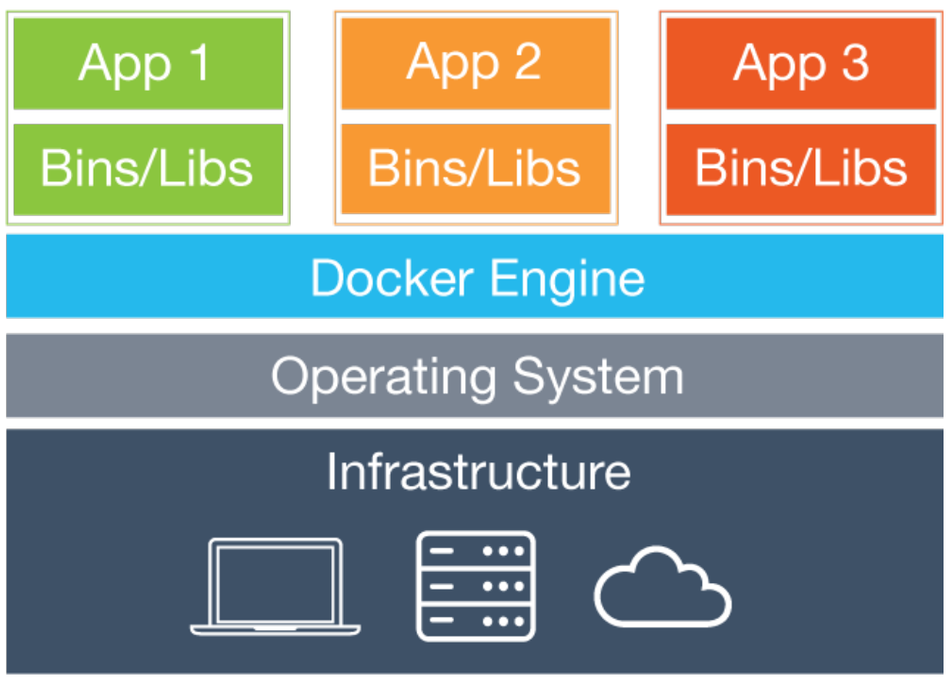
Clone in Desktop
Download ZIP

nihms-470697.pdf | yoi140096.pdf | Show All

Reproducible computing: VMs and containers



Virtual machines



Docker containers

BIDS Apps

A collection of containerized neuroimaging workflows and pipelines that accept datasets organized according to the Brain Imaging Data Structure (BIDS).

http://bids.neuroimaging.io

Repositories People 25 Teams 1 Settings

Filters ▾ Find a repository...

New repository

ndmg

Updated 2 hours ago

TeX ★ 0 0

niak

Updated 3 days ago

Python ★ 0 0

fmriprep

Updated 3 days ago

★ 0 0

People 25 >

Invite someone

Includes apps for: Freesurfer, SPM, HCP pipelines, Tracula, CPAC, FSL

Gorgolewski et al., 2017, *PLOS Comp Bio*

Running freesurfer via BIDS Apps

1. Install Docker

2. Run command:

```
docker run -ti --rm \
    -v /Users/filo/data/ds005:/bids_dataset:ro \
-v /Users/filo/outputs:/outputs \
bids/freesurfer:v6.0.0-2 \
/bids_dataset /outputs participant --participant_label 01 \
--license_key “<your license key here>"
```

OpenNEURO BETA

A free and open platform for analyzing and sharing neuroimaging data

Sign in with Google

Browse Public Datasets

Get Data

Browse and download datasets from contributors all over the world.

MORE

Share Data

Upload your data and collaborate with your colleagues or share it with users around the world.

MORE

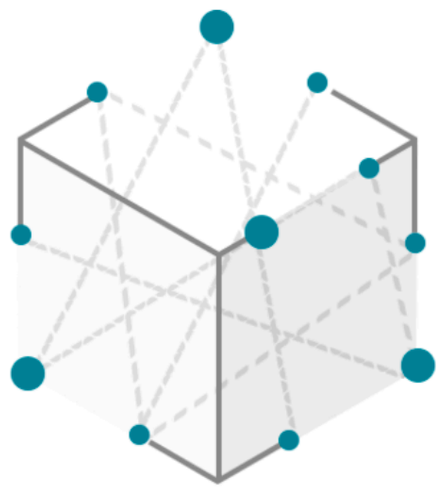
Use Data

Use our available pipelines to process any data on the site.

