Statistical inference for brain imaging

Bertrand Thirion

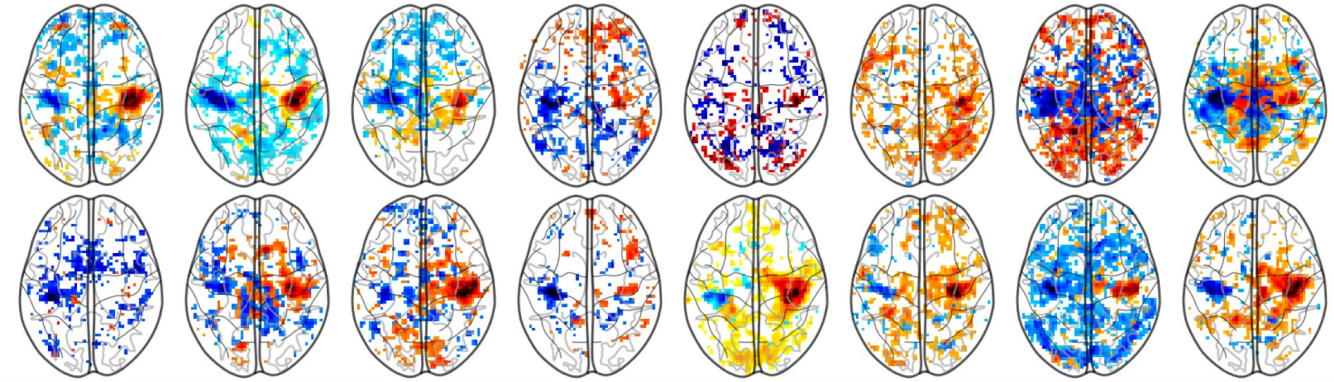
MIND team, Inria Saclay, CEA Neurospin
bertrand.thirion@inria.fr

Outline

- Statistical inference: the group-level setting
- Statistical inference: P-values
- Multiple comparisons: FDR, FWER
- Control of false discovery proportion (FDP)
- Multivariate models

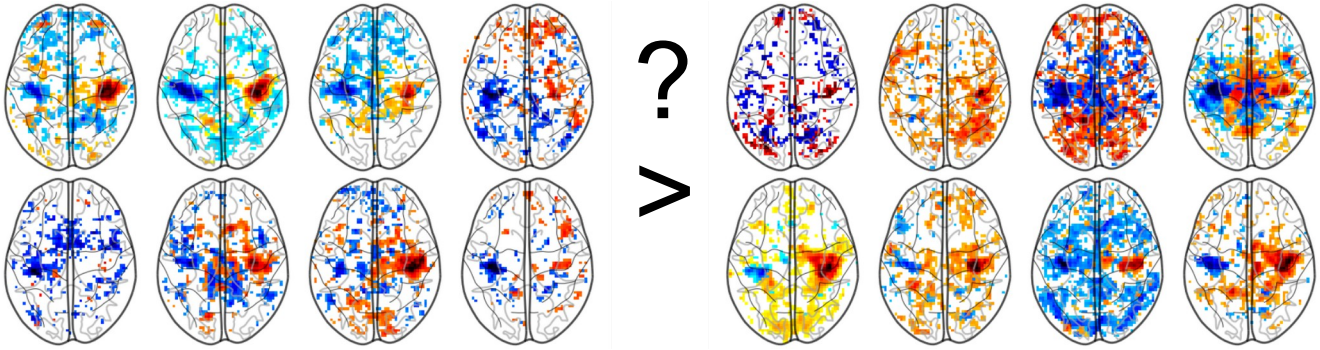
Problem statement

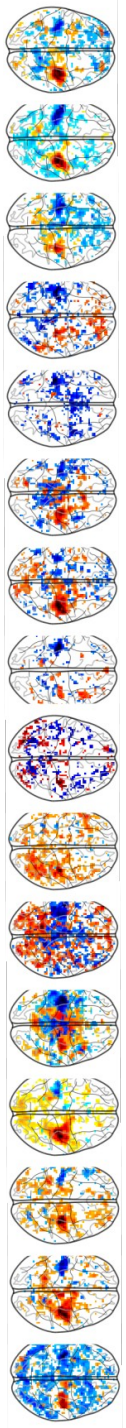
- One-sample test

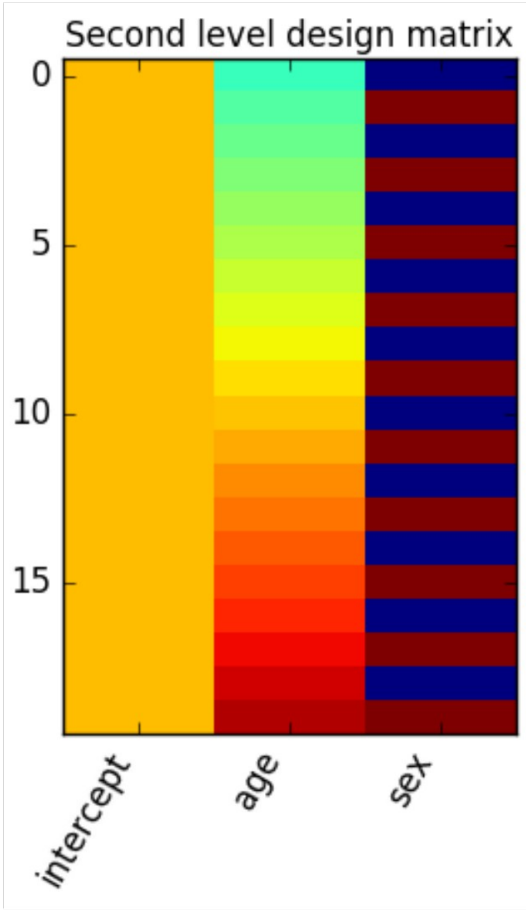


? > 0

- Two-sample test







Problem statement

What is the effect of age / sex on the activations ?

- brain/behavior correlations
- brain/genetics correlations
- diagnosis

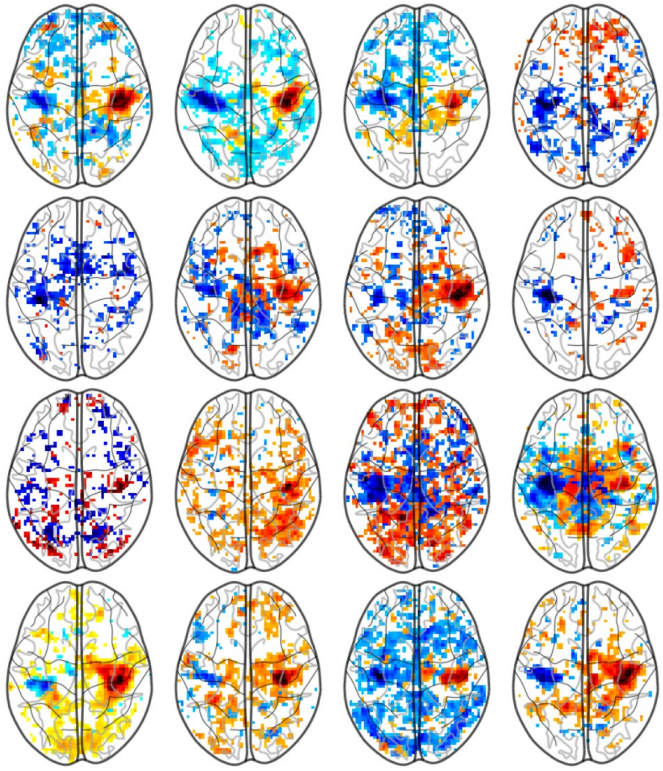
The limitations of spatial registration

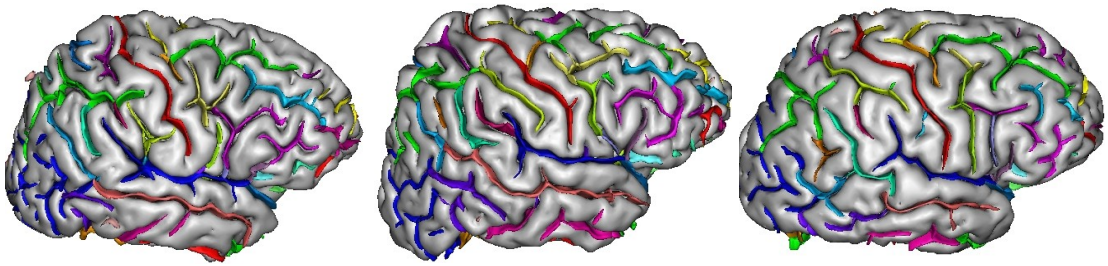
- Analysis on populations : multi subjects studies

  → Realignment to a common template

  - MNI/Talairach coordinate system
  - Assume the spatial coordinates match the functional architecture

- Problem: inter-subject functional and anatomical variability

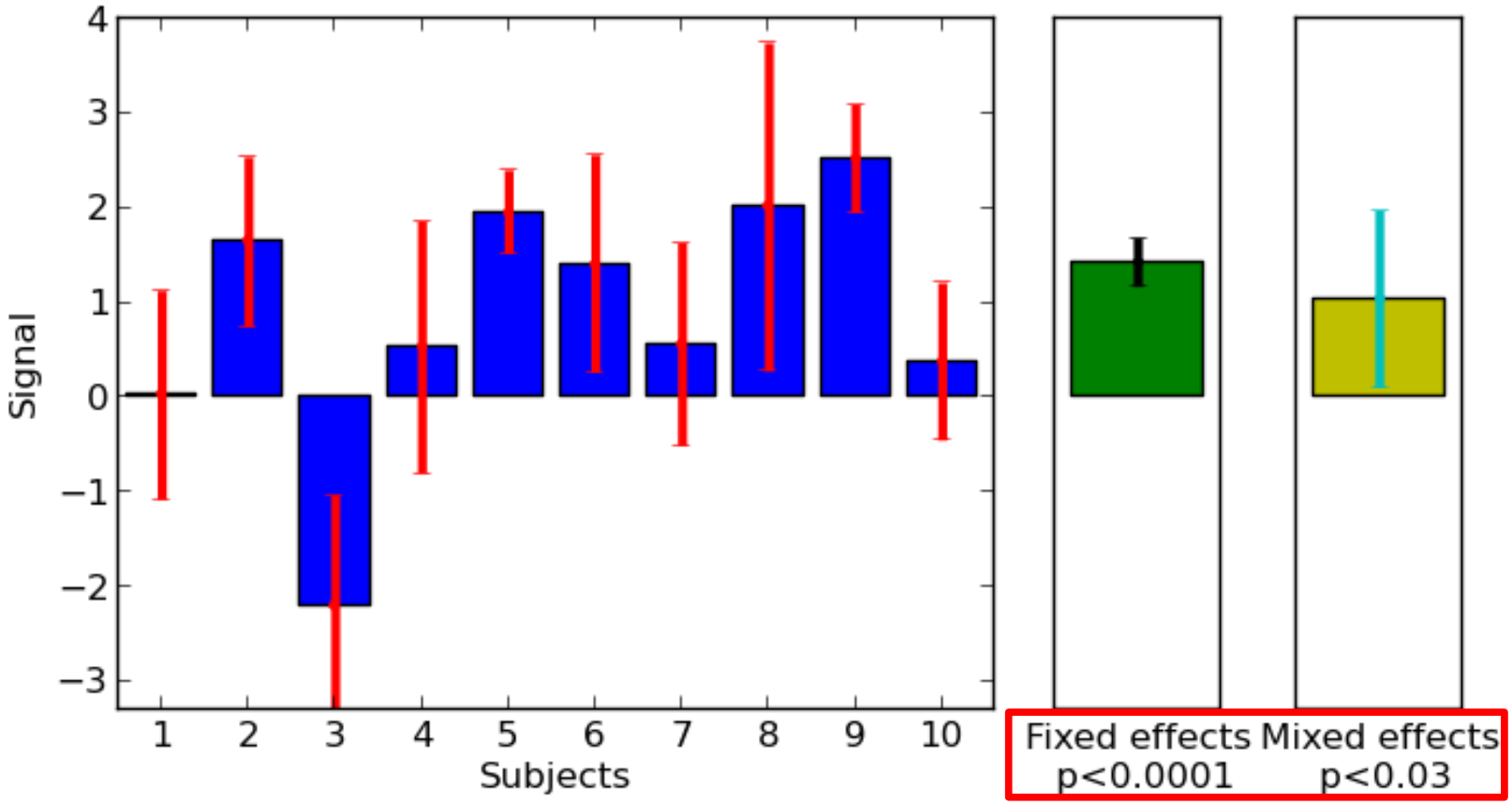




Fixed- versus random-effects analysis

- Do we want to make an inference about
  - The particular group we sampled? (FFX)
  - The population they were sampled from? (RFX)
- These correspond to different null hypotheses:
  - FFX: (H0) “all subjects are inactive”
  - RFX: (H0) “the mean effect is zero”

