Varieties of "Voodoo"

Ed Vul
MIT

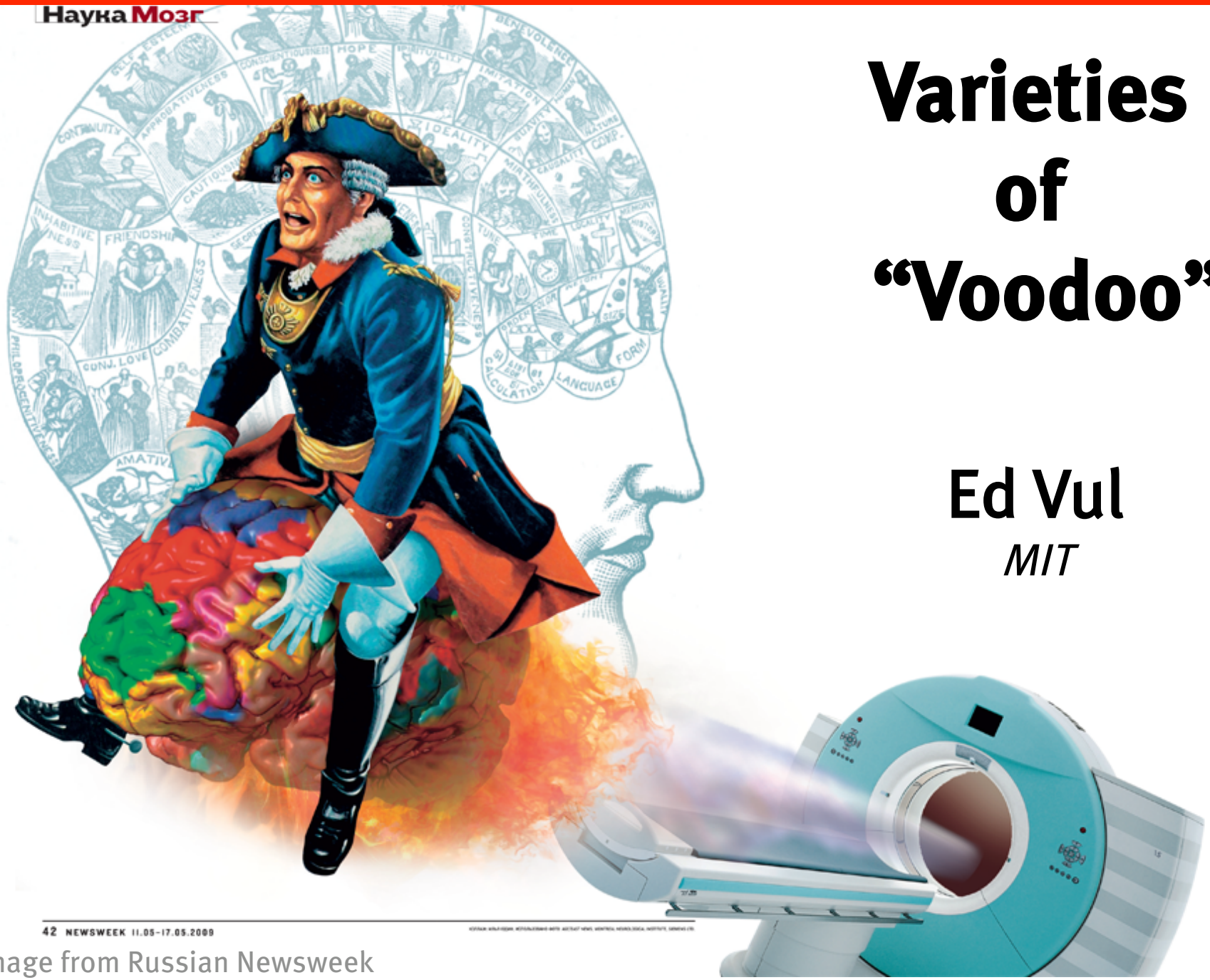
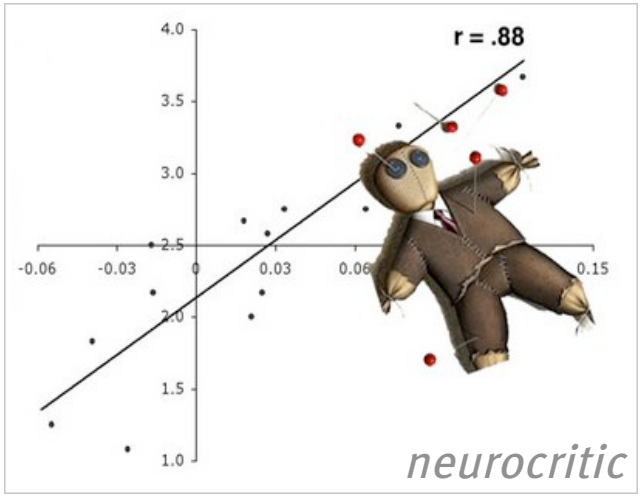


Image from Russian Newsweek

Reception

"such interchanges, in my experience, lead to more heat than light"



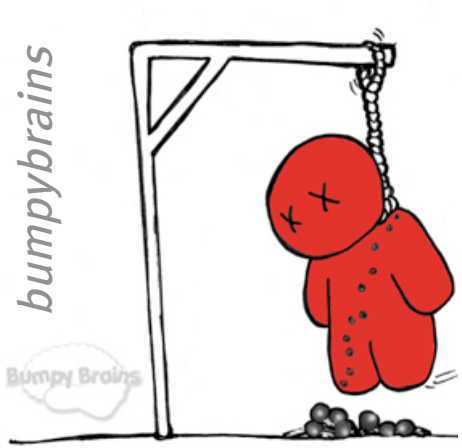
"What is the most we can learn about the mind by studying the brain?"

"grossly over-strong pronouncements"

"I haven't had so much fun since Watergate!"

"an unrepentant Vul"

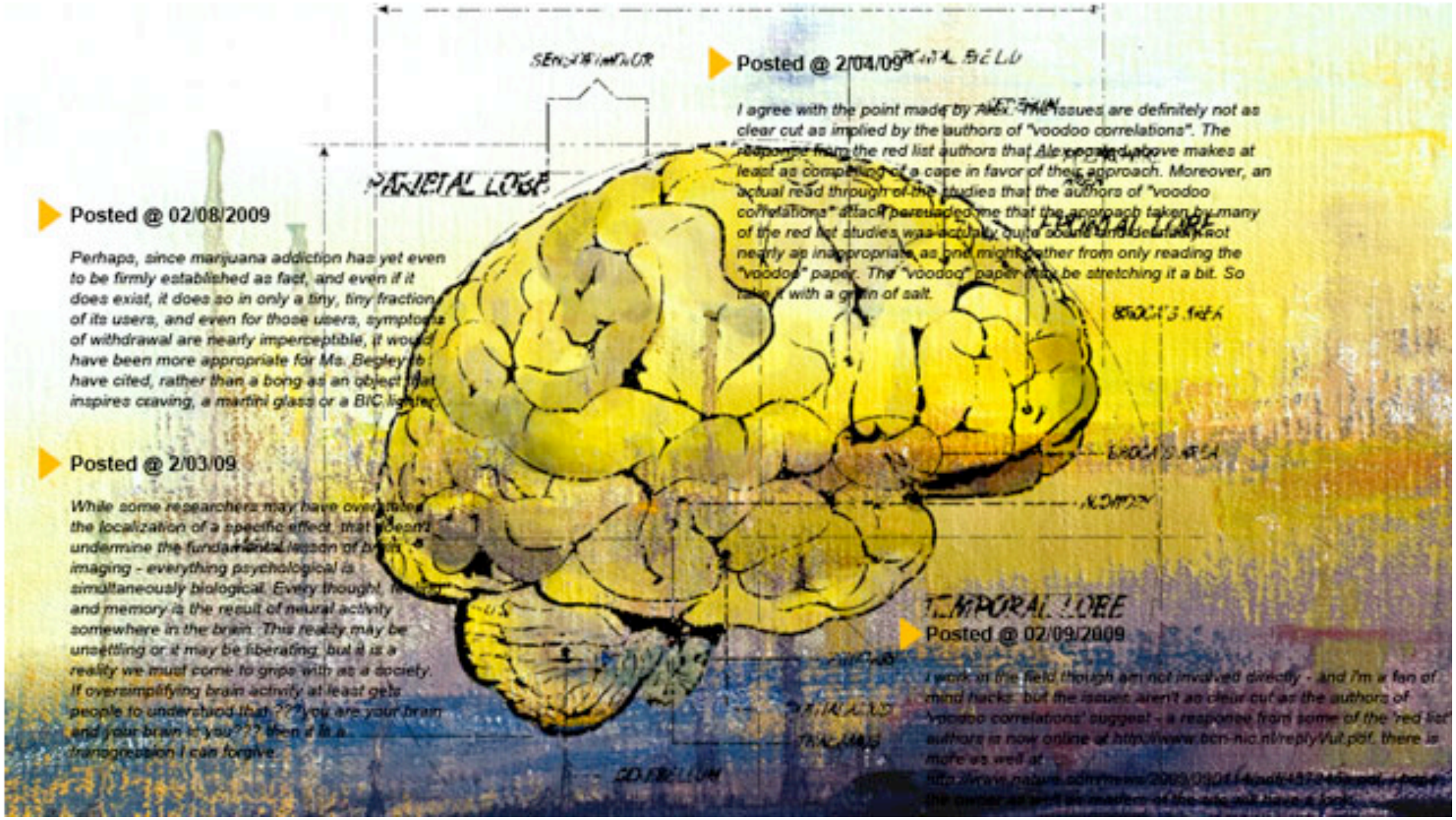
"they look like idiots because they're out of their statistical depth"



"there are people who would otherwise be dead if we adopted Vul's opinions"

Reception

"such interchanges, in my experience, lead to more heat than light"



From SEED magazine

Varieties of “voodoo”

85 Historical perspective

$r = 0.96$ What captured our attention?

The “voodoo correlation” method and generalizations

Objections

! Solutions

B-Projective Psychokinesis Test

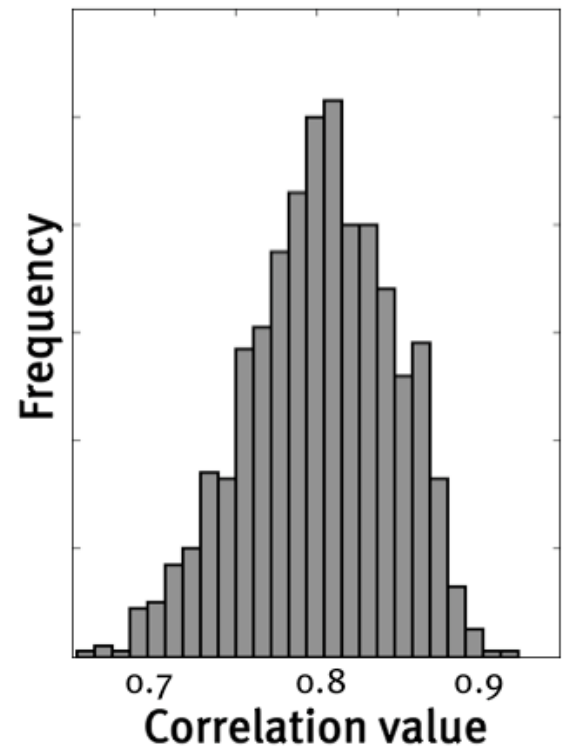
Edward Cureton (1950)

Can we use the same data to make an item analysis and to validate the test?