MORE

OpenNeuro Suite

- A set of workflows developed for the OpenNeuro project
  - Glass-box philosophy
    - Expose as many details as possible through detailed reports
  - Reproducibility
    - Versioned containers for Docker/ Singularity
  - Robustness
    - Testing with continuous integration
  - Community-driven
    - Heavily based on user feedback



OpenNEURO BETA

mriqc: a robust quality control workflow

Robust Image Quality Metrics (IQMs)

Quality Assessment of T1w and fMRI

Automated classification of T1w

(Esteban et. al 2017)

Visual reports

Ease and speed up individual eyeballing

Group reports: distribution of each IQM

Easy to use

I/O Standardization:

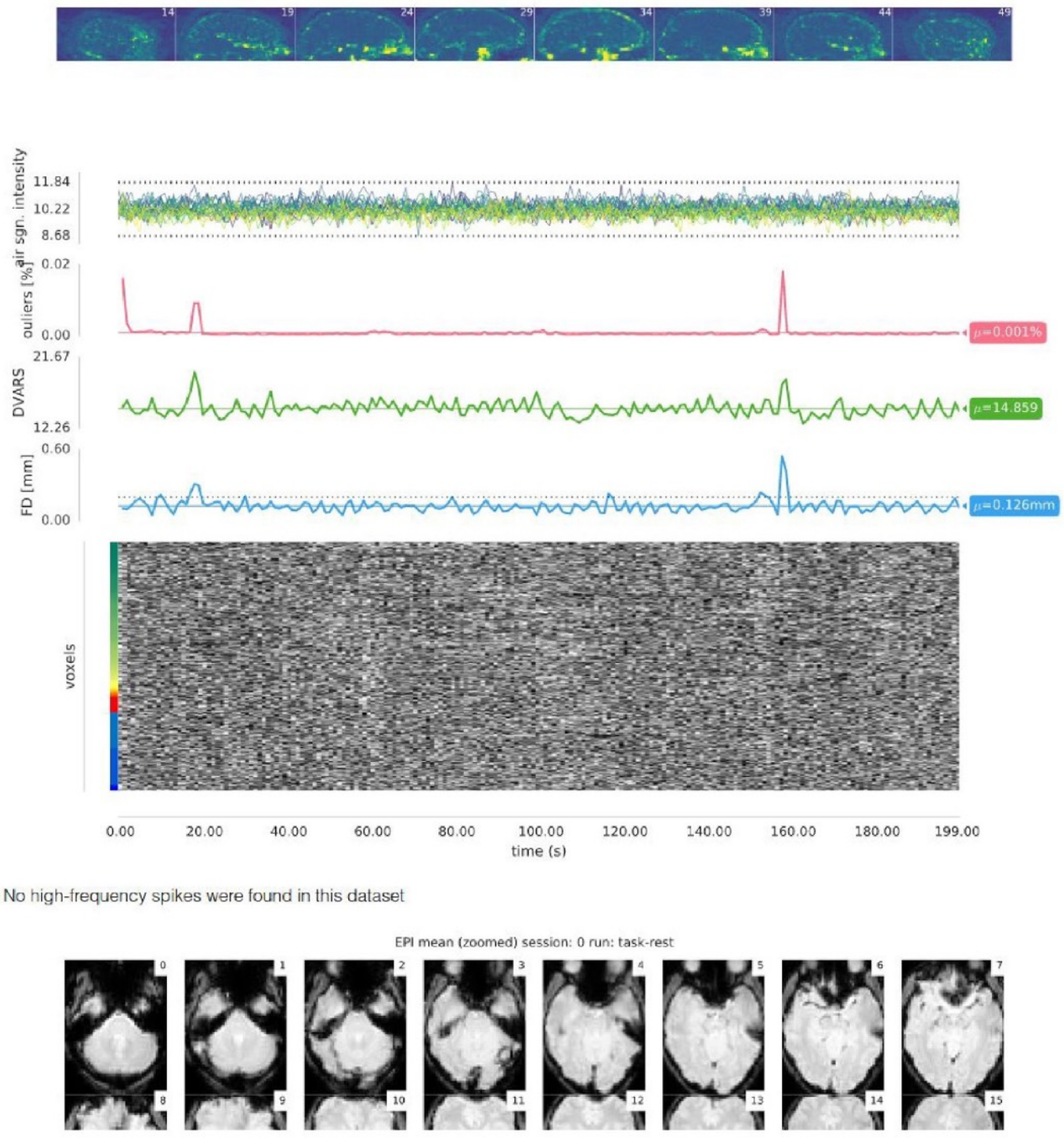
BIDS for inputs

BIDS-Apps for command line

Containerized:

Docker (poldracklab/mriqc)

Singularity (HPC friendly containers)



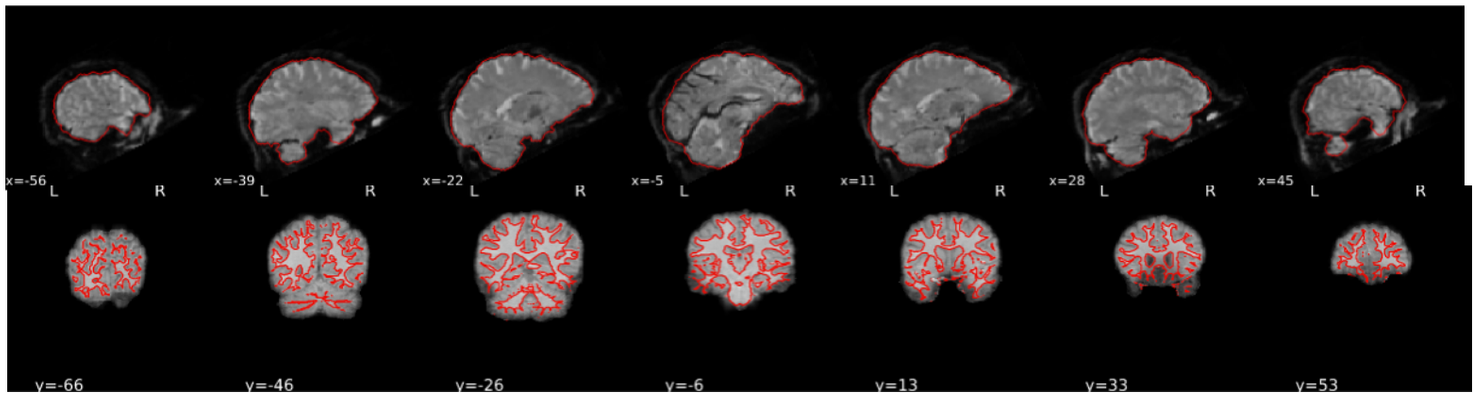
mriqc.org

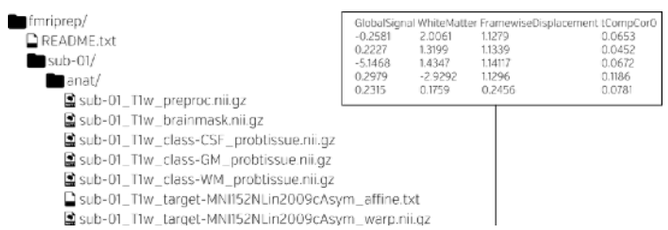
Esteban et al., in press, *PLOS One*

fmriprep: a robust and transparent preprocessing pipeline

Robust

takes any dataset,
combines well tested tools
across packages to provide
the best results





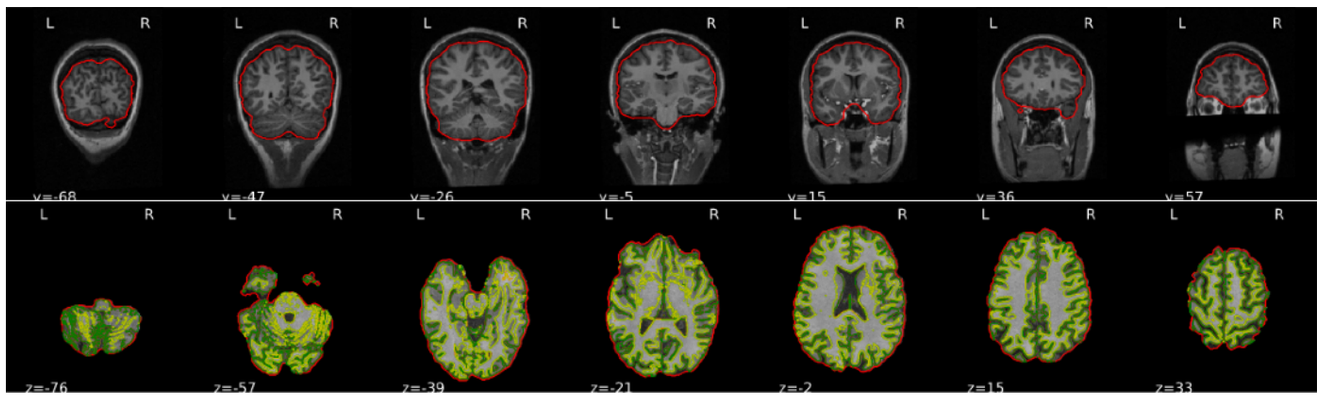
GlobalSignal	WhiteMatter	FramewiseDisplacement	tCompCor0
-0.2581	2.0061	1.1279	0.0653
0.2227	1.3199	1.1339	0.0452
-5.1468	1.4347	1.1417	0.0672
0.2979	-2.9292	1.1296	0.1186
0.2315	0.1759	0.2456	0.0781