Fixed- versus random-effects analysis



Distribution of each subject's estimate

Distribution of the mean estimate

Parametric *t*-test

- Compute the *t*-statistic, $t = \frac{\hat{\mu}}{\hat{\sigma}\sqrt{n}}$

$$\text{with } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\beta_i,\ \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\beta_i - \mu)^2$$

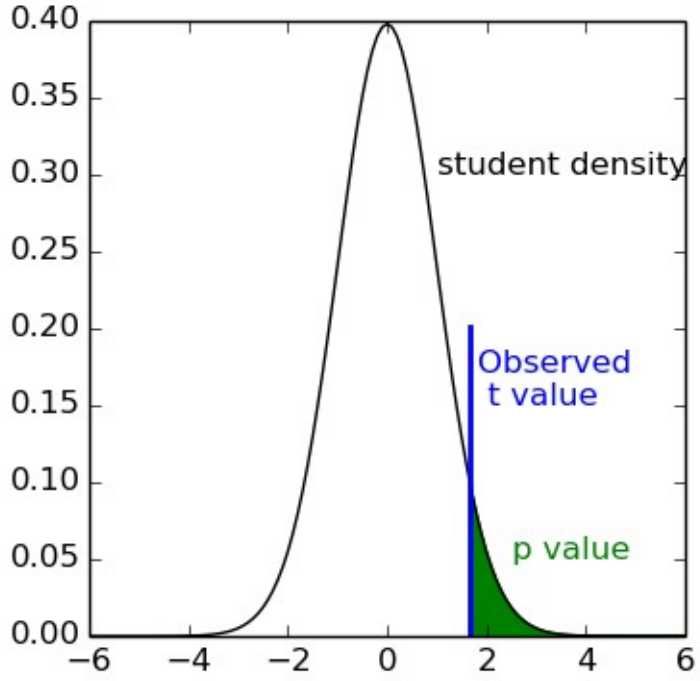
- Reject null hypothesis of zero mean if

$$T_{n-1} \geq 1 - \alpha$$

T_{n-1} : cumulative Student distribution with *n-1* degrees of freedom

α: accepted rate of false positives

$$f_n(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$



Multiple Comparison problem

- Specificity = α for one test.
- But we perform n tests (10^4 to 10^5)
- Probability of one false detection (assuming independence):

$$1 - (1 - \alpha)^n \simeq n\alpha \text{ if } n\alpha \ll 1$$

- Correct the significance of test by n: *Bonferroni correction*
- Very conservative inference, little power to detect true effects

Multiple comparisons

- Notations

Number	Truly active voxels	Inactive voxels
$t > t_\alpha$	tp	fp
$t \leq t_\alpha$	fn	tn

- False positive rate $\alpha = \frac{fp}{fp + tn} < \frac{fp}{V}$

Multiple comparisons

- Notations

Number	Truly active voxels	Inactive voxels
$t > t_\alpha$	tp	fp
$t \leq t_\alpha$	fn	tn

- False positive rate

$$\alpha = \frac{fp}{fp + tn} < \frac{fp}{V}$$

- Family-wise error rate

$$fwer = p(fp > 0)$$

Multiple comparisons

- Notations

Number	Truly active voxels	Inactive voxels
$t > t_\alpha$	tp	fp
$t \leq t_\alpha$	fn	tn

- False positive rate
- Family-wise error rate
- False discovery rate

$$\alpha = \frac{fp}{fp + tn} < \frac{fp}{V}$$

$$fwer = p(fp > 0)$$

$$fdr = \frac{fp}{fp + tp}$$

False Discovery Rate

Rationale

Control the number of false positives as a proportion q of the number of detected voxels

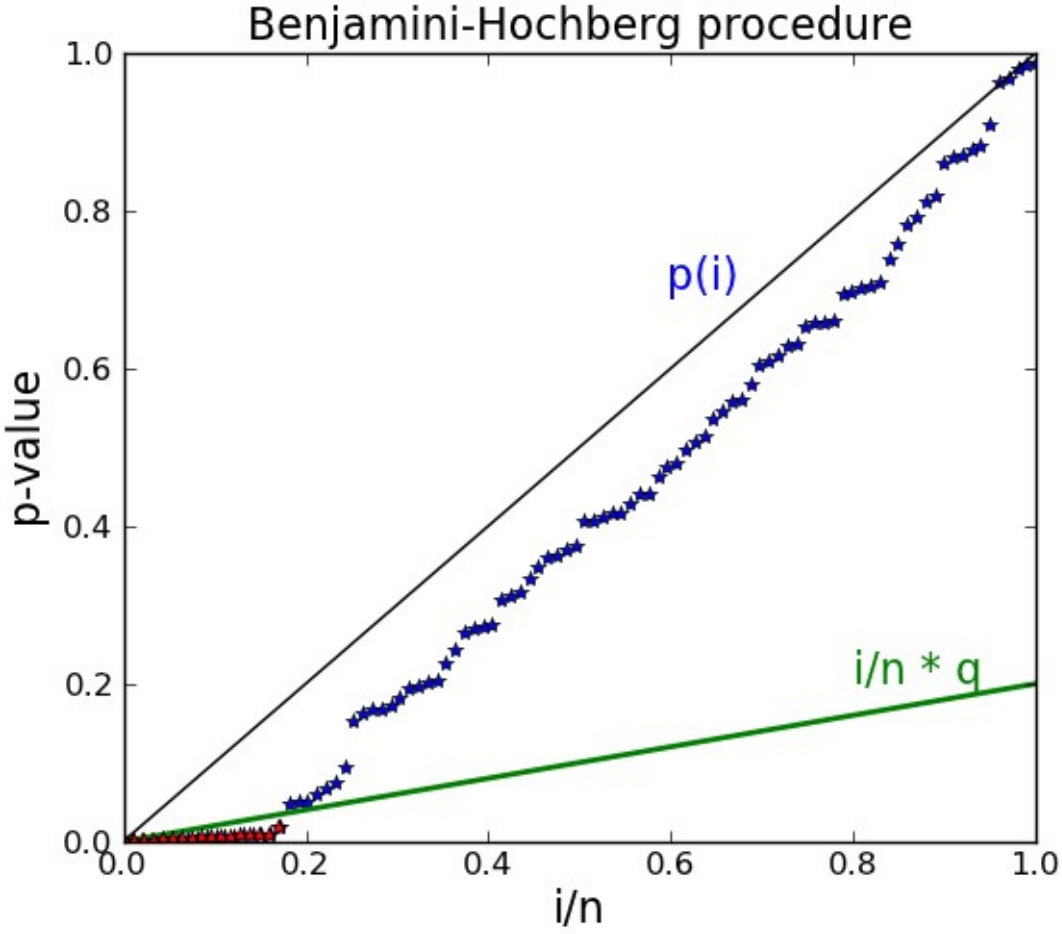
[Benjamini Hochberg 1995]

[Genovese NeuroImage 2002]

Method

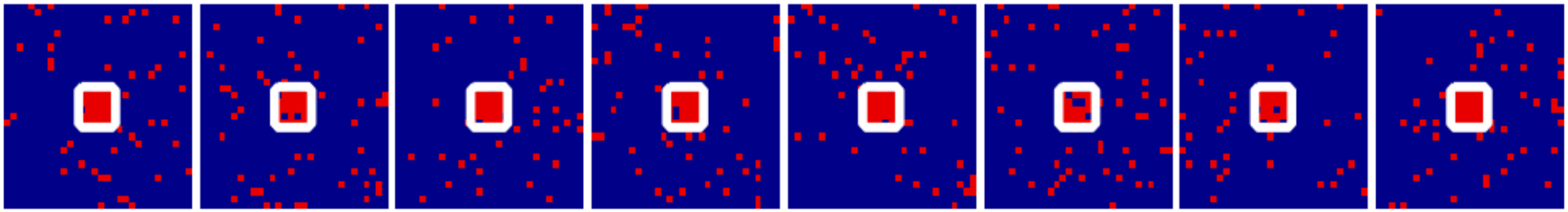
Sort the p-values

$$i^{\star} = argmax_{i \in [1,n]} p(i) < \frac{i}{n} q$$

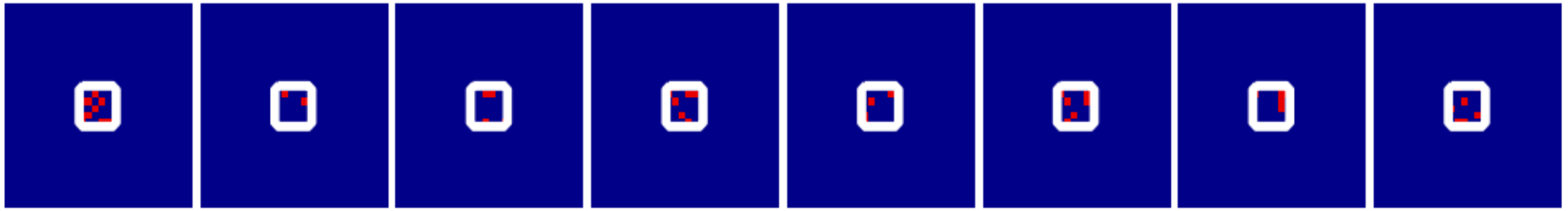


FPR, FWER, FDR

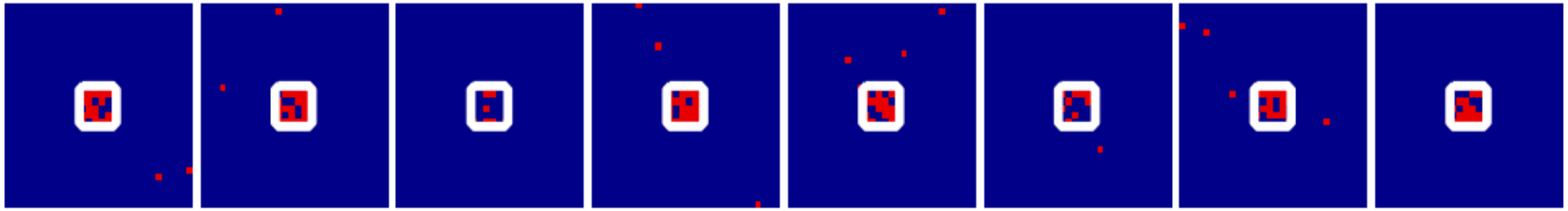
FPR <= 5%



FWER <= 5%



FDR 5%



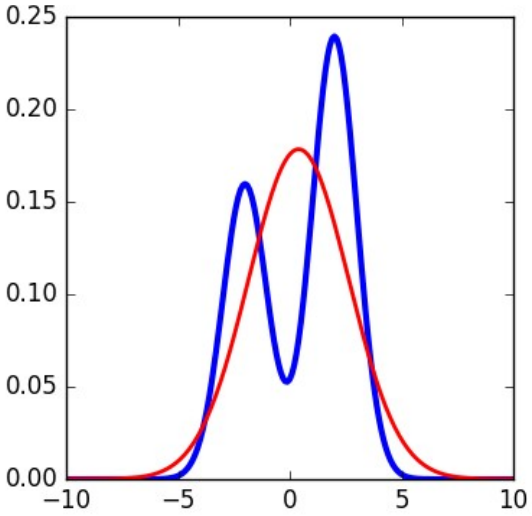
[Genovese et al. Nimg 2002]

Shortcomings

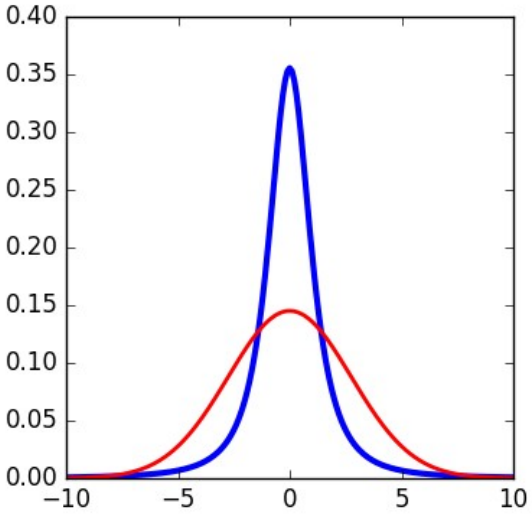
- Parametric *t*-test → assumes observations drawn independently from the same *normal* law
- If assumptions do not hold,
  - Biased *specificity* (inaccurate false positive rate)
  - Lack of *sensitivity*
- Problems are severe for small samples
  - Normality is not crucial for large samples
  - W.S. Gosset derived the Student distribution in 1908 as a large sample approximation of a permutation test!