1) Noise data:
   29 students & grades
   85 random flips/student
2) Find tags predicting GPA
3) Evaluate validity on same data

Result: validity = 0.82

This analysis produces high validity from pure noise.



Thanks to: Dirk Vorberg!

The moral of this story, I think, is clear. When a validity coefficient is computed from the same data used in making an item analysis, this coefficient cannot be interpreted uncritically. And, contrary to many statements in the literature, it cannot be interpreted "with caution" either. There is one clear interpretation for all such validity coefficients. This interpretation is—

"Baloney!"

Cureton (1950)

A new name in each field

- Auctioneering: "the winner's curse"
- Machine learning: "testing on training data"
  "data snooping"
- Modeling: "overfitting"
- Survey sampling: "selection bias"
- Logic: "circularity"
- Meta-analysis: "publication bias"
- fMRI: "double-dipping"
  "non-independence"

Varieties of "voodoo"

Historical perspective

$r = 0.96$

What captured our attention?

The "voodoo correlation" method and generalizations

Objections

! Solutions

Large Correlations

Interoceptive accuracy, an outcome associated with neuroticism, was better accounted for by dACC reactivity ($r^2 = .74$) than by self-reported neuroticism ($r^2 = .16$), suggesting that neural reactivities may provide a more direct measure of personality than self-reports do.

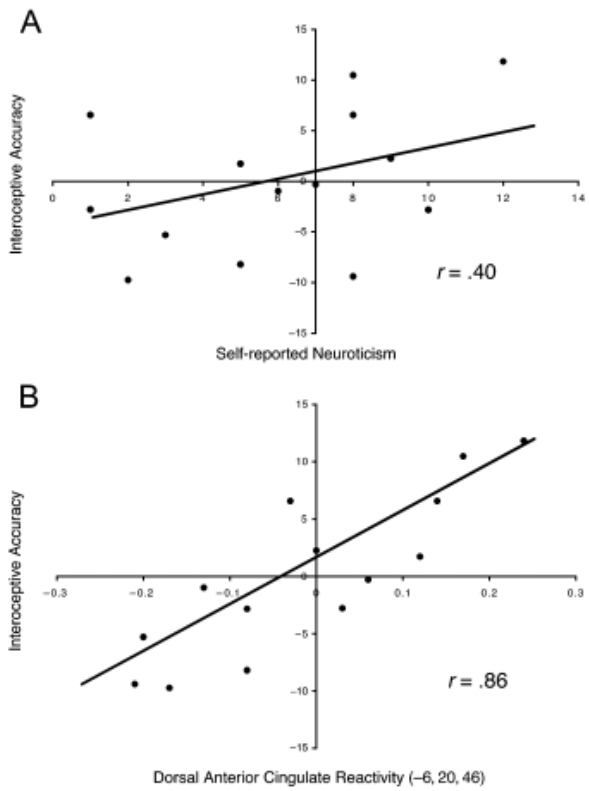


Figure 4. (A) Scatterplot showing self-reported neuroticism scores, plotted against interoceptive accuracy (controlling for arousal levels) across participants. (B) Scatterplot showing dorsal anterior cingulate cortex activations to oddball trials relative to nonoddball trials, plotted against interoceptive accuracy (controlling for arousal levels) across participants.

r = 0.86

Eisenberger, Lieberman, & Satpute (2005)

Large Correlations

Neural Correlates of Human Virtue Judgment

Hidehiko Takahashi[1,2], Motoichiro Kato[3], Masato Matsuura[2], Michihiko Koeda[4], Noriaki Yahata[5], Tetsuya Suhara[1] and Yoshiro Okubo[4]

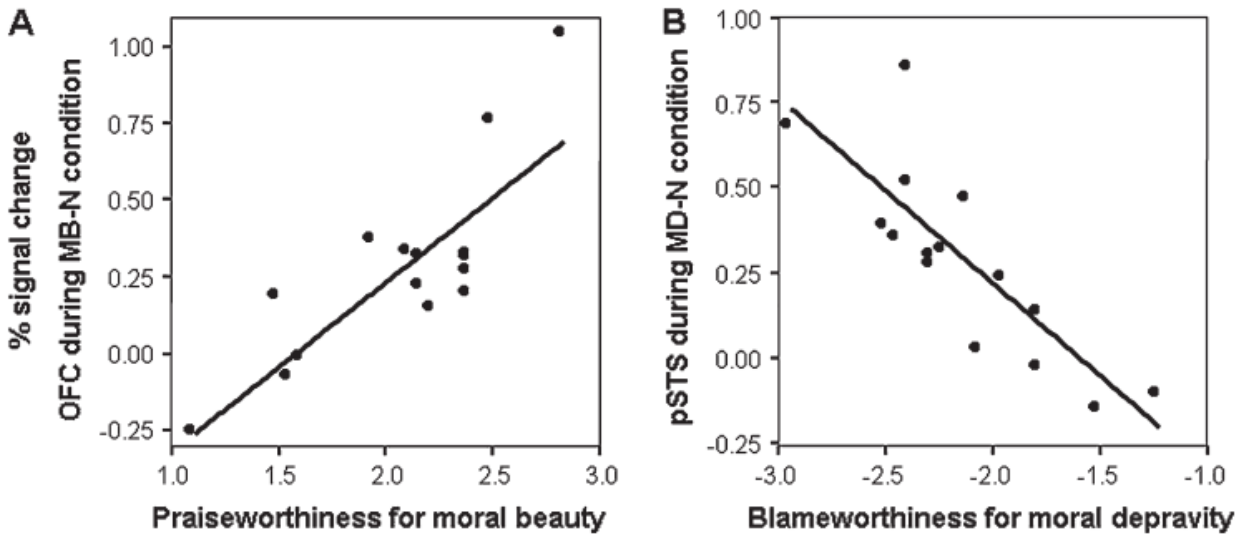


Figure 3. Regression lines of correlations between (*A*) praiseworthiness (*B*) blameworthiness and degree of brain activation. (*A*) There were correlations ($r = 0.82$, degrees of freedom [df] = 13, $P < 0.001$) between self-rating of praiseworthiness and degree of activation in OFC. (*B*) There were positive linear correlations ($r = -0.83$, df = 13, $P < 0.001$) between self-rating of blameworthiness and degree of activation in pSTS.

$r = 0.82$

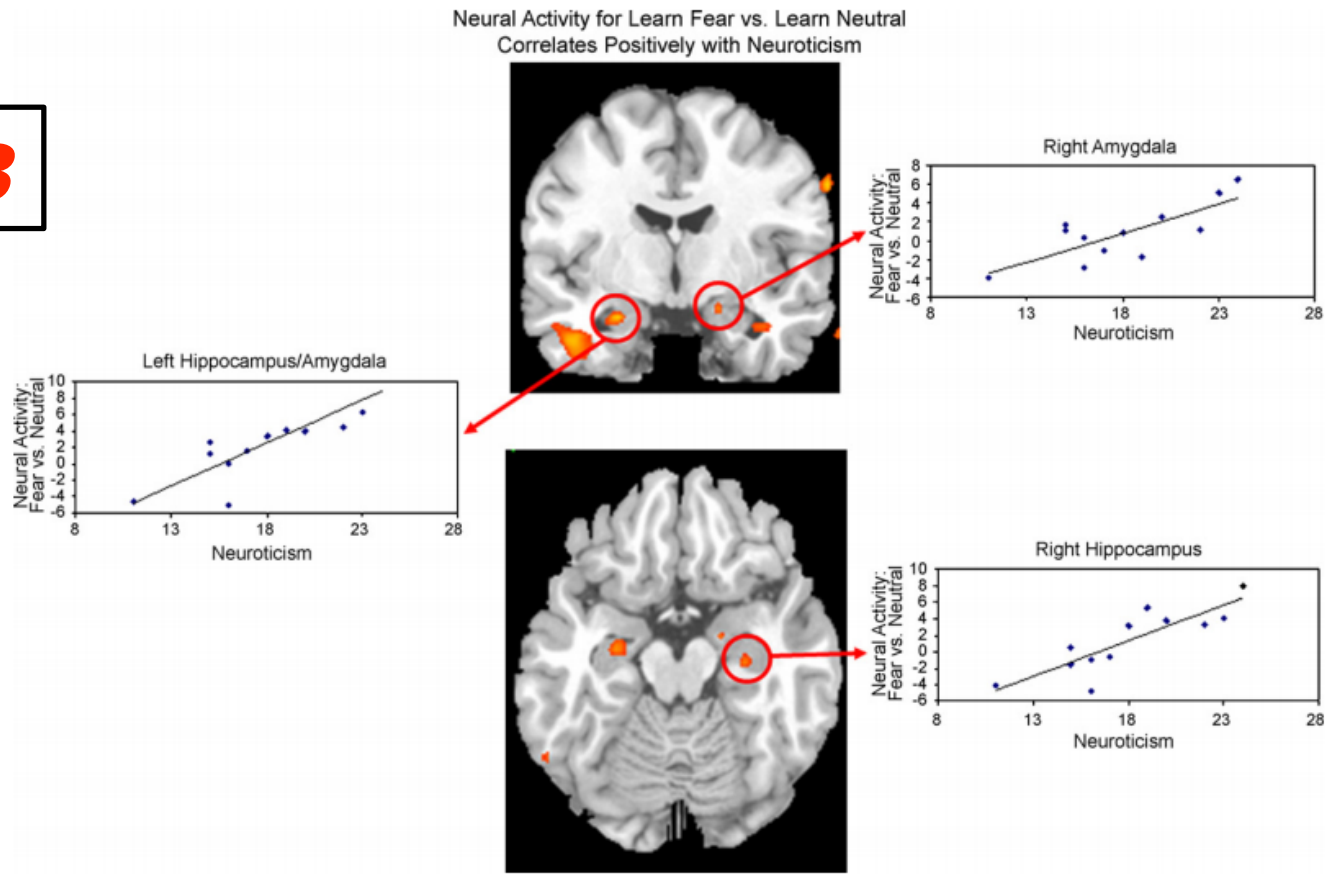
Takahashi, Kato, Matsuura, Koeda, Yahata, Suhara, & Okubo (2008)

Large Correlations

The influence of personality on neural mechanisms of observational fear and reward learning

Christine I. Hooker[a,*], Sara C. Verosky[b], Asako Miyakawa[c], Robert T. Knight[c,d], Mark D'Esposito[c,d]