Easy to use

Uses containers, works on Win, Mac, Linux and HPCs. Takes standardized datasets (BIDS) and outputs standardized derivatives.

Transparent

Produces interactive reports
that allow you to check
quality in minutes.



http://fmriprep.org

OPEN ACCESS Freely available online

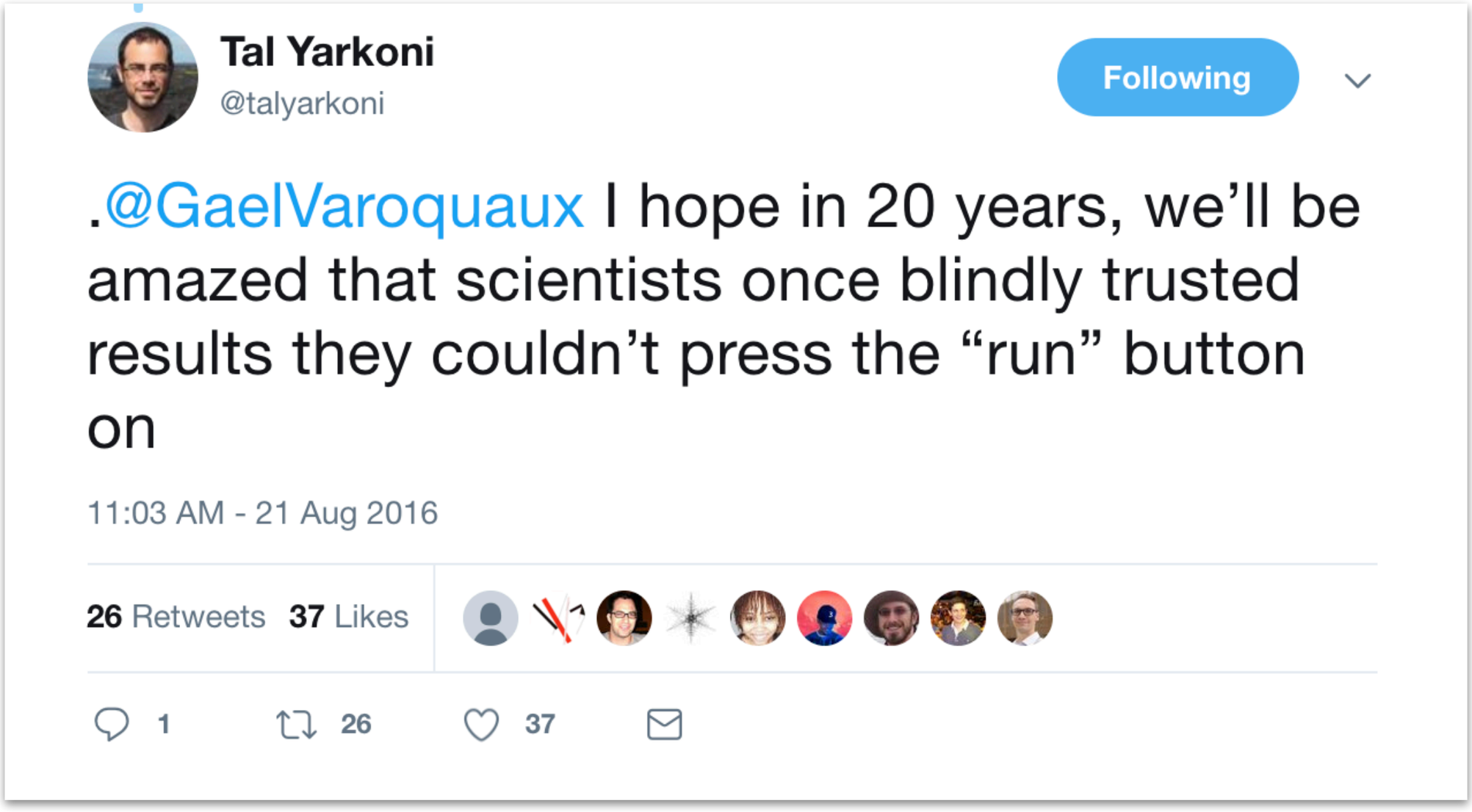
PLOS | COMPUTATIONAL BIOLOGY

Editorial

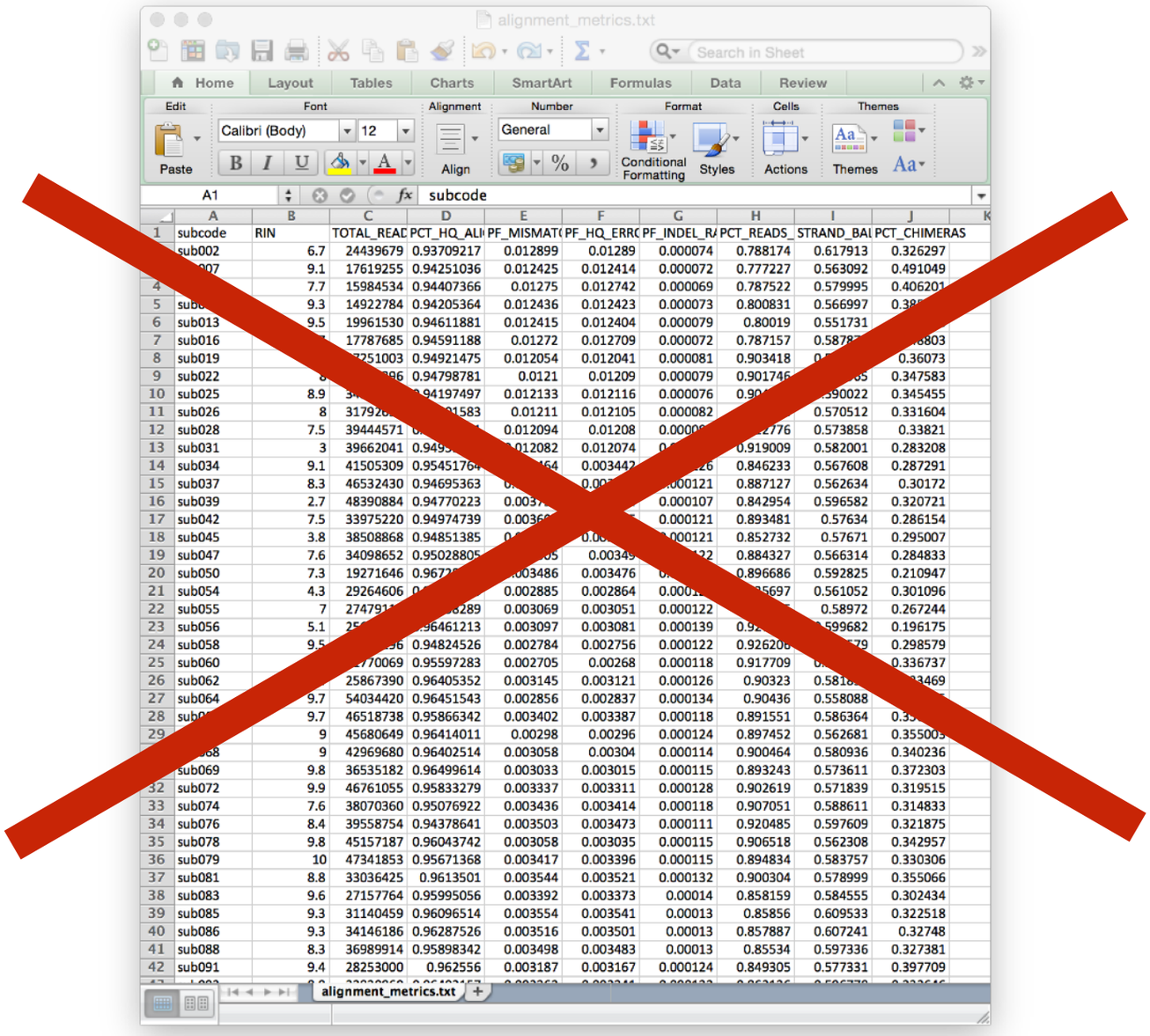
4 ~~Ten~~ Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