Inference on non-Gaussian data

Non-Gaussian densities



Sub-Gaussian
Bi-modal mixture
of Gaussians



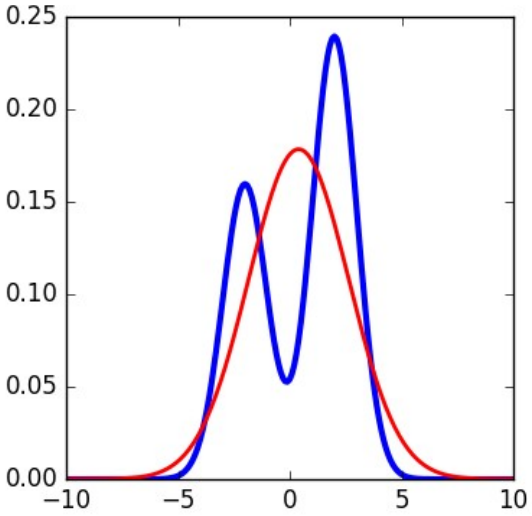
Super-Gaussian
Student with
low dofs

One-sample test (is the mean larger than zero ?)

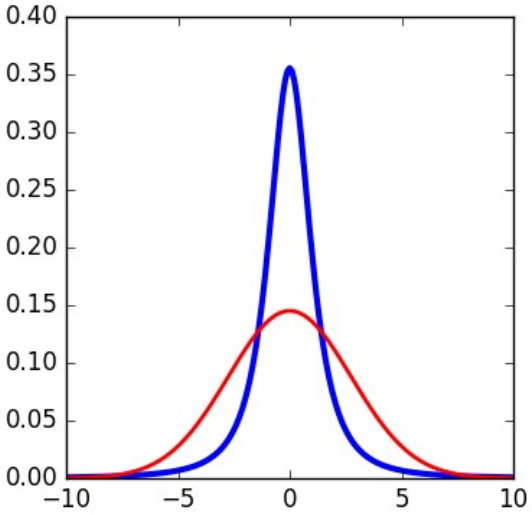
$p(T > T(X) \mid \mu = 0)$?

Inference on non-Gaussian data

Non-Gaussian densities

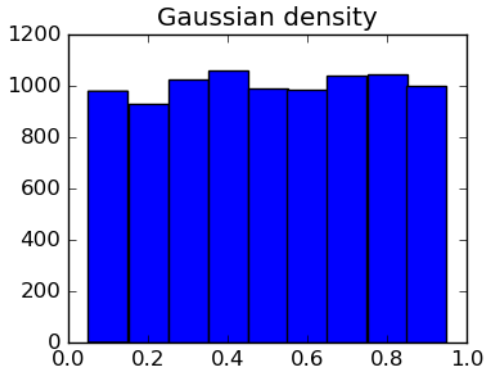


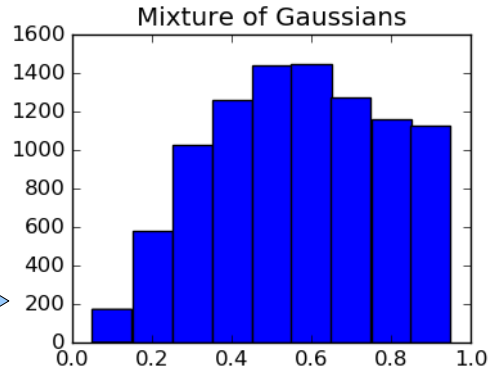
Sub-Gaussian
Bi-modal mixture
of Gaussians

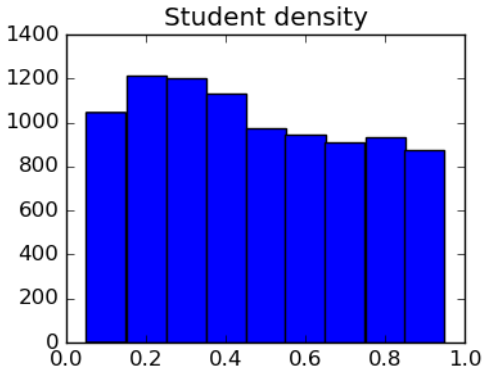


Super-Gaussian
Student with
low dofs

One-sample test (is the mean larger than zero ?)

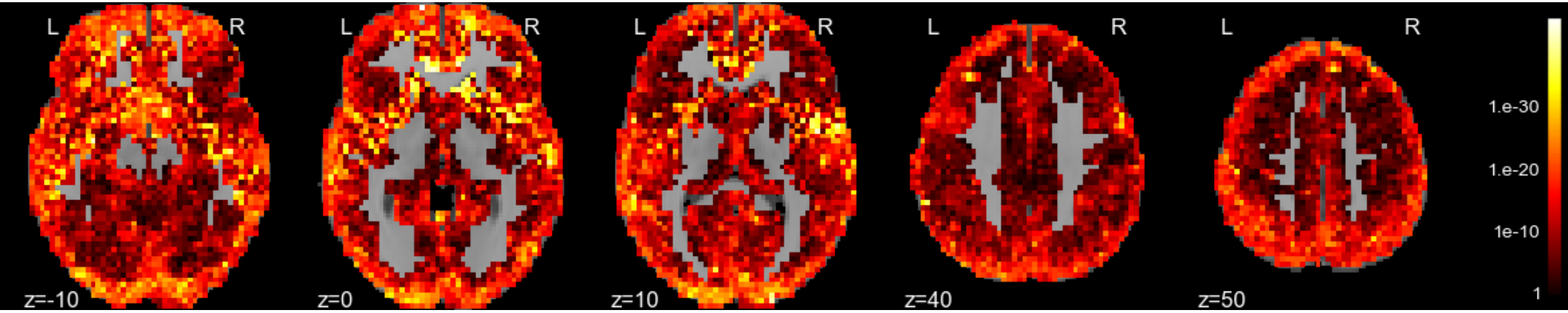






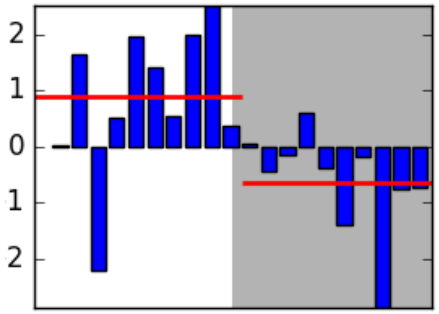
$p(T > T(X) \mid \mu = 0)$?

Population-level fMRI data are non-Gaussian



A standard normality test rejects uniformly the null hypothesis:
"data are Gaussian distributed"
in a large sample of subjects (n=543)

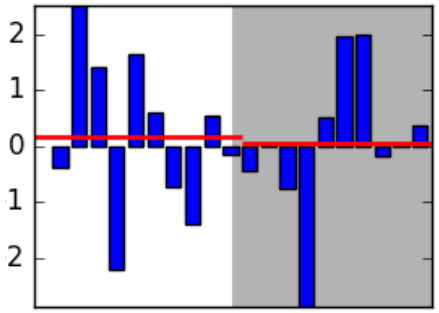
Permutation for two-sample tests



t=2.87

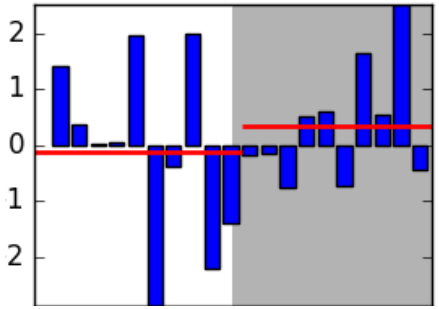
Is the difference significant ?

Shuffle the sample and re-compute the statistic



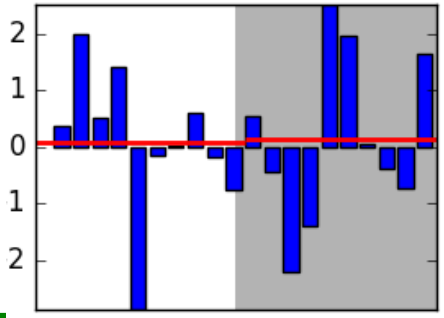
t=0.18

iterate



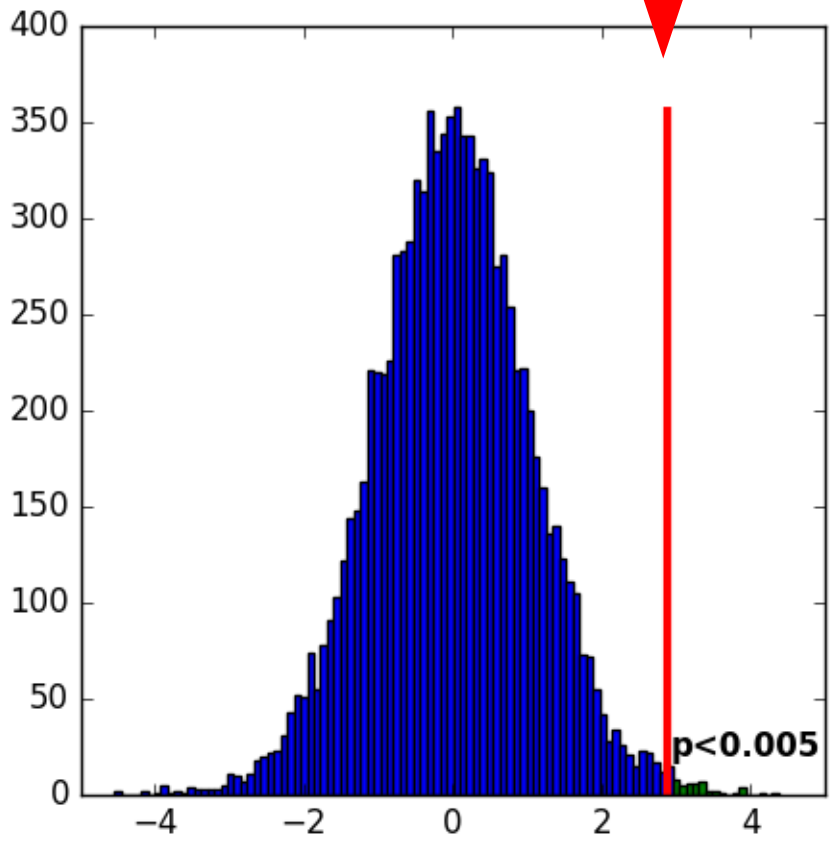
t=-0.74

...