$r = 0.88$



Hooker, Verosky, Miyakawa, Knight, & D'Esposito (2008)

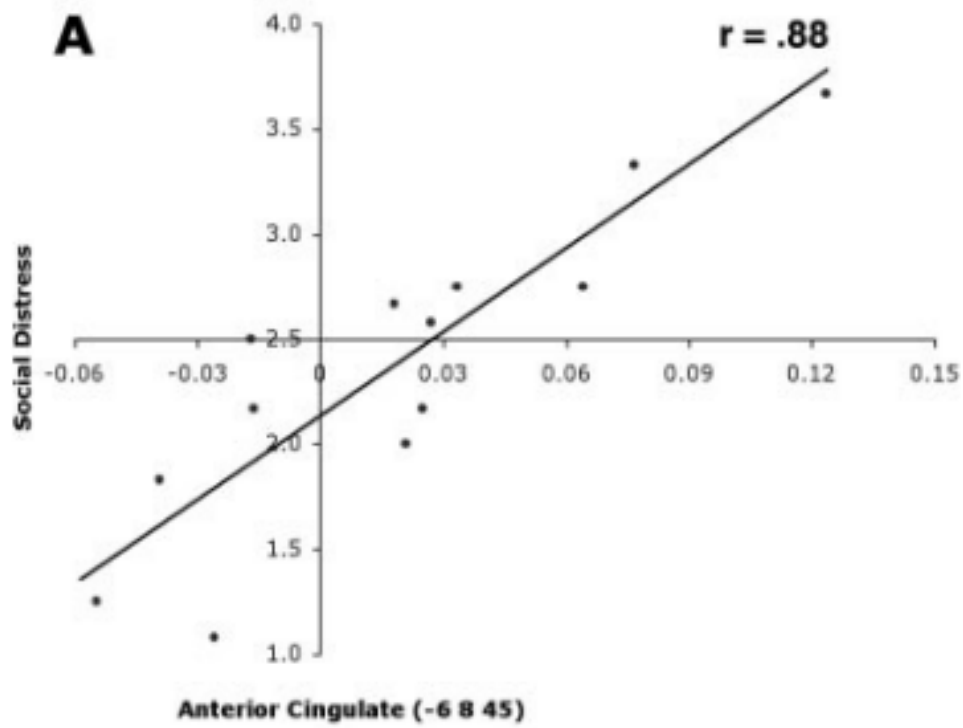
Large Correlations

Does Rejection Hurt? An fMRI Study of Social Exclusion

Naomi I. Eisenberger,[1]* Matthew D. Lieberman,[1]
Kipling D. Williams[2]

r = 0.88

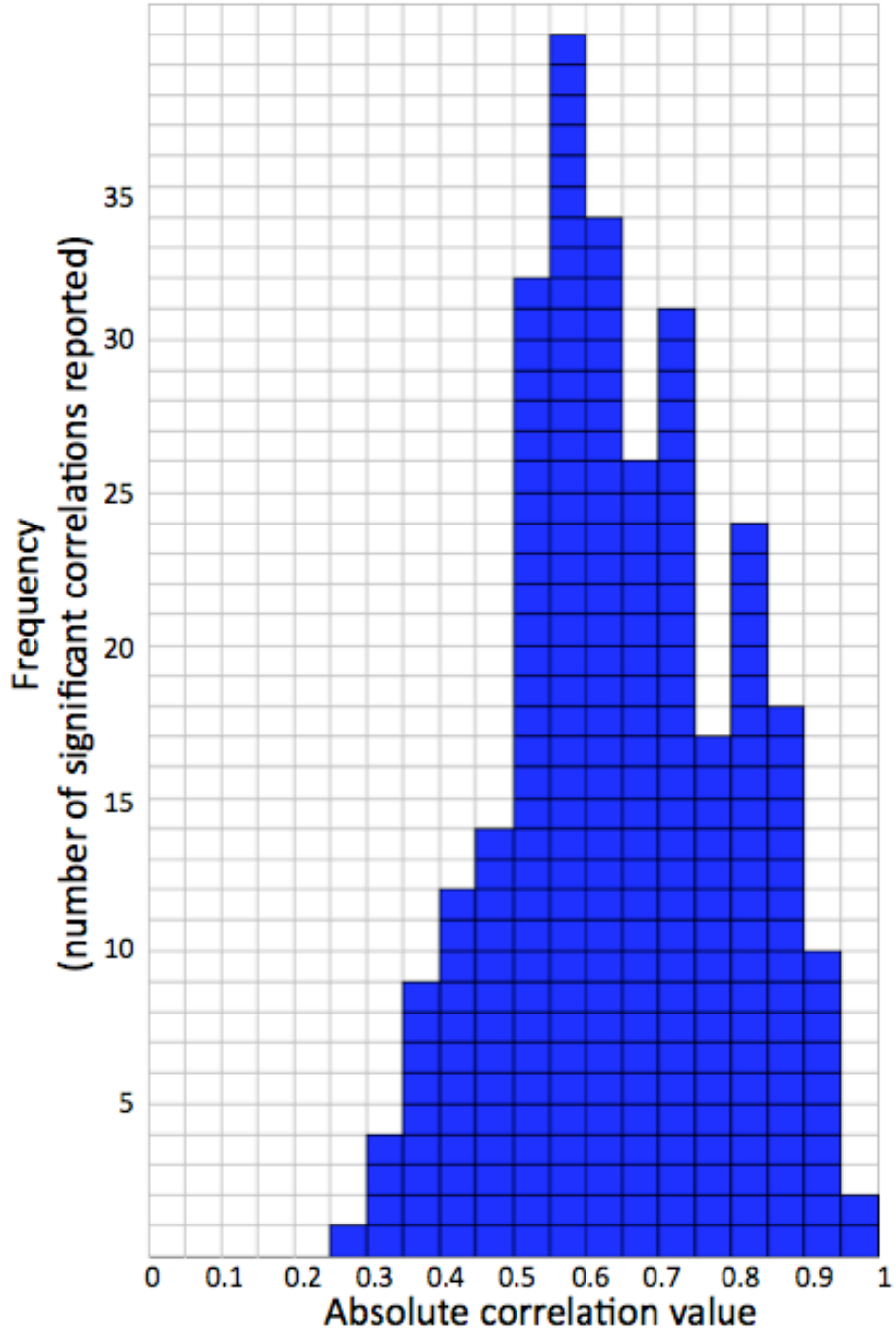
Fig. 2. Scatterplots showing the relation during exclusion, relative to inclusion, between (**A**) ACC activity and self-reported distress, (**B**) RVPFC and self-reported distress, and (**C**) ACC and RVPFC activity. Each point represents the data from a single participant.



Eisenberger, Lieberman, & Williams (2003)

How common are these correlations?

Phil Nguyen surveyed soc/aff/person fMRI, in search of these correlations.



Vul, Harris, Winkielman, & Pashler (2009)

What's wrong with big correlations?

- (Expected) observed correlations are limited by reliability:

$$r(\hat{A},\hat{B}) = r(A,B)\sqrt{r(\hat{A}_1,\hat{A}_2)r(\hat{B}_1,\hat{B}_2)}$$

Expected observed correlation

True underlying correlation

Geometric mean of measure reliabilities

- **Maximum** expected observed correlation should be the geo. mean. of reliabilities…

Reliability of personality measures

- Viswesvaran and Ones (2000)
  reliability of the Big Five: .73 to .78.

- Hobbs and Fowler (1974)
  MMPI: .66 - .94 (mean=.84)

We say:

0.8 optimistic guess for ad hoc scales

Reliability of fMRI measures

- (a lot of variability depending on measure)
- Choose estimates most relevant for across-subject correlations
  - 0-0.76 Kong et al. (2006)
  - 0.23 – 0.93 (mean=0.60) Manoach et al (2001)
  - ~0.8 Aron, Gluck, and Poldrack (2006)
  - 0.59-0.64 Grinband (2008)

- We say:

not often over 0.7

Maximum expected correlation

Expected observed correlation

True underlying correlation

Geometric mean of measure reliabilities

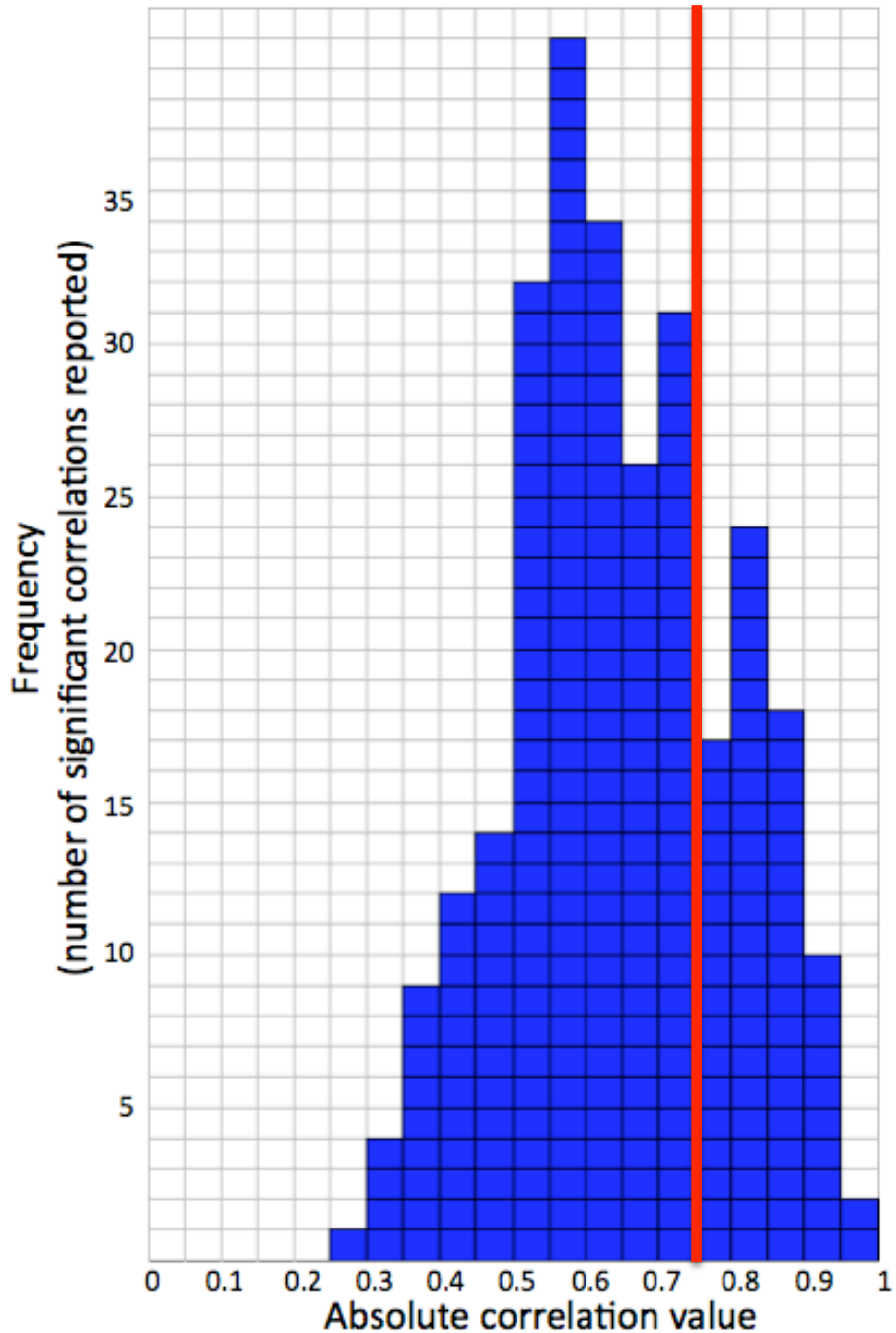
$$r(\hat{A},\hat{B}) = r(A,B)\sqrt{r(\hat{A}_1,\hat{A}_2)r(\hat{B}_1,\hat{B}_2)}$$

$$\boxed{0.74} = \left(1.0\right)\sqrt{\left(0.7\right)\left(0.8\right)}$$

“Upper bound”?