- Rule 1: For every result/figure, make sure you can reproduce it



- Corollary: Avoid Manual Data Manipulation Steps



	A	B	C	D	E	F	G	H	I	J
1	subcode	RIN	TOTAL_READ	PCT_HQ_ALI	PF_MISMAT(	PF_HQ_ERR(	PF_INDEL_R/	PCT_READS_	STRAND_BAI	PCT_CHIMERAS
2	sub002	6.7	24439679	0.93709217	0.012899	0.01289	0.000074	0.788174	0.617913	0.326297
3		9.1	17619255	0.94251036	0.012425	0.012414	0.000072	0.777227	0.563092	0.491049
4		7.7	15984534	0.94407366	0.01275	0.012742	0.000069	0.787522	0.579995	0.406201
5		9.3	14922784	0.94205364	0.012436	0.012423	0.000073	0.800831	0.566997	0.38
6	sub013	9.5	19961530	0.94611881	0.012415	0.012404	0.000079	0.80019	0.551731	
7	sub016		17787685	0.94591188	0.01272	0.012709	0.000072	0.787157	0.587	8803
8	sub019		7251003	0.94921475	0.012054	0.012041	0.000081	0.903418	0.4	0.36073
9	sub022		96	0.94798781	0.0121	0.01209	0.000079	0.901746	65	0.347583
10	sub025	8.9	3	0.94197497	0.012133	0.012116	0.000076	0.90	590022	0.345455
11	sub026	8	317920	1583	0.01211	0.012105	0.000082		0.570512	0.331604
12	sub028	7.5	39444571	0.9	0.012094	0.01208	0.0000	2776	0.573858	0.33821
13	sub031	3	39662041	0.949	012082	0.012074	0.0	0.919009	0.582001	0.283208
14	sub034	9.1	41505309	0.95451764	164	0.003442		26 0.846233	0.567608	0.287291
15	sub037	8.3	46532430	0.94695363	0.0	0.00	000121	0.887127	0.562634	0.30172
16	sub039	2.7	48390884	0.94770223	0.0037		0.000107	0.842954	0.596582	0.320721
17	sub042	7.5	33975220	0.94974739	0.0036		0.000121	0.893481	0.57634	0.286154
18	sub045	3.8	38508868	0.94851385	0.0	0.00	000121	0.852732	0.57671	0.295007
19	sub047	7.6	34098652	0.95028805	95	0.00349	22	0.884327	0.566314	0.284833
20	sub050	7.3	19271646	0.967	003486	0.003476	0.	0.896686	0.592825	0.210947
21	sub054	4.3	29264606	0.9	0.002885	0.002864	0.0001	35697	0.561052	0.301096
22	sub055	7	2747911	6289	0.003069	0.003051	0.000122		0.58972	0.267244
23	sub056	5.1	25	96461213	0.003097	0.003081	0.000139	0.92	0.599682	0.196175
24	sub058	9.5	2	0.94824526	0.002784	0.002756	0.000122	0.926208	79	0.298579
25	sub060		2770069	0.95597283	0.002705	0.00268	0.000118	0.917709	0.	0.336737
26	sub062		25867390	0.96405352	0.003145	0.003121	0.000126	0.90323	0.581	3469
27	sub064	9.7	54034420	0.96451543	0.002856	0.002837	0.000134	0.90436	0.558088	
28	sub	9.7	46518738	0.95866342	0.003402	0.003387	0.000118	0.891551	0.586364	0.3
29		9	45680649	0.96414011	0.00298	0.00296	0.000124	0.897452	0.562681	0.355003
30	58	9	42969680	0.96402514	0.003058	0.00304	0.000114	0.900464	0.580936	0.340236
31	sub069	9.8	36535182	0.96499614	0.003033	0.003015	0.000115	0.893243	0.573611	0.372303
32	sub072	9.9	46761055	0.95833279	0.003337	0.003311	0.000128	0.902619	0.571839	0.319515
33	sub074	7.6	38070360	0.95076922	0.003436	0.003414	0.000118	0.907051	0.588611	0.314833
34	sub076	8.4	39558754	0.94378641	0.003503	0.003473	0.000111	0.920485	0.597609	0.321875
35	sub078	9.8	45157187	0.96043742	0.003058	0.003035	0.000115	0.906518	0.562308	0.342957
36	sub079	10	47341853	0.95671368	0.003417	0.003396	0.000115	0.894834	0.583757	0.330306
37	sub081	8.8	33036425	0.9613501	0.003544	0.003521	0.000132	0.900304	0.578999	0.355066
38	sub083	9.6	27157764	0.95995056	0.003392	0.003373	0.00014	0.858159	0.584555	0.302434
39	sub085	9.3	31140459	0.96096514	0.003554	0.003541	0.00013	0.85856	0.609533	0.322518
40	sub086	9.3	34146186	0.96287526	0.003516	0.003501	0.00013	0.857887	0.607241	0.32748
41	sub088	8.3	36989914	0.95898342	0.003498	0.003483	0.00013	0.85534	0.597336	0.327381
42	sub091	9.4	28253000	0.962556	0.003187	0.003167	0.000124	0.849305	0.577331	0.397709