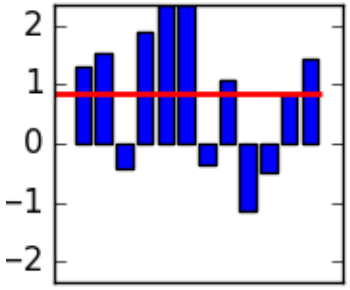


t=-0.09

Generate a histogram of the values



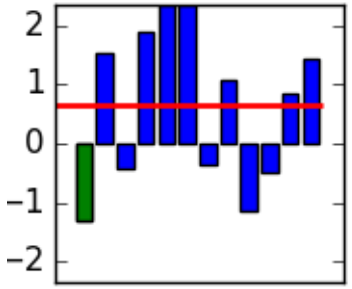
Permutations for one-sample tests



t=2.52

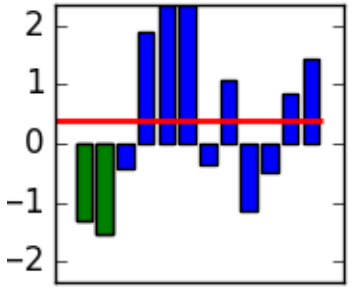
Is the mean significantly > 0 ?

swap the sign of 1 observation and re-compute the statistic



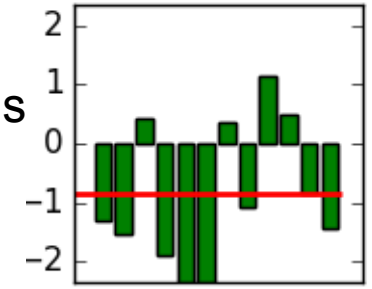
t=1.68

swap the sign of 2 observations and re-compute the statistic



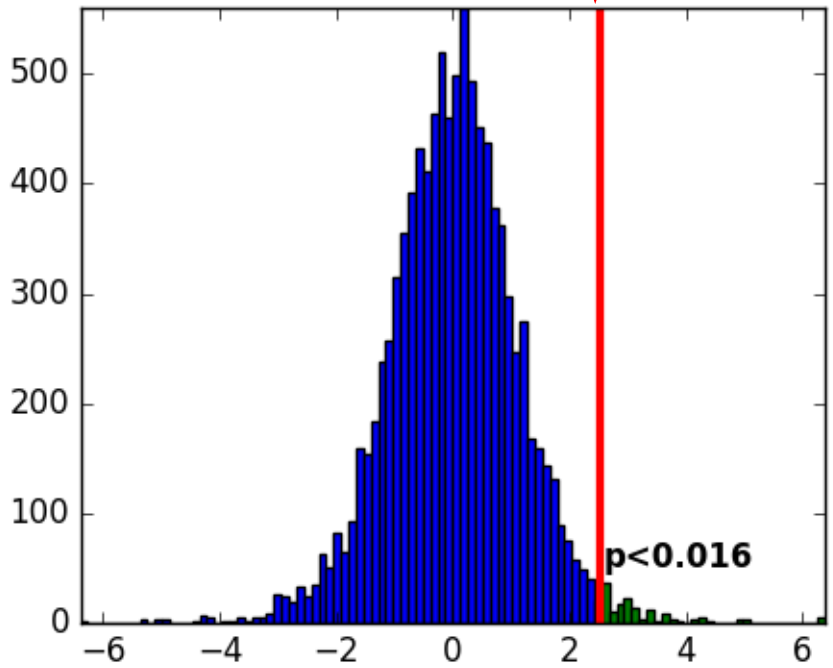
t=0.94

… swap the sign of all observations and re-compute the statistic

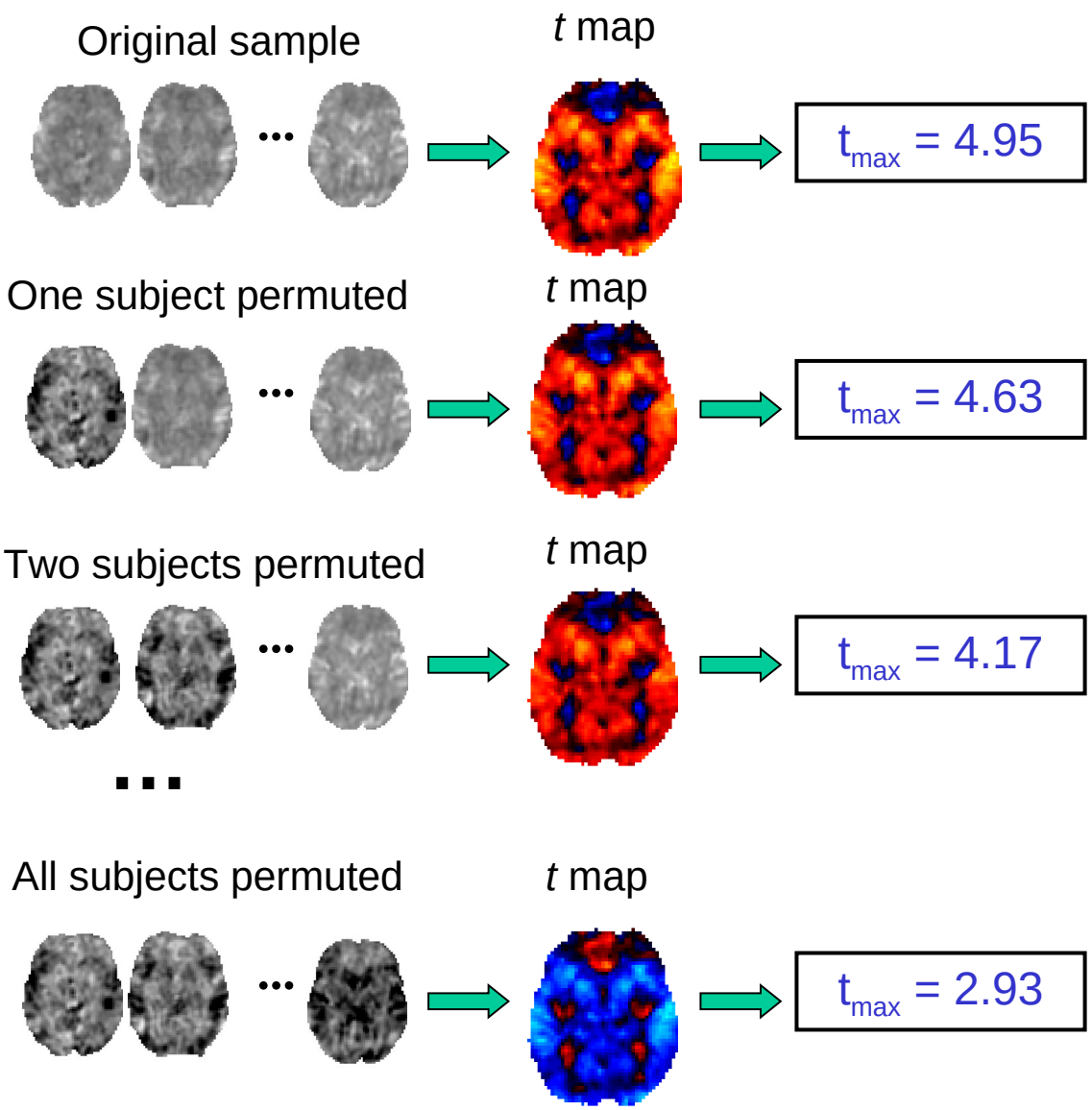


t=-2.52

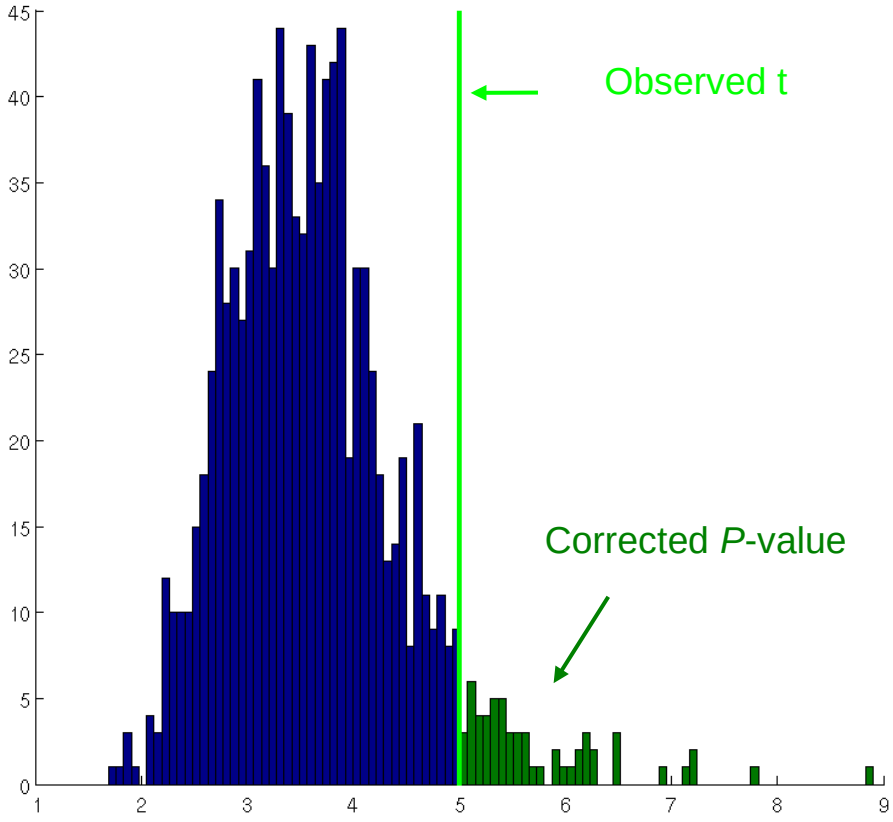
Generate a histogram of the values



Familywise error correction



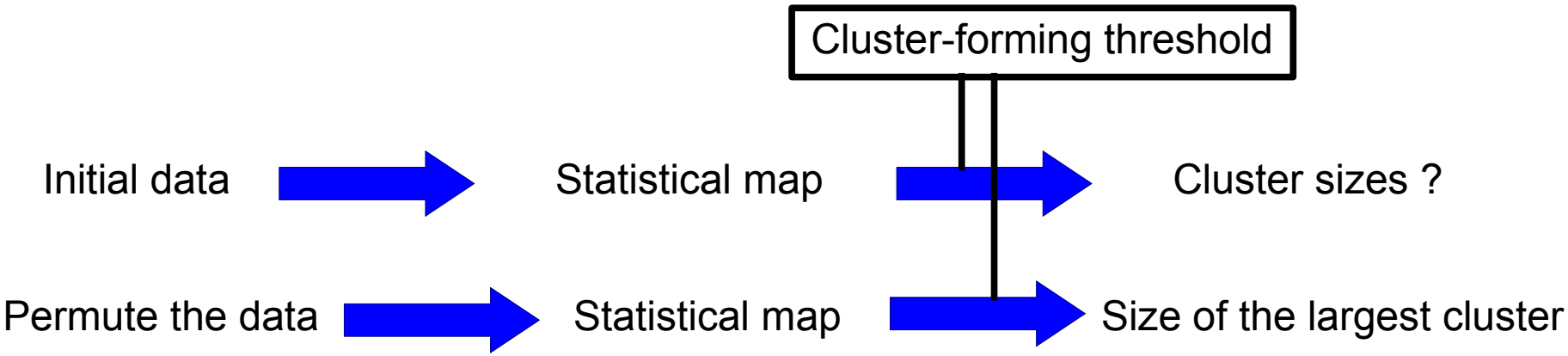
Histogram of the t_{max} statistics

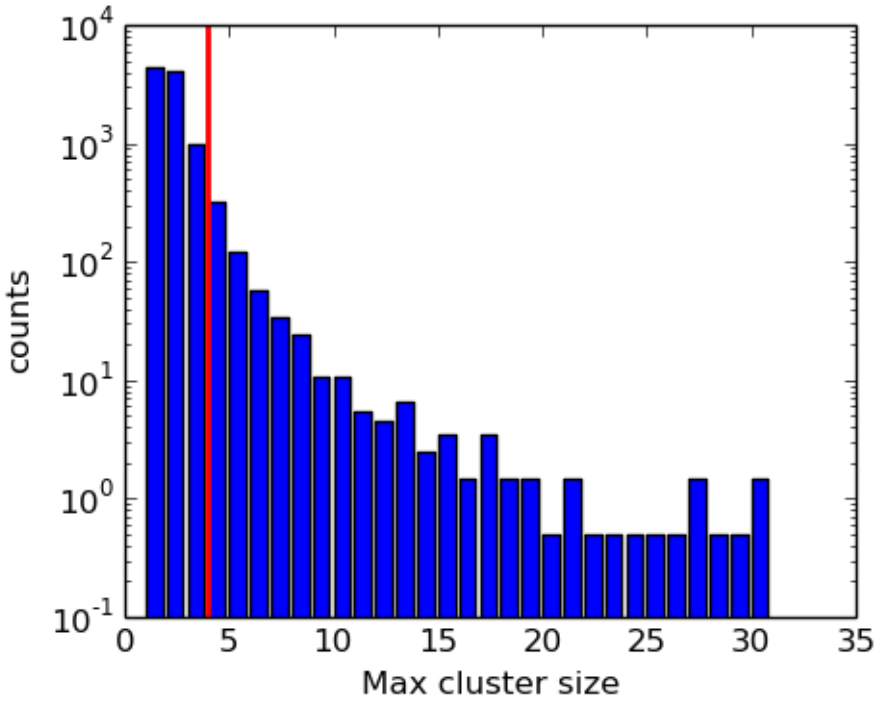


Sign permutation tests: advantages

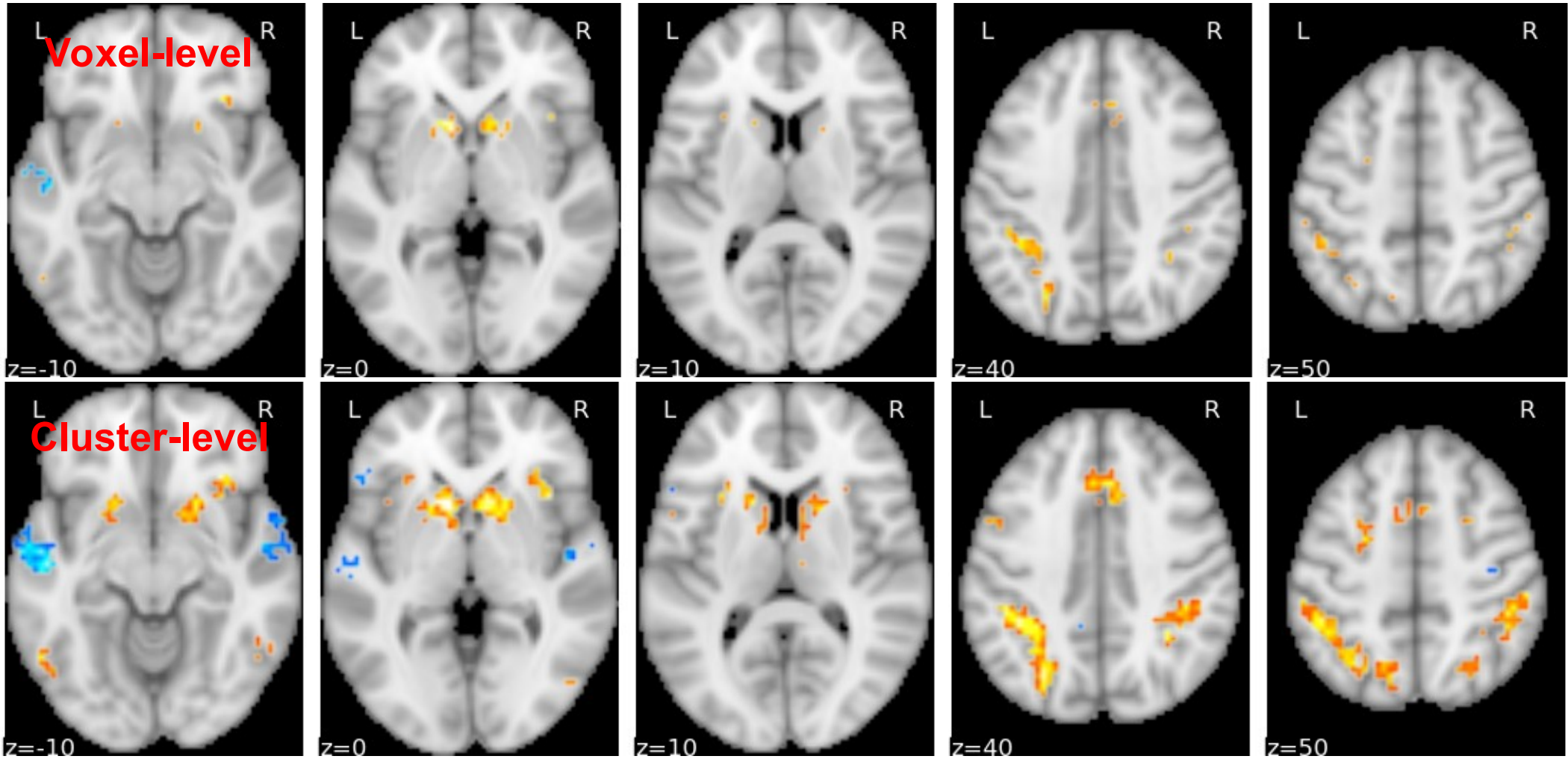
- They produce accurate *P*-values
  - Whatever the chosen test statistic
- They work under mild assumptions
  - One-sided tests require distribution symmetry