Many exceeding the **maximum possible** expected correlation.

(variability is possible, but this struck us as exceedingly unlikely)



Vul, Harris, Winkielman, & Pashler (2009)

Varieties of "voodoo"

Historical perspective

$r = 0.96$

What captured our attention?

The "voodoo correlation" method and generalizations

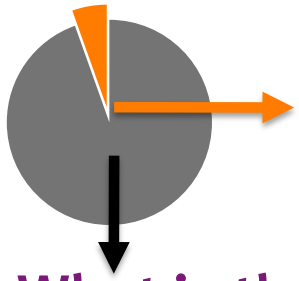
Objections

! Solutions

How are these correlations produced?

Response?

Would you please be so kind as to answer a few very quick questions about the analysis that produced, i.e., the correlations on page XX?...



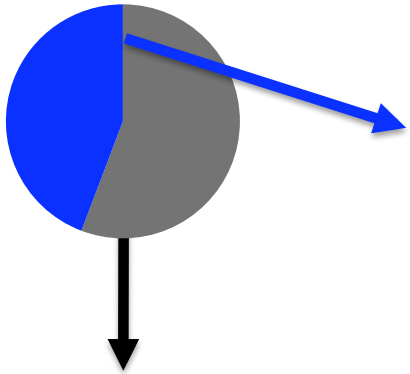
"No response"

1 What is the brain measure: One peak voxel, or average of several voxels?

The data plotted reflect the percent signal change or difference in parameter estimates (according to some contrast) of...

1 ...the average of a number of voxels

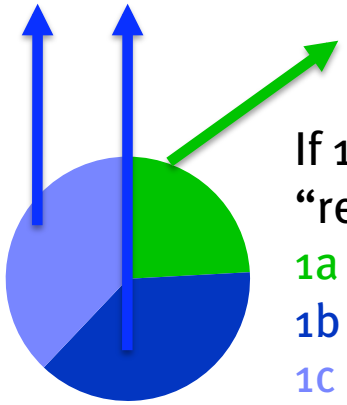
2 ...one peak voxel that was most significant according to some functional measure



Go to question 2

"Independent"

1.1 How was this set of voxels selected?



If 1: The voxels whose data were plotted (i.e., the "region of interest") were selected based on...

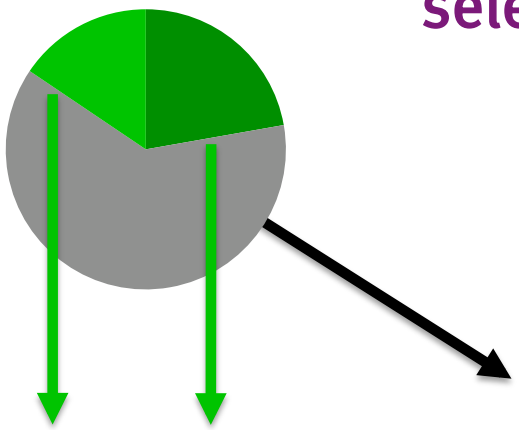
1a ...*only* anatomical constraints

1b ...*only* functional constraints

1c ... anatomical and functional constraints

How are these correlations produced?

2. What sort of functional selection?



"Independent"

The functional measure used to select the voxel(s) plotted in the figure was...

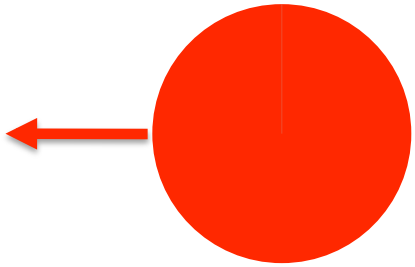
A ...a contrast within individual subjects

B ...the result of running a regression, across subjects, of the behavioral measure of interest against brain activation (or a contrast) at each voxel

C ...something else?

3. Same or different data?

"Non-Independent"



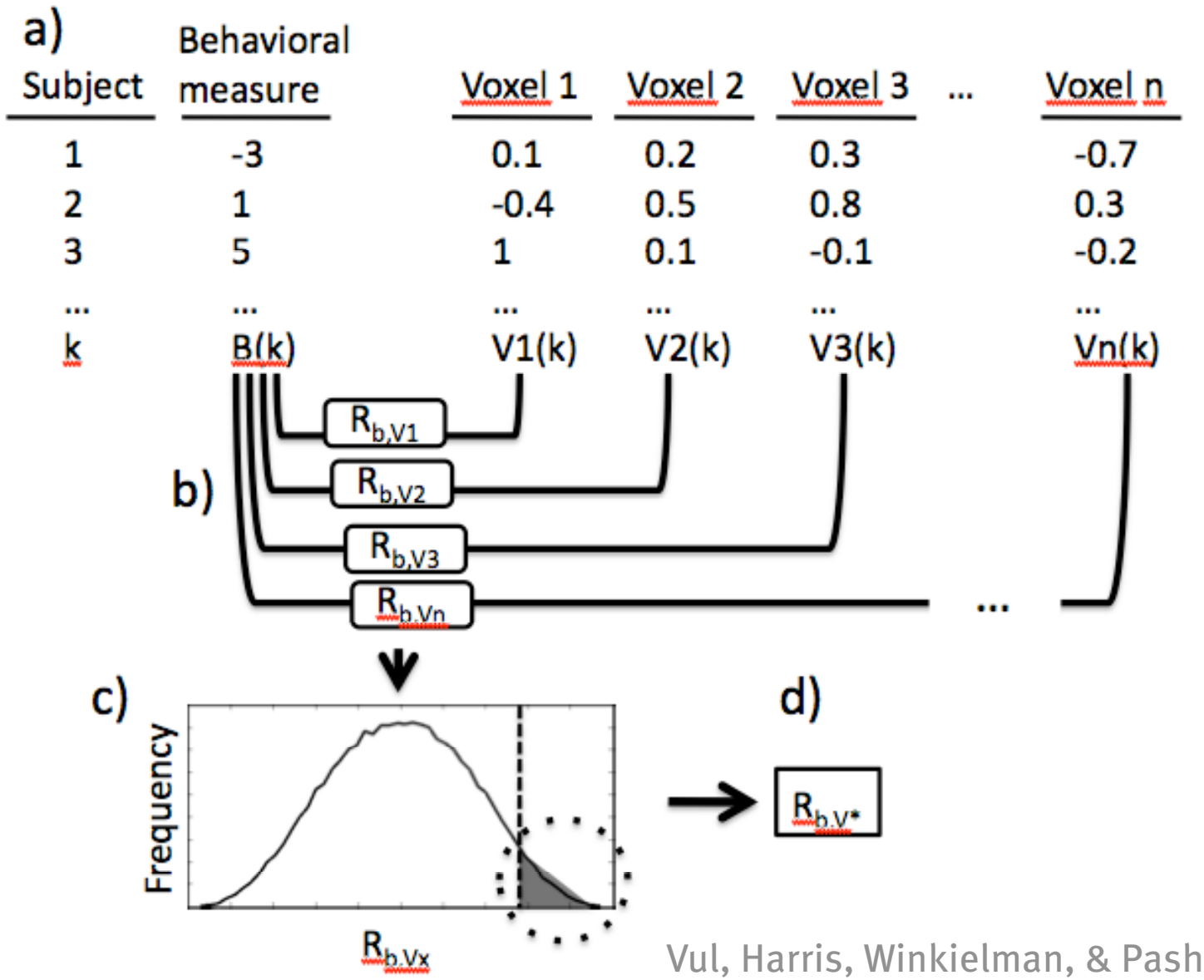
Finally: the fMRI data (runs/blocks/trials) displayed in the figure were...

A ...the same data as those employed in the analysis used to select voxels

B ...different data from those employed in the analysis used to select voxels.

54% said correlations were computed in voxels that were selected for containing high correlations.

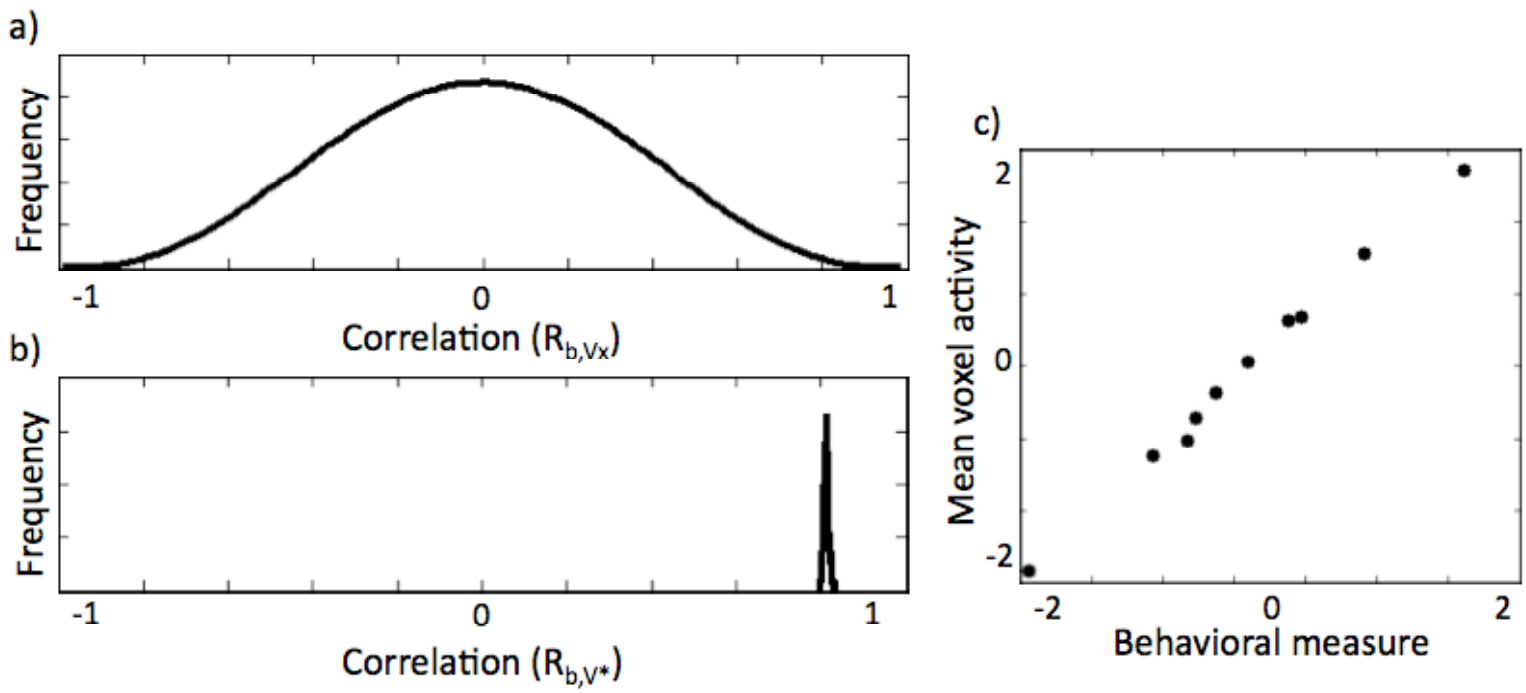
54% did what?