Growth in a Time of Debt

By Carmen M. Reinhart and Kenneth S. Rogoff*

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

Reinhard & Rogoff have clearly exerted a major influence in recent years on public policy debates over the management of government debt and fiscal policy more broadly. Their findings have provided significant support for the austerity agenda that has been ascendant in Europe and the United States since 2010. - Herndon et al., 2013

The Washington Post

Wonkblog

Is the evidence for austerity based on an Excel spreadsheet error?

By Brad Plumer April 16, 2013

"Reinhart and Rogoff appear to have made an error with one of their Excel spreadsheet formulas. By typing AVERAGE(L30:L44) at one point instead of AVERAGE(L30:L49), they left out Belgium, a key counterexample [to their claim]"

Debt, Growth and the Austerity Debate

By CARMEN M. REINHART and KENNETH S. ROGOFF APRIL 25, 2013

Last week, three economists at the University of Massachusetts, Amherst, released a paper criticizing our findings. They correctly identified a spreadsheet coding error that led us to miscalculate the growth rates of highly indebted countries since World War II.

- Rule 2: Use version control for all computer code

vs.

RTanalysis_script3_June1_good_try4.R

- Rule 3: Build quality control into your analyses

PLOS ONE

CORRECTION

Correction: The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science

Stephan Lewandowsky, Gilles E. Gignac, Klaus Oberauer

The dataset included two notable age outliers (reported ages 5 and 32757).

Specifically, the statement on page 9 “age turned out not to correlate with any of the indicator variables” is incorrect. It should read instead “age correlated significantly with 3 latent indicator variables (Vaccinations: .219, $p < .0001$; Conservatism: .169, $p < .001$; Conspiracist ideation: -.140, maximum likelihood $p < .0001$, bootstrapped $p = .004$), and straddled significance for a fourth (Free Market: .08, p%.05).”

```
In [1]: age=32757

In [2]: assert age>12 and age<120
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-2-37de876b5fda> in <module>()
----> 1 assert age>12 and age<120

AssertionError:
```

- Rule 4: Make your data, code, and results public

Improvement: Severe testing

The road to wisdom? --- Well, it's plain
and simple to express:
Err
and err
and err again
but less
and less
and less.

- Piet Hein

- We need to subject our ideas to "severe testing" (D. Mayo)
  - Tests that have a high likelihood of showing that we are wrong, if in fact we are
  - Need to apply this to our scientific ideas as well as our software

Conclusions

- Humans are biased in ways that work against the goals of science
- We need to redesign the choice architecture of neuroimaging methods so that it prevents rather than affords fooling ourselves
- Doing these things will make your life harder but make your science better
    - If every experiment "works" then are we really doing interesting science?

Things you can do today

- Never run another underpowered study
- Pre-register your research
- Use nonparametric corrections for multiple comparisons in fMRI
- Share your data and code
- Use version control and software testing
    - If you don't know how, sign up for a Software Carpentry course

Acknowledgments

laura and john arnold foundation®

NIH National Institute on Drug Abuse

NSF

The Poldrack Lab @ Stanford

http://reproducibility.stanford.edu