  - Yet symmetry is more general than normality
- They manage multiple comparisons
  - Alternative to random field theory

Cluster-level inference



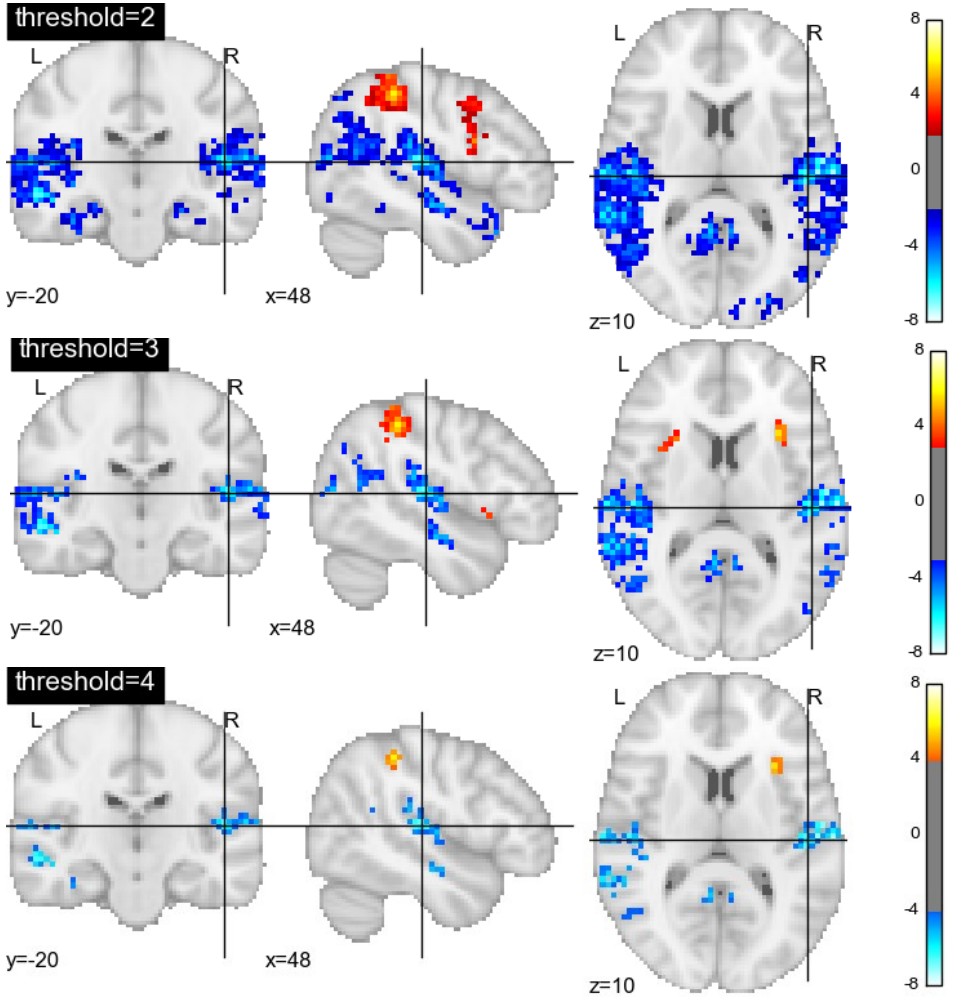


Cluster-level inference increases sensitivity



(p<.05 corrected for multiple comparisons)

Cluster-level detection depends on an arbitrary threshold



- T maps thresholded at t>2, t>3 and t> 4 respectively
- Clusters detected at FWER <0.05

A second look at cluster-level inference

- Large clusters → more likely to discover effects

But

- Large clusters → each cluster weakly informative

Solution: control the false discovery proportion per cluster

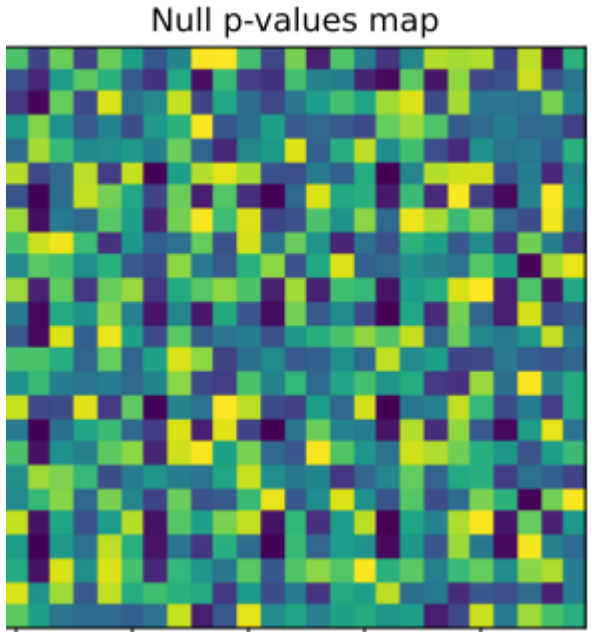
$$FDP(S) = |S \cap H_0| \,/\, |S|$$

Circularity: clusters S are defined from the data.

Solution: Control for all clusters simultaneously (“post-hoc”)

Intermezzo: double dipping

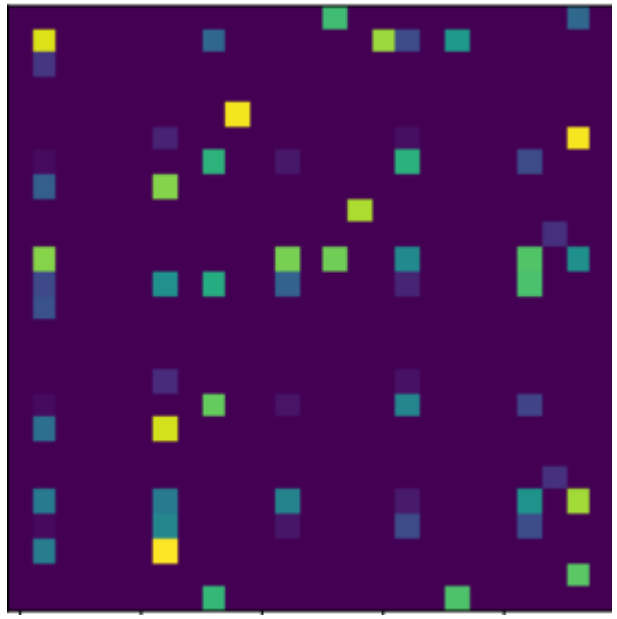
Image of null p-values



Benjamini-Hochberg procedure

No detection

Selection of the 60 lowest p-values

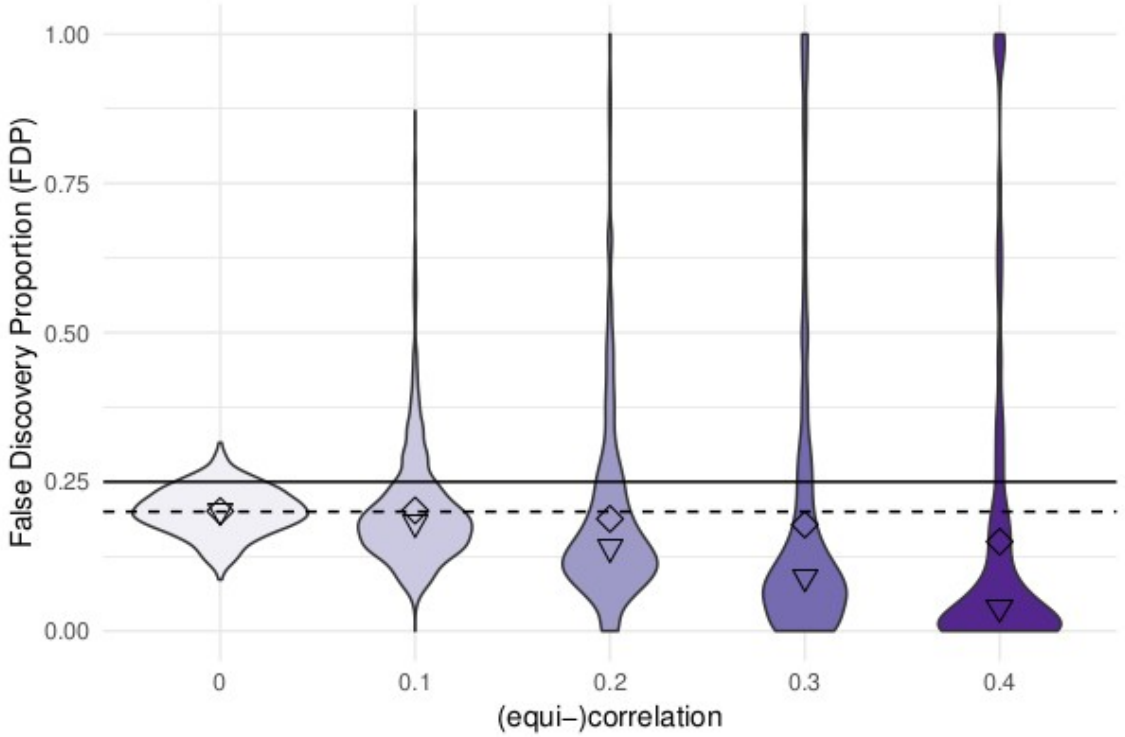


62% of p-values rejected
FDP=100% !