Vul, Harris, Winkielman, & Pashler (2009)

What would this do on noise?

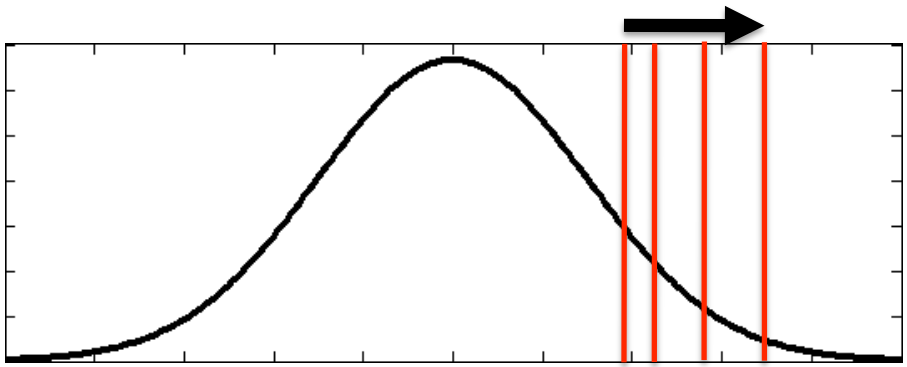
Simulate 10 subjects, 10,000 voxels each.

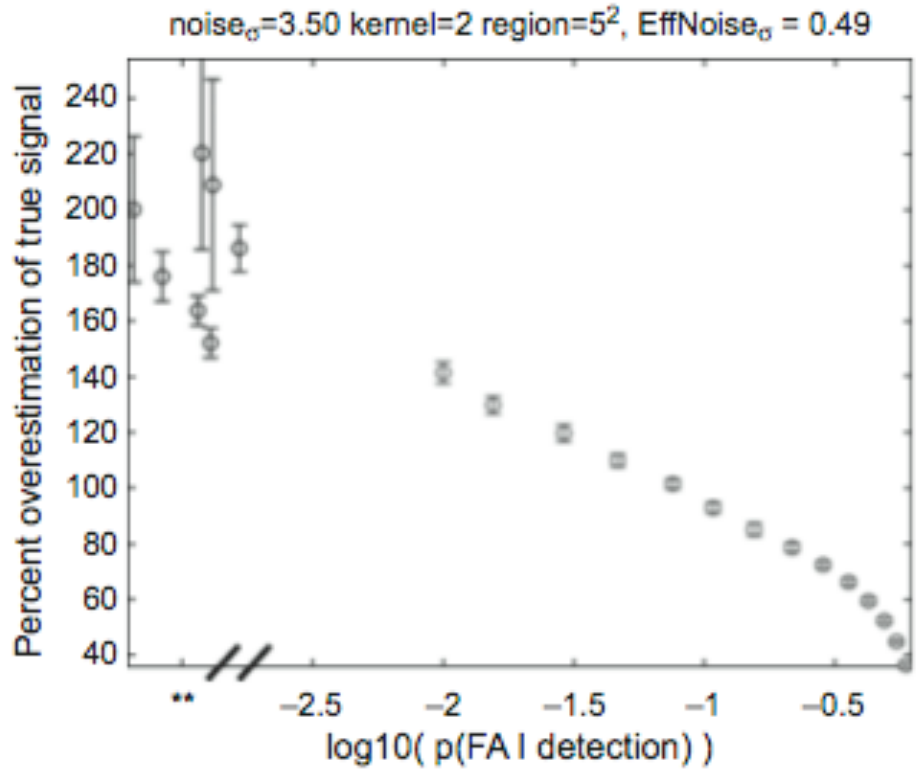


This analysis produces high correlations from pure noise.

Vul, Harris, Winkielman, & Pashler (2009)

False Alarms vs. Inflation





Multiple comparisons correction:

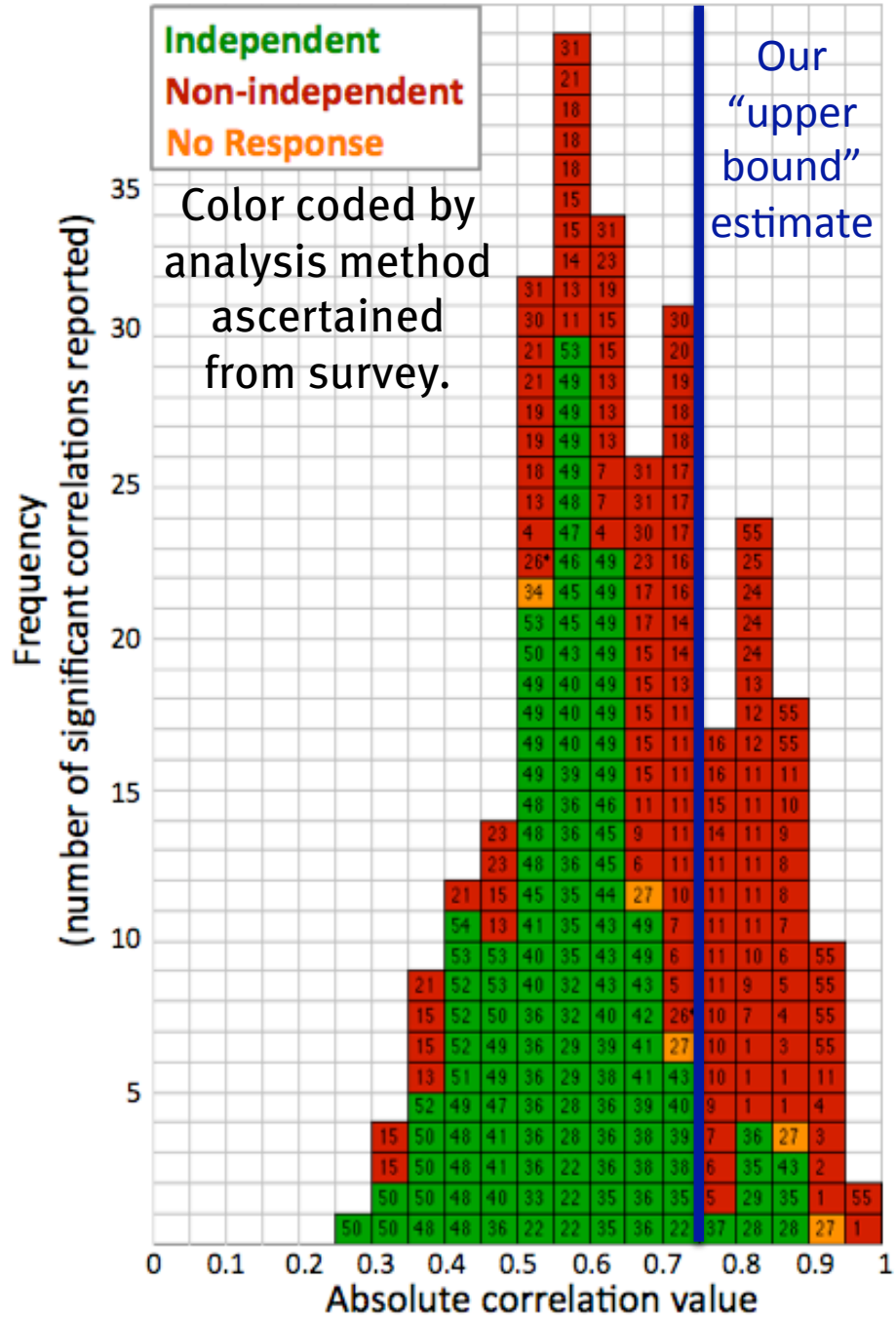
1) Minimizes false alarms

2) Exacerbates effect-size inflation

Vul, Harris, Winkielman, & Pashler (2009)

Source of high correlations

Biased method produced the majority of the "surprisingly high" correlations



Vul, Harris, Winkielman, & Pashler (2009)

Varieties of "voodoo"