False Discovery Proportion vs False Discovery Rate

$$FDP(S) = |S \cap H_0| \,/\, |S|$$

$FDR = \mathrm{E}(FDP)$



Empirical FDP for a given FDR controlling procedure

FDR control does not imply FDP control [Neuvial et al. HDR 2020]

Controling the FDP: interpolation

- For given α find V: S $\rightarrow$V(S) s.t.

$$P\left(\forall S,\ |S \cap H_0| \leq V(S)\right) \geq 1-\alpha$$

- How to build such a V ?
  - Consider Simes Family, under H_0 and PRDS:

$$\mathbb{P}\left(\exists k \in \{1, \cdots, m\} : p_{(k:m)} < \frac{\alpha k}{m}\right) \leq \alpha$$

  - Possible construction $\mathbf{R}_k = \left\{ i : p_i \leq \frac{\alpha k}{m} = t_k^{Simes} \right\}$

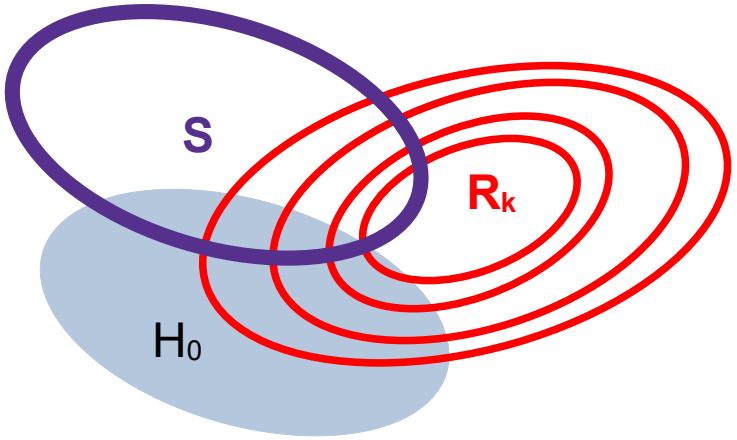
$$\mathbb{P}\left(\forall k, |\mathbf{R}_k \bigcap H_0| \leq k-1\right) \geq 1-\alpha$$

The condition holds for a family ($\mathbf{R}_k$) only, but simultaneously

Controling the FDP: interpolation

- need to generalize from (R_k) to arbitrary sets S

$$
\begin{aligned}
|S \cap H_0| &= \left|S \cap \overline{R_k} \cap H_0\right| + |S \cap R_k \cap H_0| \\
&\leq \left|S \cap \overline{R_k}\right| + |R_k \cap H_0| \\
&\leq \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + |R_k \cap H_0| \\
&\leq \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + k - 1
\end{aligned}
$$

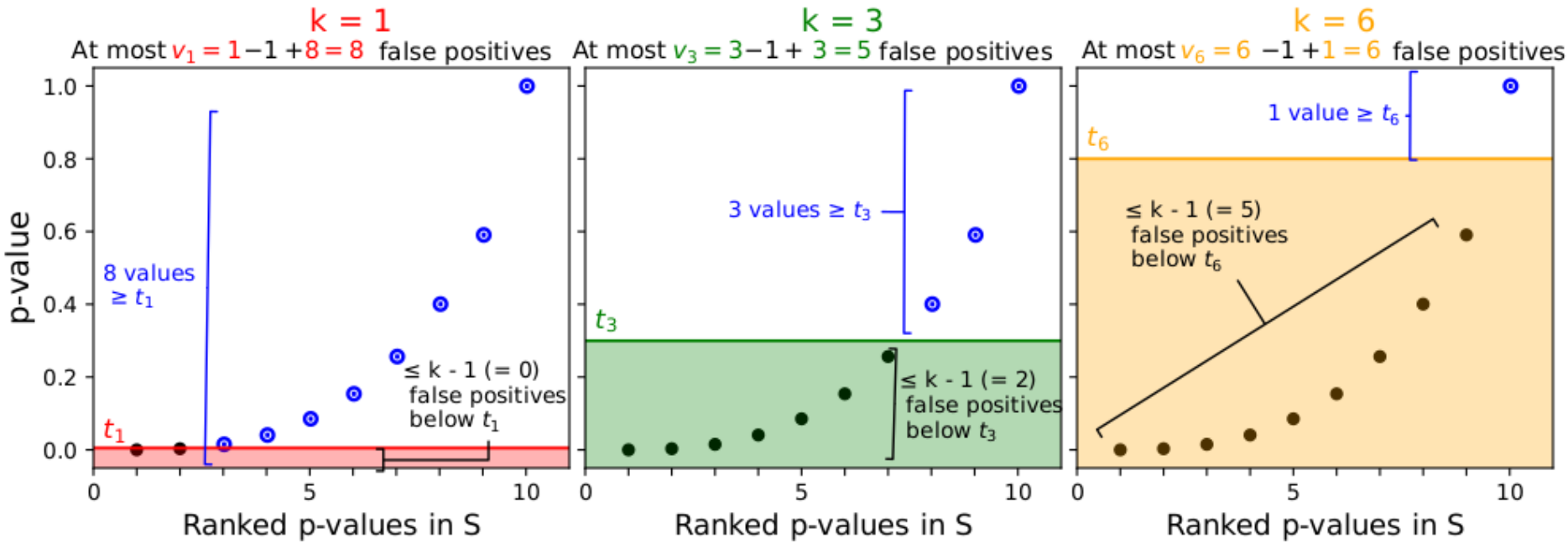


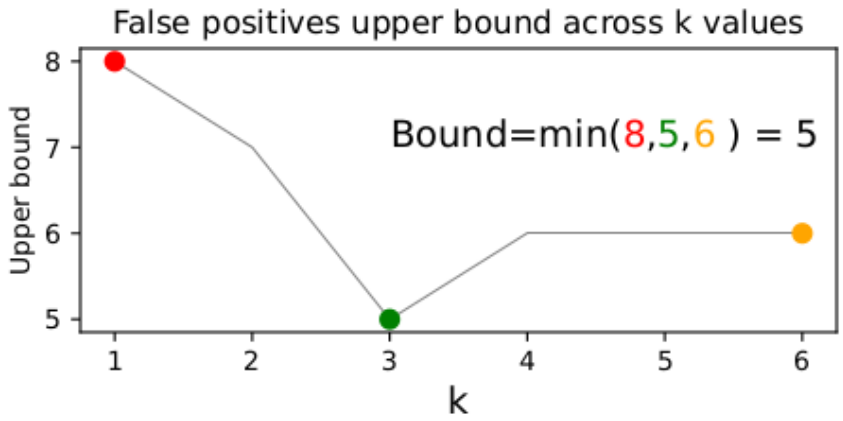
- Thus for any set S:

With probability > 1- α

$$
V^S(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + k - 1 \right\}
$$

Controling the FDP: interpolation





Controlling the FDP: joint error rate

- For a suitable threshold family $t := (t_k)_k$

$$JER(t) = \mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:m_0)} \leq t_k\right)$$

- We want to control $JER(t) \leq \alpha$

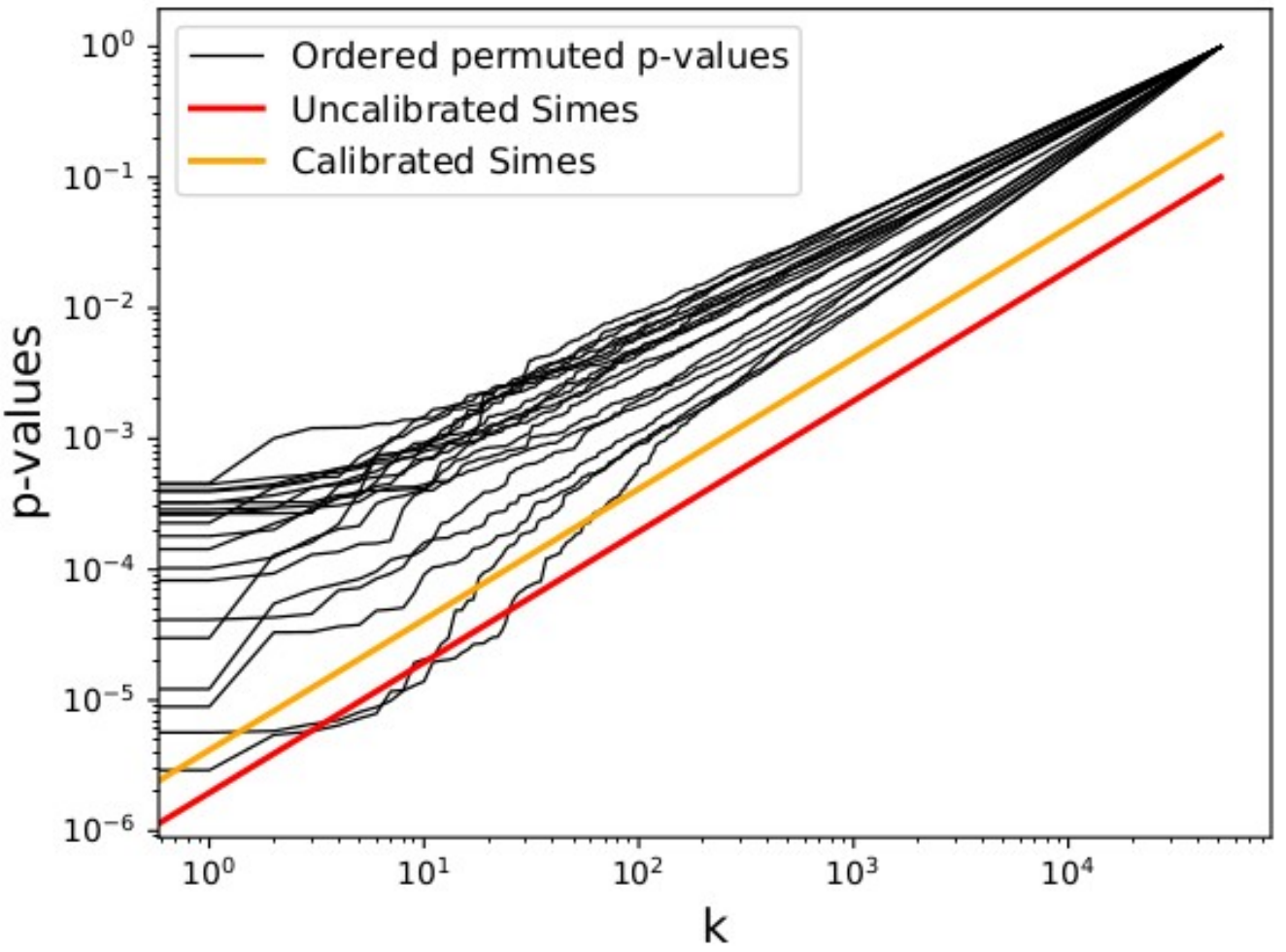
- Simes family holds under general conditions, but it is too conservative

  → use a “calibration” to estimate it.