Historical perspective

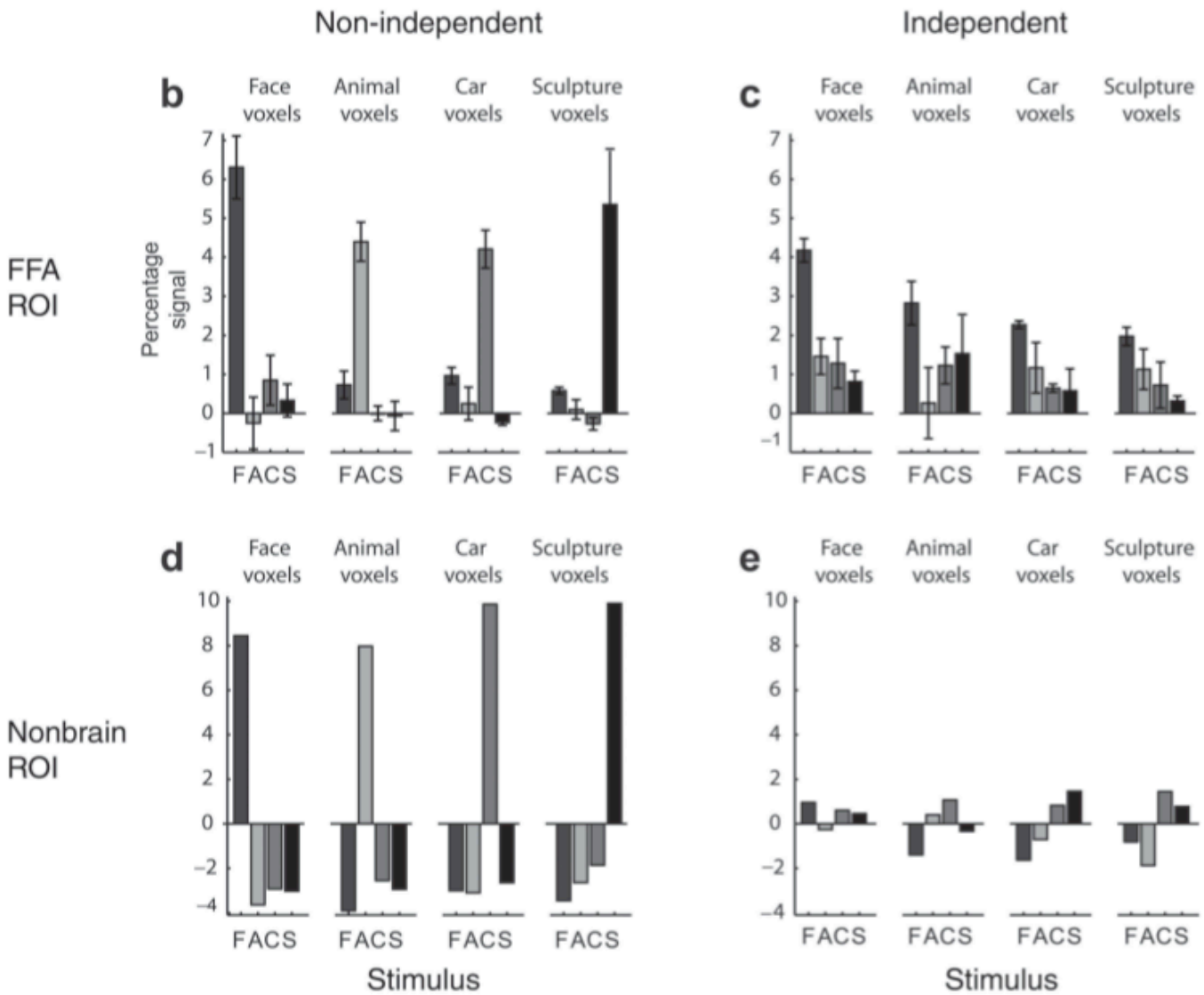
$r = 0.96$

What captured our attention?

The "voodoo correlation" method and generalizations

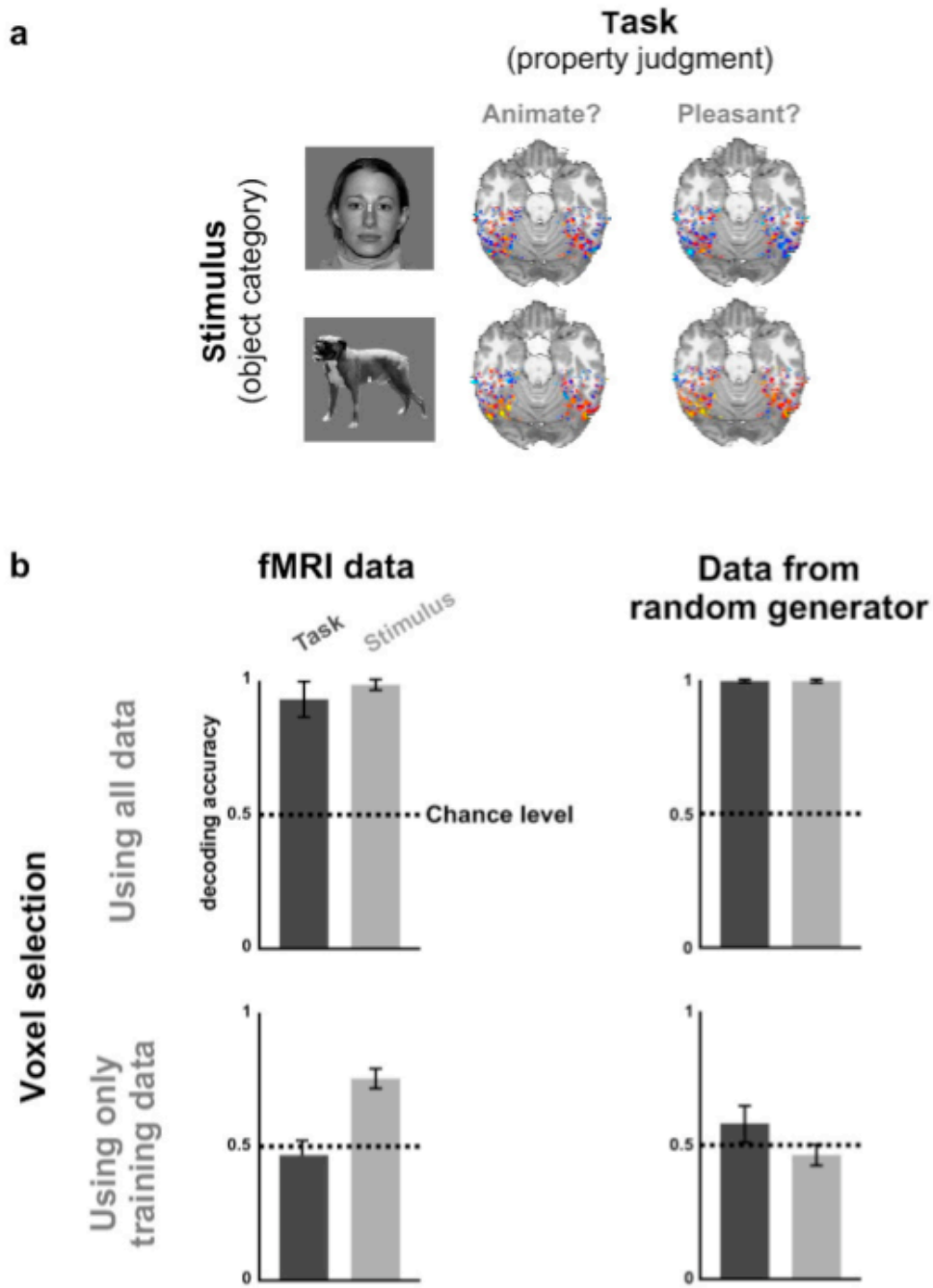
Objections

! Solutions



This analysis produces high selectivity from pure noise.

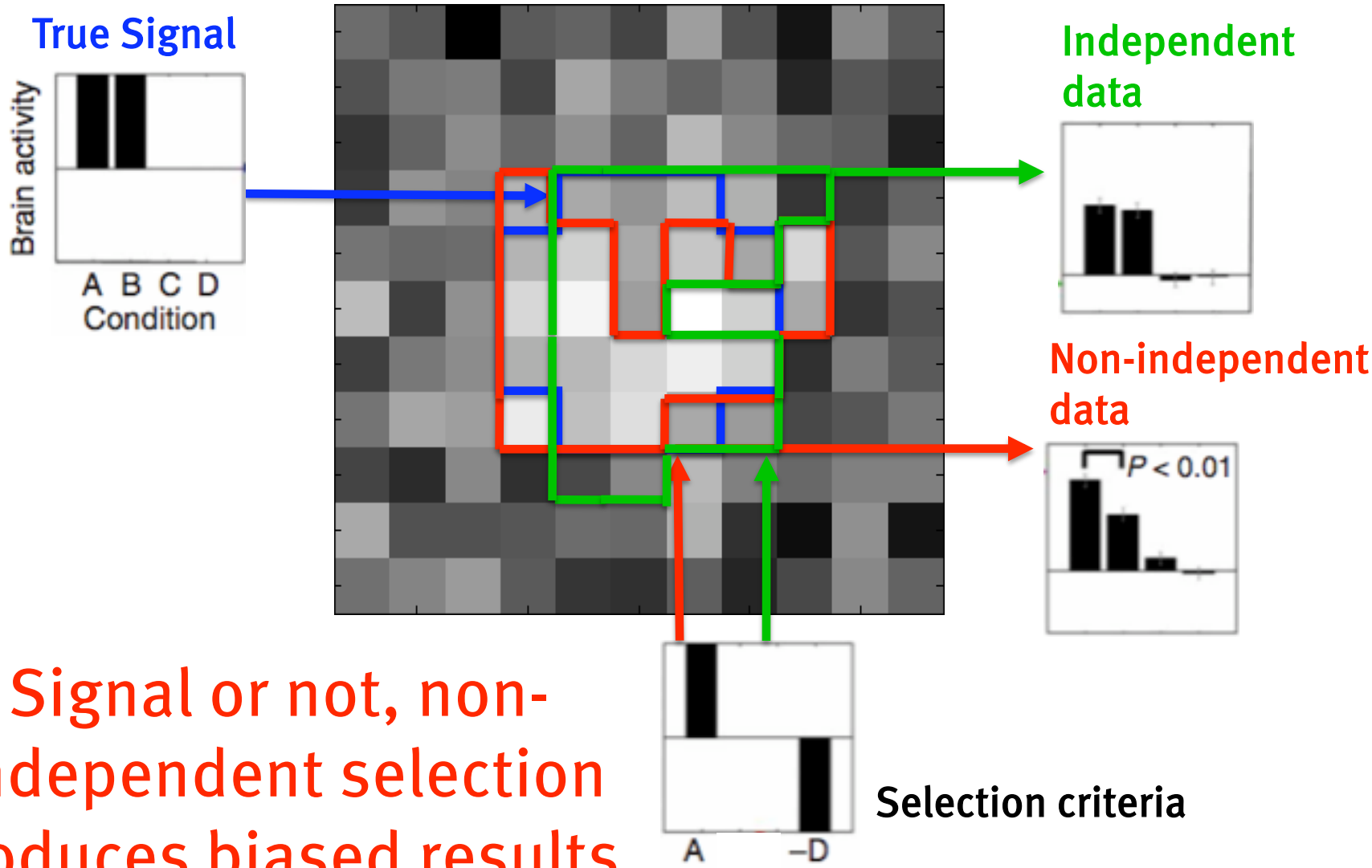
Baker, Hutchison, & Kanwisher (2007)



This analysis produces high classification accuracy from pure noise.

Kriegeskorte, Simons, Bellgowan & Baker (2009)

And if there is some signal?



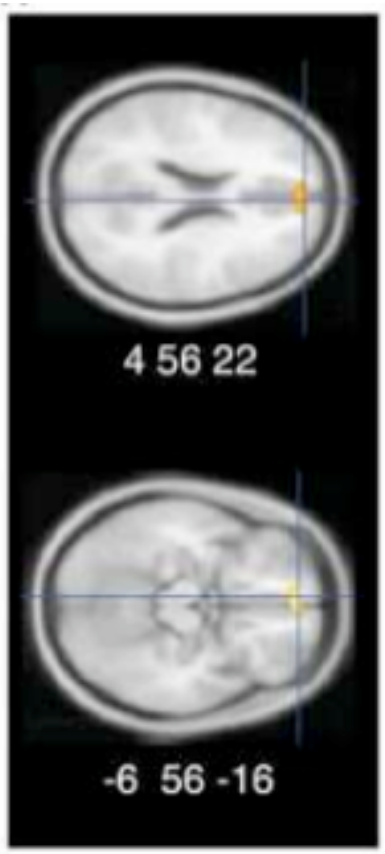
Signal or not, non-independent selection produces biased results.