  [Blanchard et al. Ann. Statist 2020]

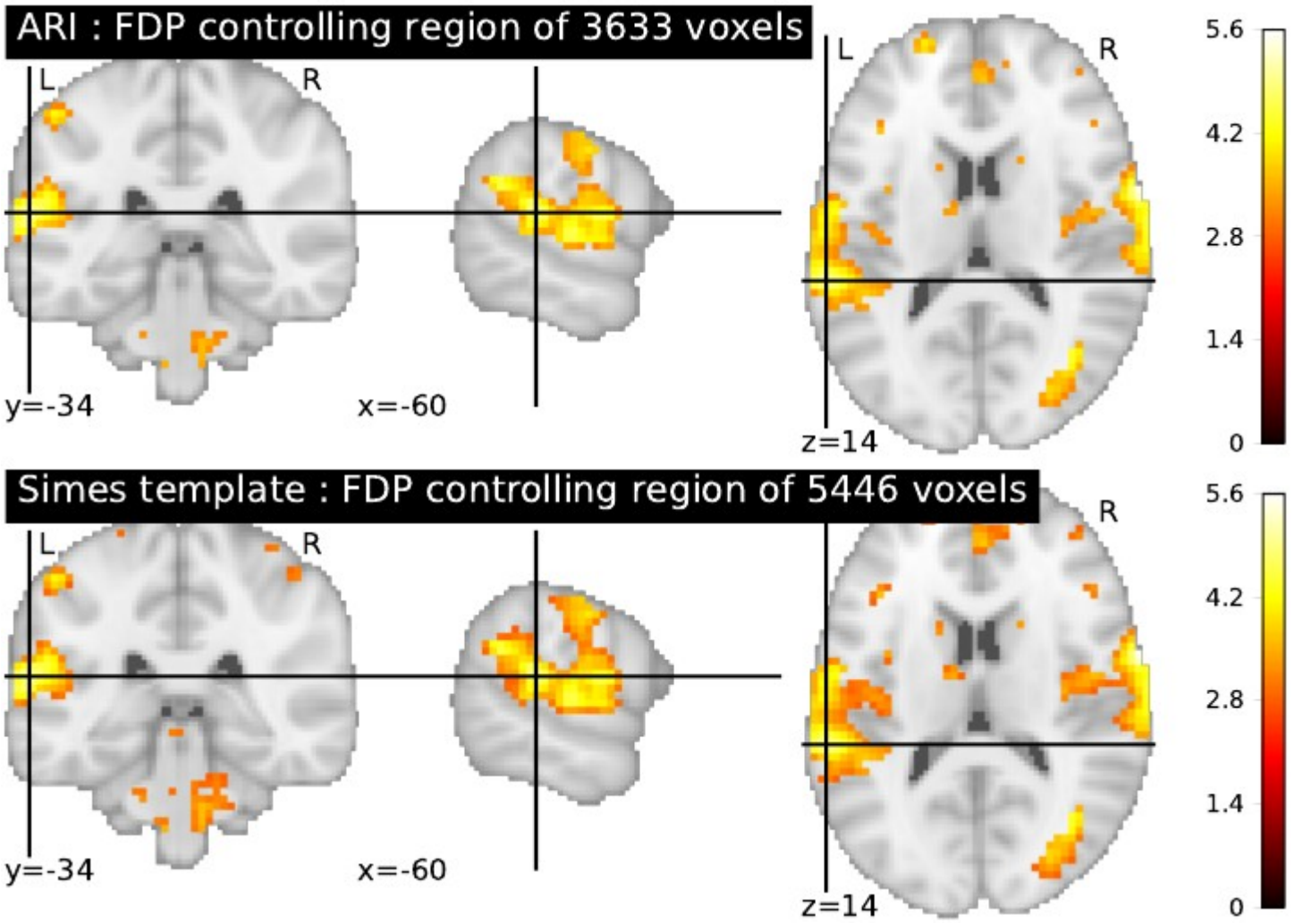
Tight control of the FDR

- Idea: calibrate the template family by comparing it to p-values obtained under permutation.

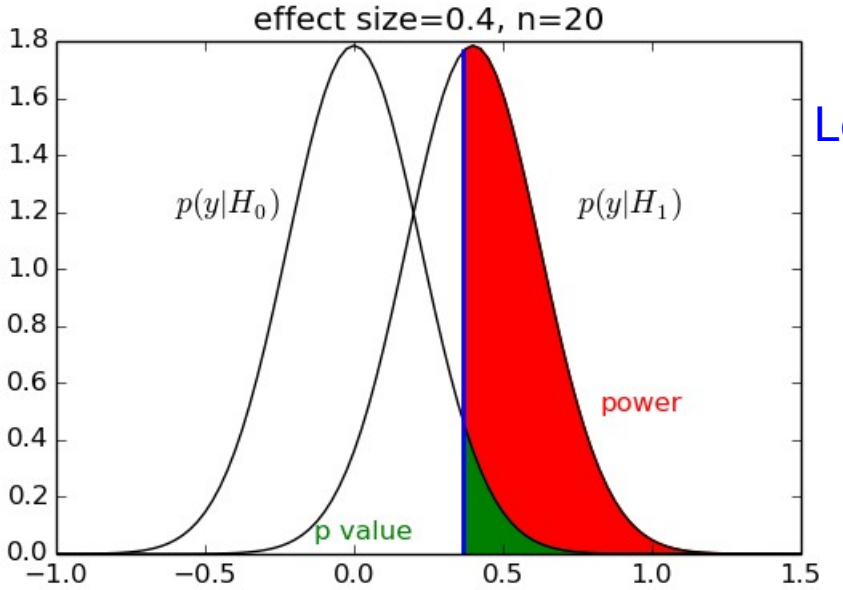


FDP control: results

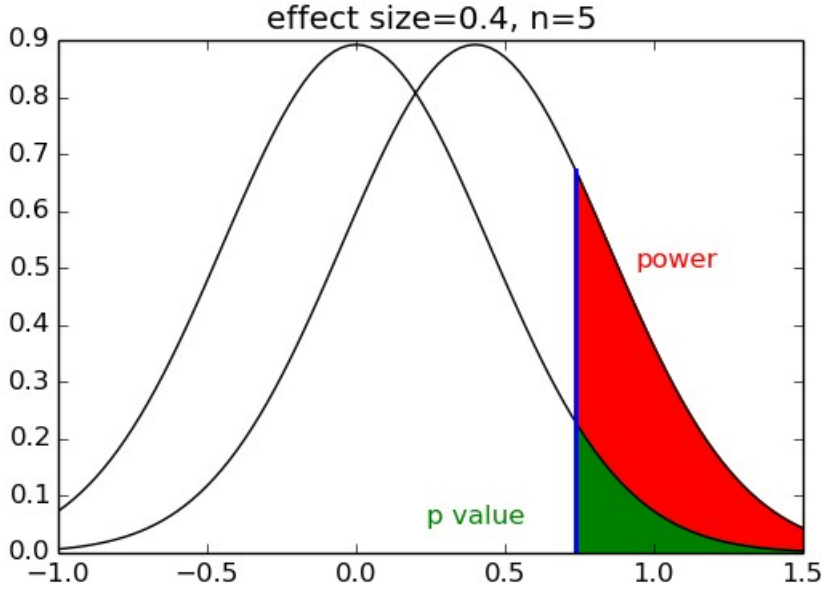
- FDP<0.1 regions
- Shifting from uncalibrated Simes (ARI) to calibrated Simes enhances sensitivity.



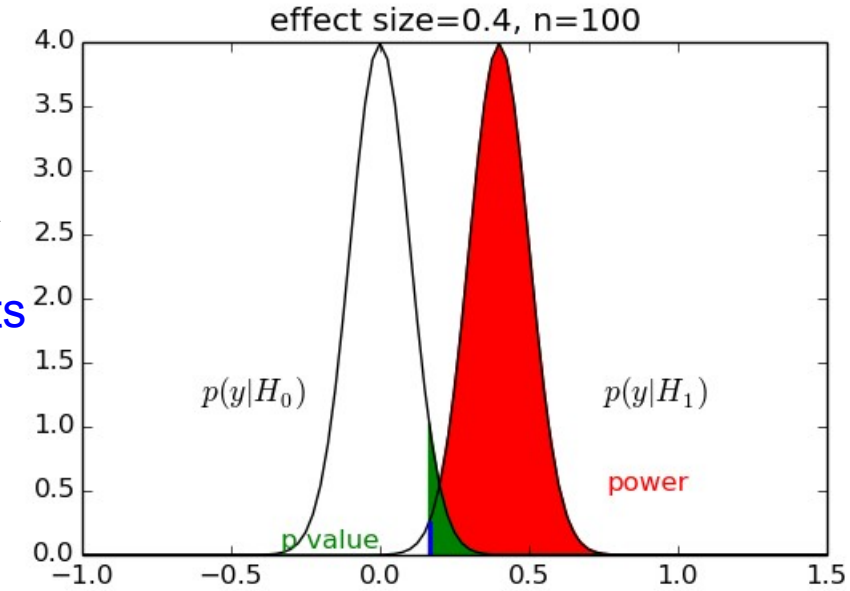
Power & reproducibility



Less subjects



more subjects



Power and reproducibility

- Currently a critical issue

Nature Reviews Neuroscience | AOP, published online 10 April 2013; doi:10.1038/nrn3475

nature REVIEWS

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]

ONLINE FIRST

Excess Significance Bias in the Literature on Brain Volume Abnormalities

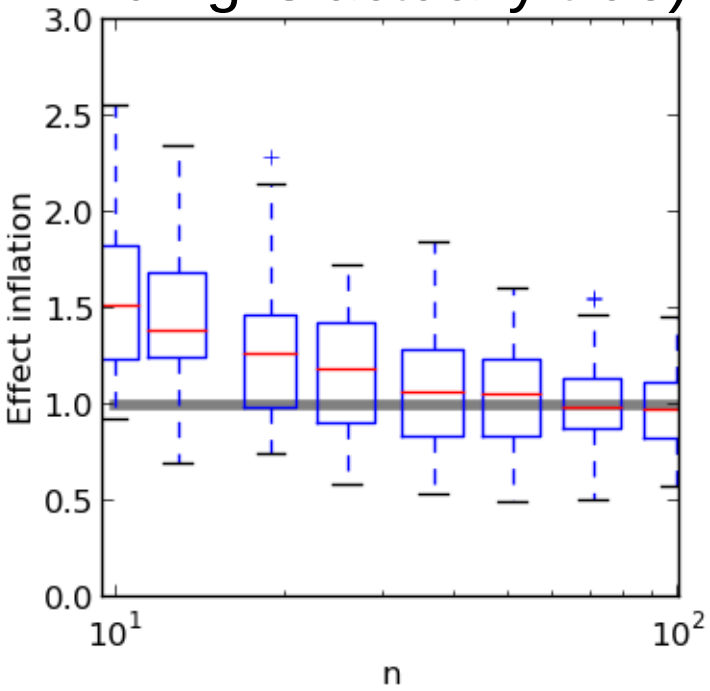
John P. A. Ioannidis, MD, DSc

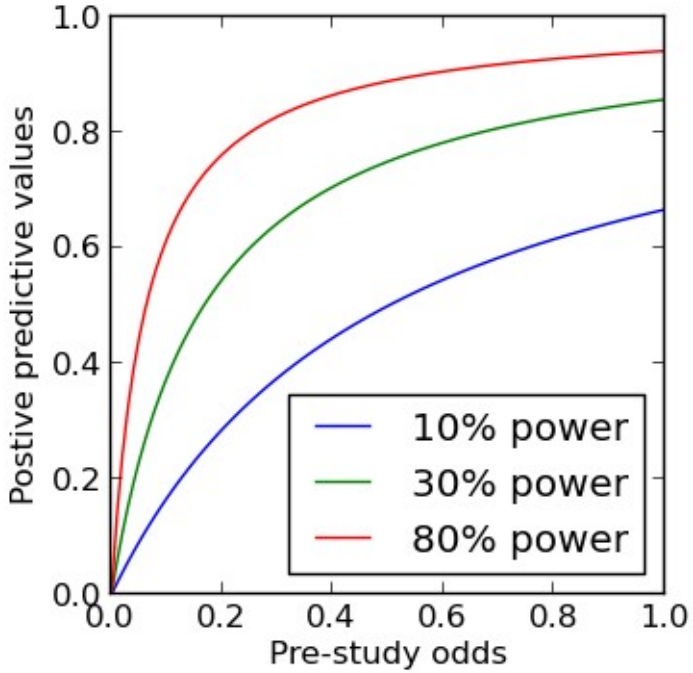
Power and reproducibility

Low power → unreliable findings

False negatives: low probability of finding true effects

Low positive predictive value: lower likelihood that a statistically significant finding is actually true)