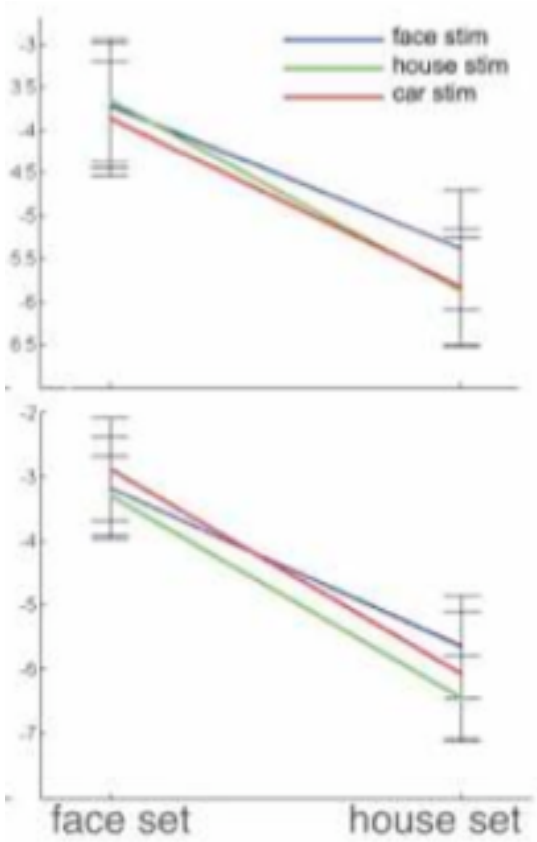
Adapted from Kriegeskorte, Simons, Bellgowan & Baker (2009)

Same problem in different guises

- Null-hypothesis tests on non-independent data.
- Computing meaningless effect sizes (e.g., correlations).
- Showing uninformative/misleading graphs… (with error bars?)

Testing null-hypothesis





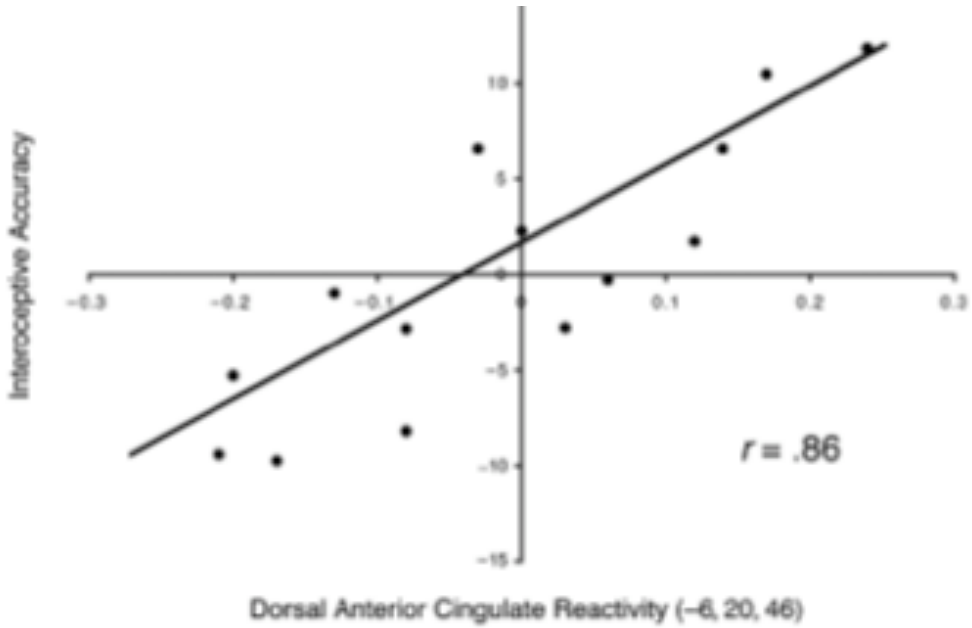
$p < 0.005$
uncorrected

peak voxel:
$F_{(2,14)} = 22.1$;
$p < .0004$

Significant results, potentially out of nothing!

Summerfield et al, 2006

Computing effect sizes



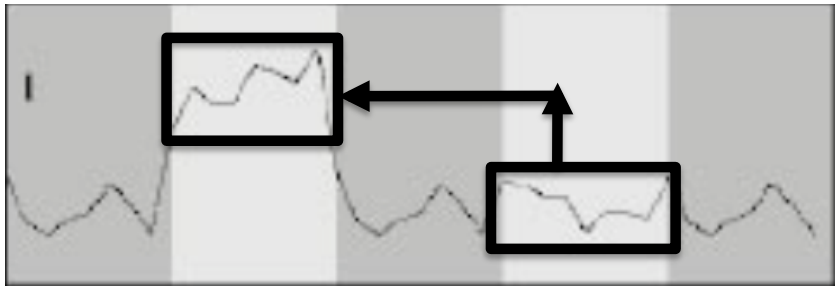
"$r^2 = 0.74$"

A "voodoo" correlation

Eisenberger, Lieberman, & Satpute (2005)

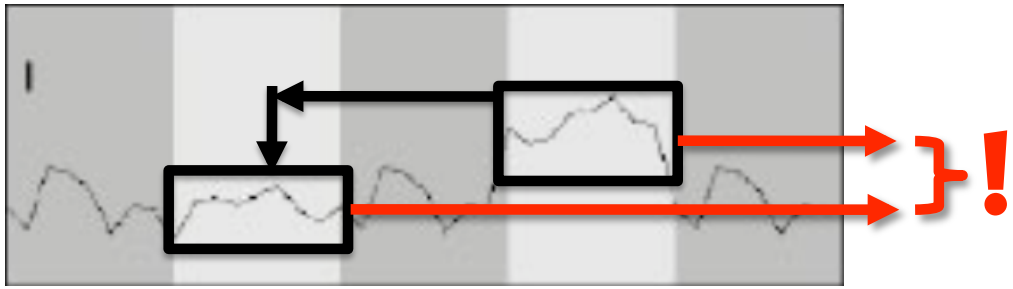
Modulating motion perception

Low load



Motion No motion

High load

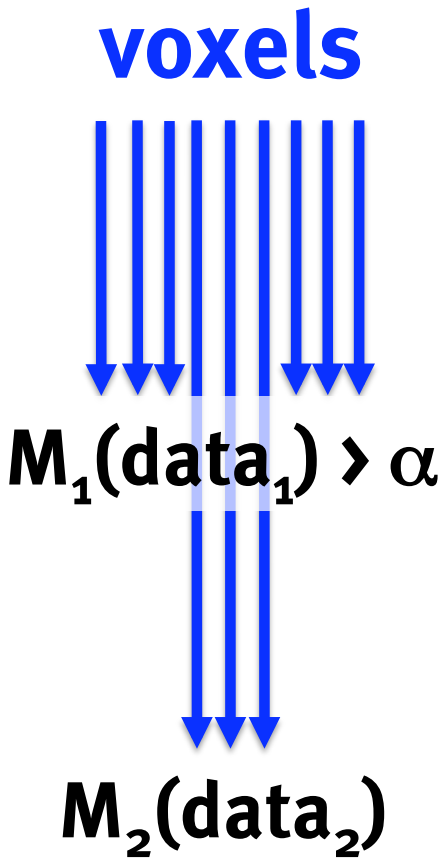


Motion No motion

Non-independent selection produces bizarre spurious patterns

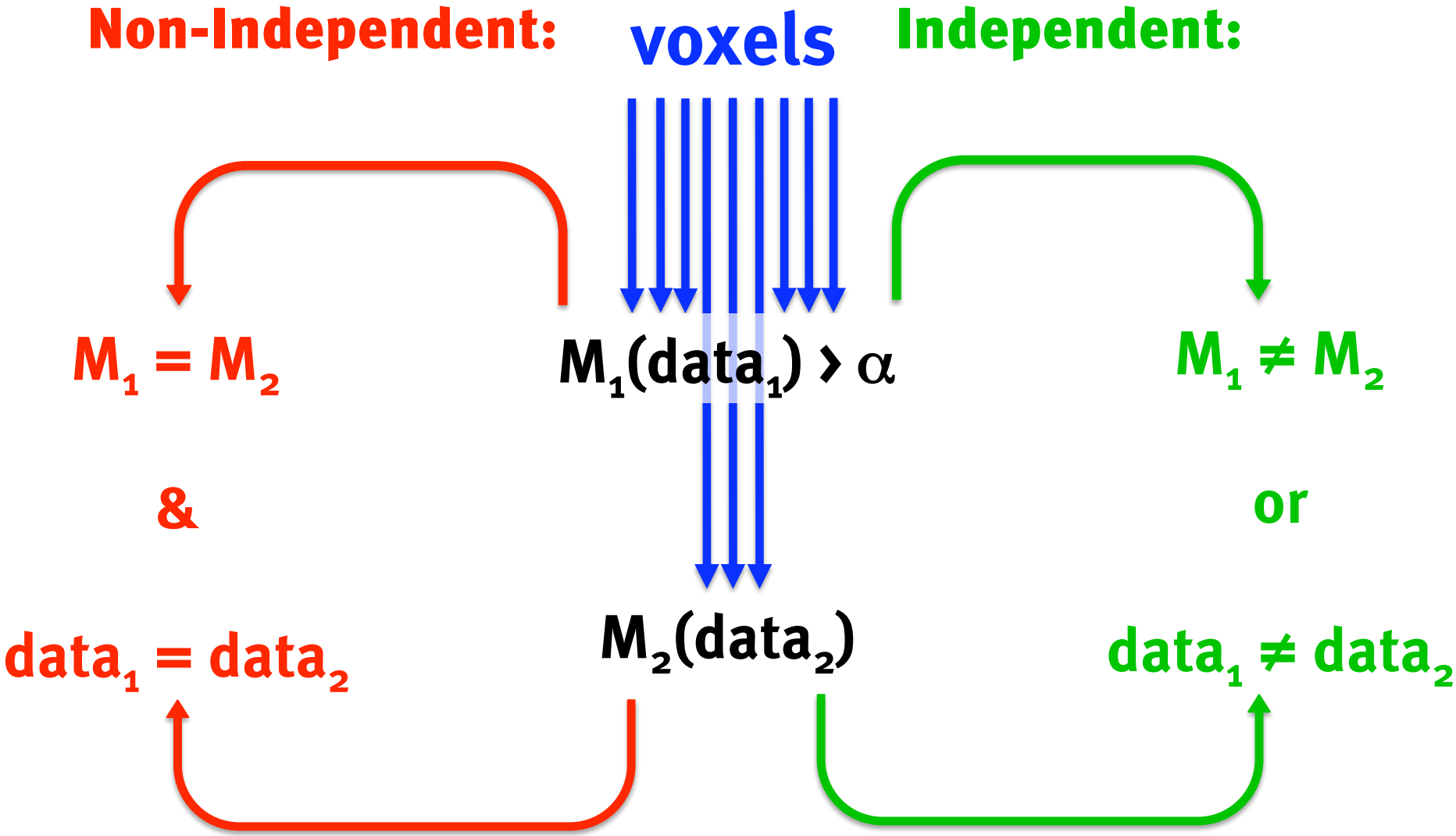
Rees, Frith, & Lavie (1997)

Data Analysis



Reporting peak voxel, cluster mean, etc. amounts to two analysis steps!

Data Analysis

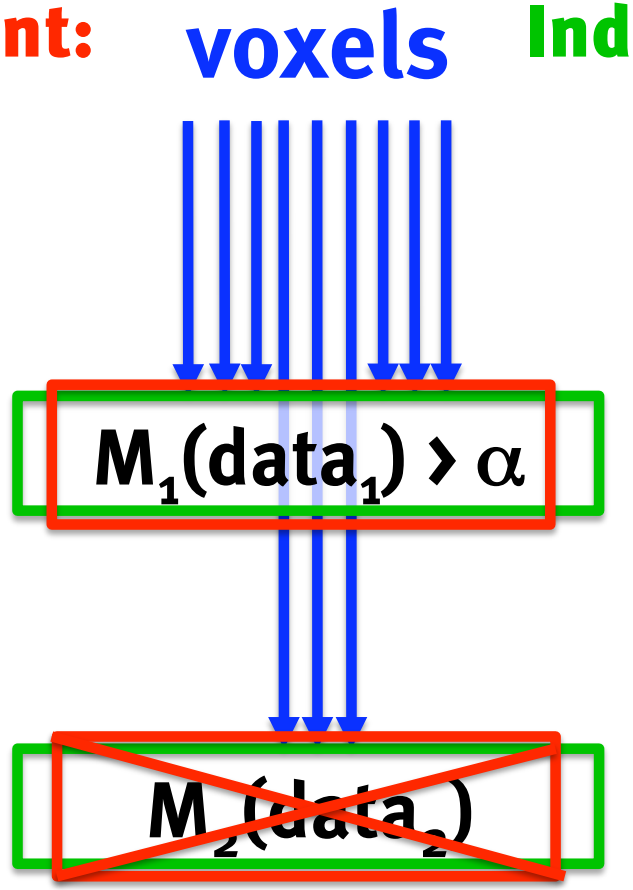


Data Analysis

Non-Independent:

Conclusions can only be based on selection step

M_2 is biased



Independent:

Conclusions can be based on both measures

M_2 provides unbiased estimates of effect size

Some fields are more susceptible

- fMRI, genetics, finance, immunology, etc.
- A whole lot of data are available
- Only some contains relevant signal
- Goals:
  - Find the relevant bits of data (brain, genes, etc.)
  - Estimate their signal
  - Without extra costs of data
- Data are reused, non-independence is common.

Varieties of "voodoo"

Historical perspective

$r = 0.96$

What captured our attention?

The "voodoo correlation" method and generalizations

Objections

! Solutions

Inflation isn't so bad

- Cannot estimate inflation in general – must reanalyze data.
- We know of one comparison (in real data) of independent – non-independent correlations

~0.8 down to ~0.5

(non-indep. r^2 would be inflated by more than a factor of 2)

run 1			run 2			run 3		
voxels	test r	train r	voxels	test r	train r	voxels	test r	train r
257	0.563	0.761	216	0.558	0.872	216	0.558	0.872
276	0.228	0.783	304	0.676	0.767	304	0.676	0.767
311	0.486	0.760	377	0.606	0.788	377	0.606	0.788
331	0.473	0.875	473	0.724	0.85	473	0.724	0.850
346	0.614	0.827	590	0.766	0.812	590	0.766	0.812
492	0.329	0.861	698	0.466	0.810	698	0.466	0.81
498	0.470	0.793	806	0.677	0.822	806	0.677	0.822
634	0.551	0.825	829	0.728	0.845	829	0.728	0.845
711	0.666	0.787	1341	0.510	0.748	1341	0.510	0.748
1135	0.552	0.806	2151	0.282	0.808	2151	0.282	0.808
bias	**0.315**			**0.213**			**0.337**	

Poldrack and Mumford (submitted)

Location, not magnitude, matters

- Do we want to study the brain for its own sake?
  - Only location matters, but…

    Localization seems to be overestimated by underpowered studies: Studies with adequate power yield small widespread correlations
    (Yarkoni and Braver)
- Do we want to use measures/manipulations of the brain to predict/alter behavior?
  - Effect size really matters!

Correlations not used for inference?

In the present study, the first neuroimaging study to include an outcome measure against which both self-reported and neurally assessed neuroticism could be compared, it was found that dACC reactivity was a substantially better predictor of interoceptive accuracy than self-reported neuroticism was, accounting for nearly five times the variance in interoceptive accuracy (74% vs. 16%). With the utilization of these types of methods, future personality research may have the potential to account for a substantially larger portion of the variation in human experience and behavior than has been possible with self-report measures alone.

Eisenberger, Lieberman, & Satpute (2005)

Varieties of "voodoo"

Historical perspective

$r = 0.96$

What captured our attention?

The "voodoo correlation" method and generalizations

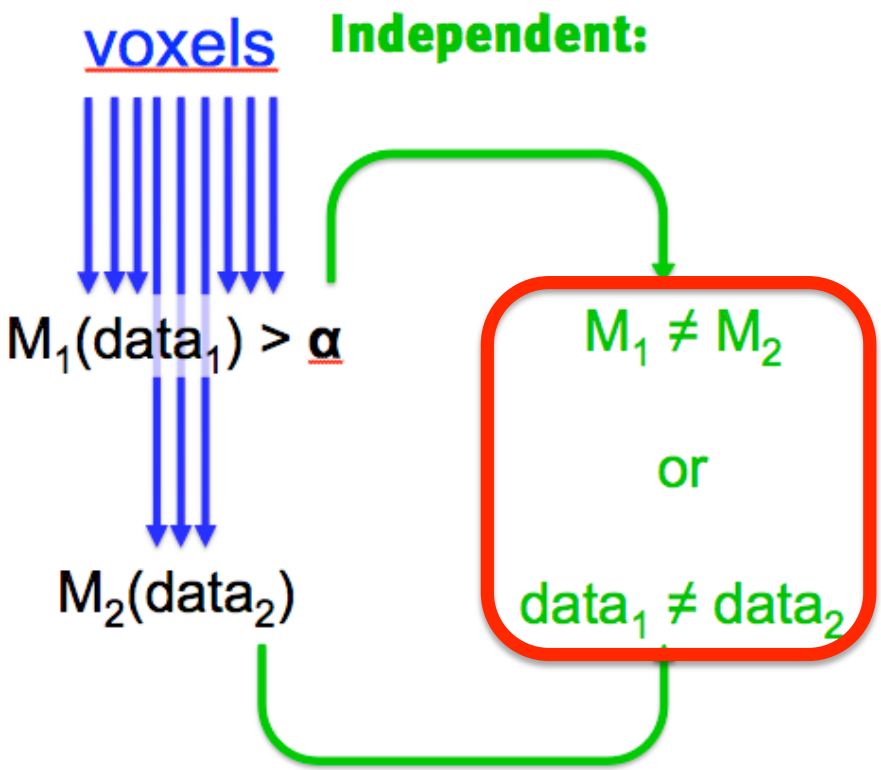
Objections

! Solutions

Positive suggestions

1: Independent localizers

Use different data or orthogonal measure to select voxels.



Positive suggestions

1: Independent localizers

2: Cross-validation methods

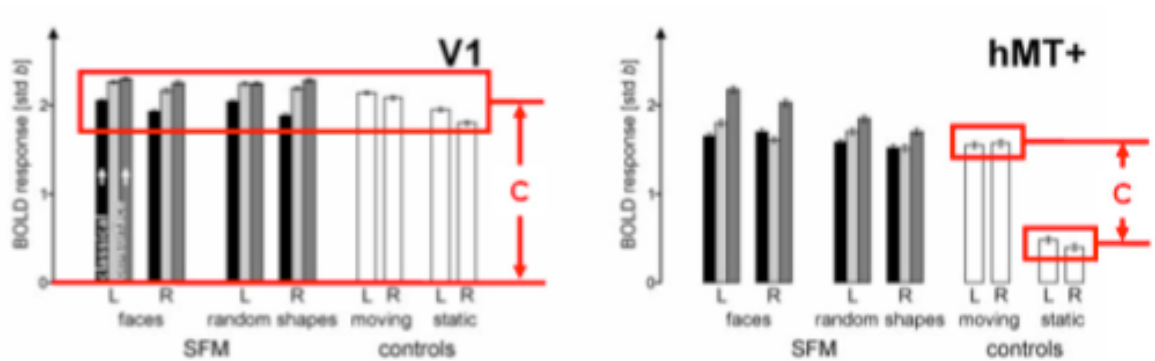
Use one portion of data to select, another to obtain a validation measure, repeat.

Positive suggestions

1: Independent localizers

2: Cross-validation methods

3: Be careful with graphs



From Kriegeskorte, Simons, Bellgowan and Baker (2009)

Positive suggestions

1: Independent localizers

2: Cross-validation methods

3: Be careful with graphs

4: Random/Permutation simulations to evaluate independence

If orthogonality of measures or independence of data are not obvious, try the analysis on random data
(Generates empirical null-hypothesis distributions.)

Closing

- Reusing the same data and measure to
  (a) select data and
  (b) estimate effect sizes leads to overestimation:
  "Baloney" – Cureton (1950)
- Most high correlations in individual differences research in fMRI are obtained in this manner:
  "Voodoo"
- Many fMRI papers contain variants of this error
  42% in 2008 (Kriegeskorte et al, 2009)

Simple solutions exist to avoid these problems!

Thanks!

Thanks to:
Nancy Kanwisher, Hal Pashler,
Chris Baker, Niko Kriegeskorte
and many others for
very engaging discussions.

Vul, Harris, Winkielman, Pashler (2009) Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition, *Perspectives on Psychological Science.*

Vul, Kanwisher (in press) Begging the question: The non-independence error in fMRI Data Analysis, *Foundational issues for human brain mapping.*

Cureton (1950) Validity, Reliability, and Baloney. *Educational and Psychological Measurement*.

Kriegeskorte, Simmons, Bellgowan, Baker (2009) Circular analysis in systems neuroscience: the dangers of double dipping, *Nature Neuroscience*.

Poldrack, Mumford (submitted) Independence in ROI analyses: Where's the voodoo*?*

Lieberman, Berkman, Wager (2009) Correlations in social neuroscience aren't voodoo..., *Perspectives on Psychological Science*.

Vul, Harris, Winkielman, Pashler (2009) Reply to comments on "Puzzlingly high correlations...", *Perspectives on Psychological Science.*

Yarkoni (2009) Big correlations in little studies..., *Perspectives on Psychological Science*.

Yarkoni, Braver (in press) Cognitive neuroscience approaches to individial differences in working memory... *Handbook of individual differences in cognition.*

Nicholas, Poline (2009) Commentary on Vul et al..., *Perspectives on Psychological Science*.