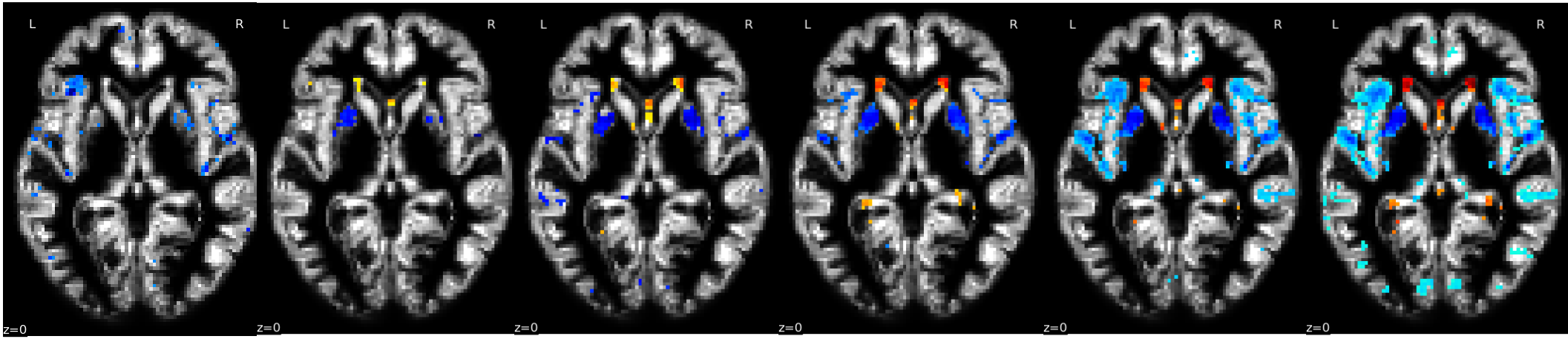


Exaggerated estimate of the magnitude of discovered effects "*winner's curse*"

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

Example

OASIS dataset, effet of age on gray matter density

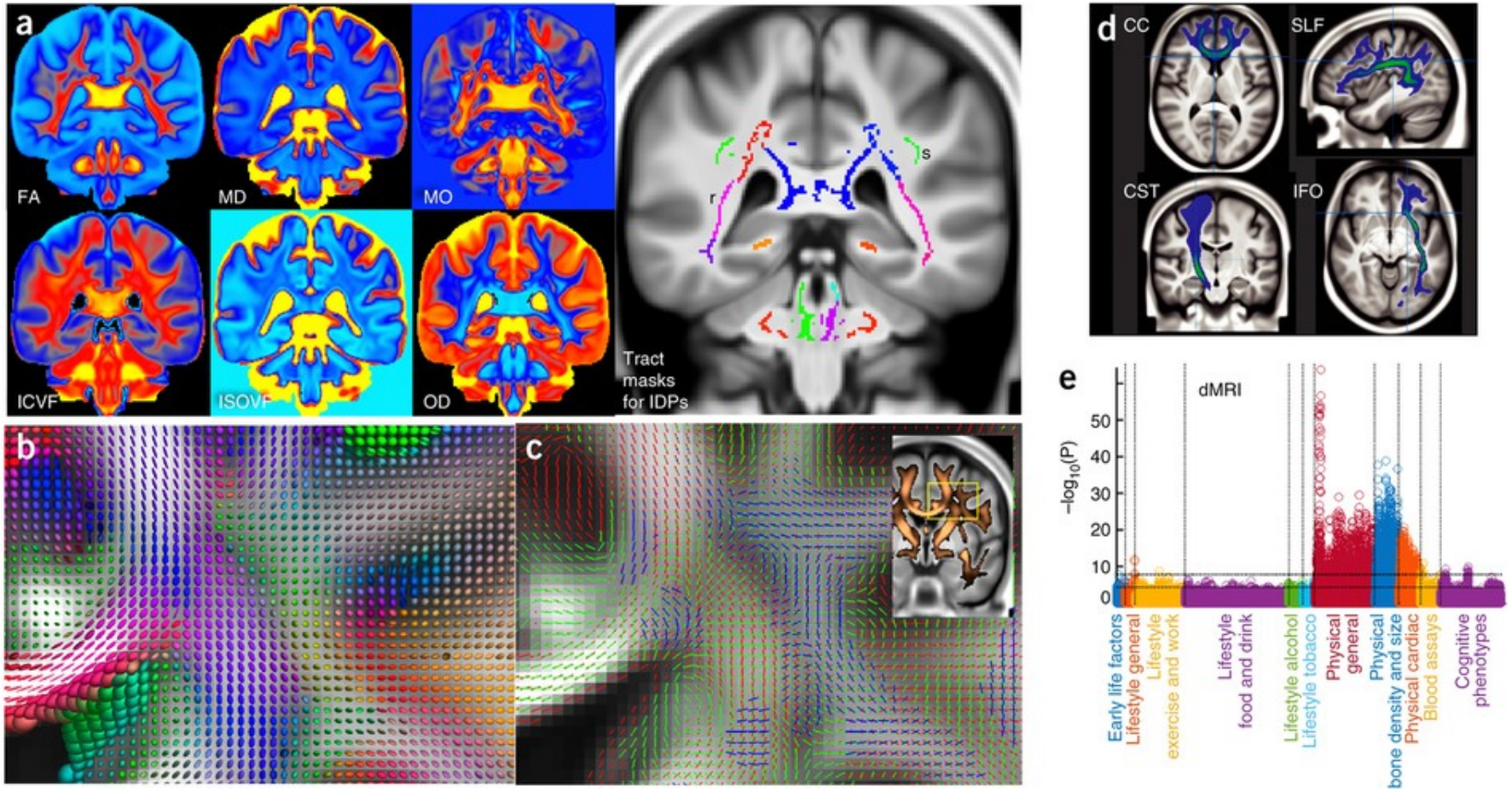


n=10 n=20 n=30 n=50 n=70 n=100

Reproducibility by bootstrap:

7% 19% 32% 53% 66% 75%

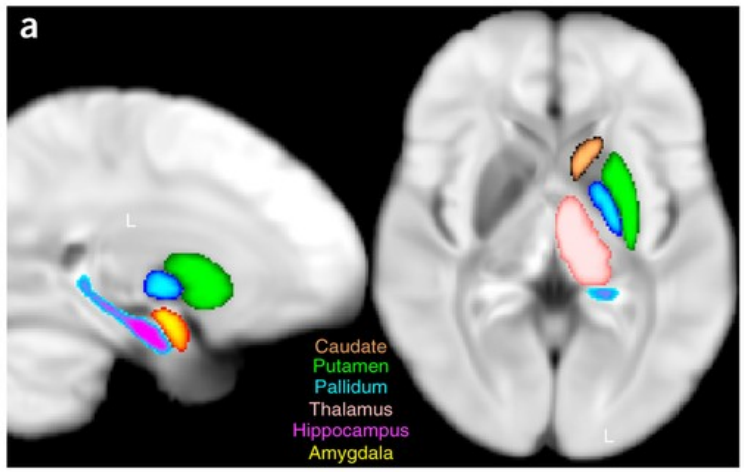
Example: UK biobank



5K participants 09/2016 → 100K in 2021

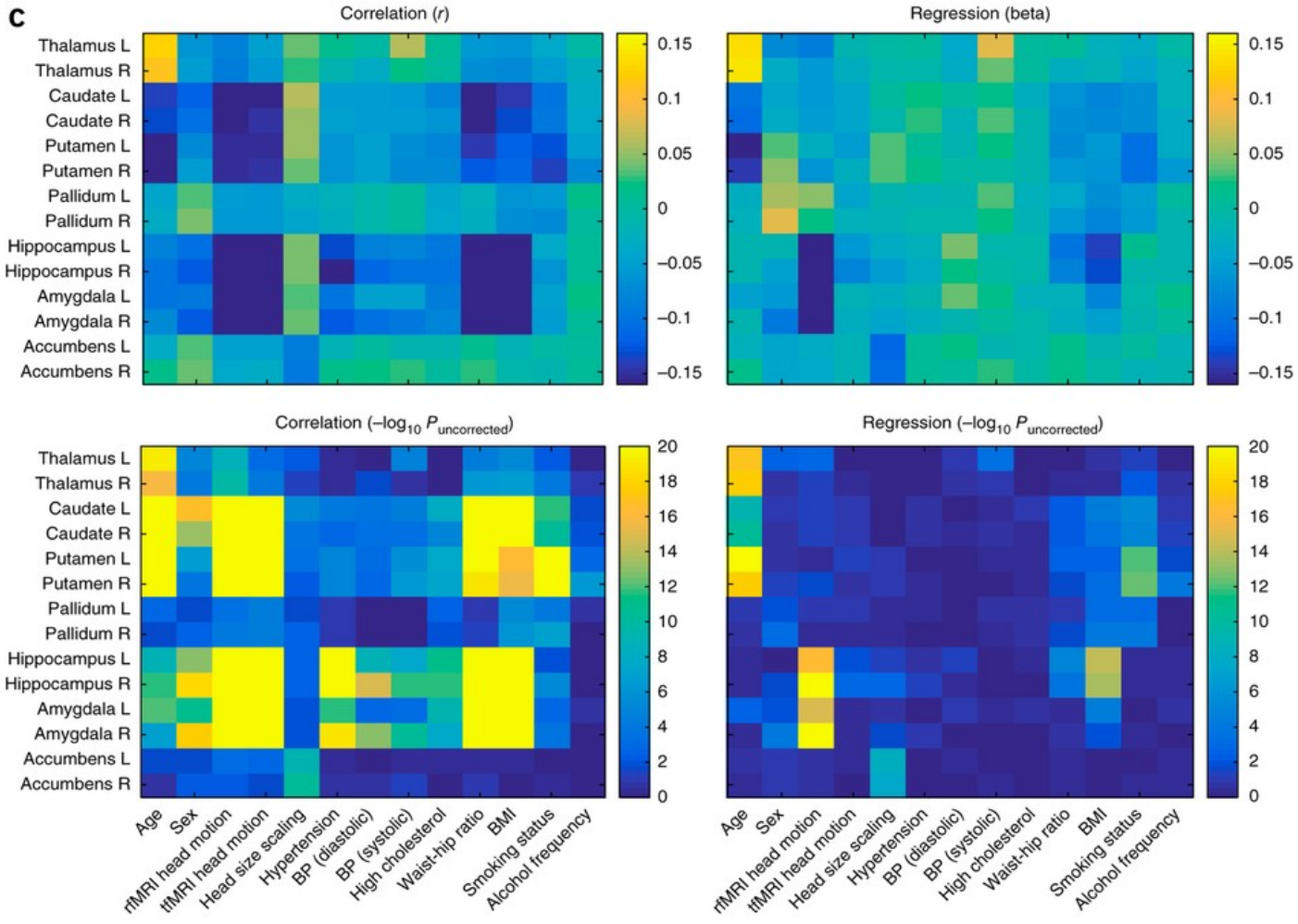
[Miller et al. Nat. Neurosci. 2016]

Example: UK biobank replicates previous studies



b

	ASPS R2* –BMI regression beta	ASPS R2* –BMI regression $-\log_{10}P_{FDR}$		UK Biobank T2* –BMI regression beta	UK Biobank T2* –BMI regression $-\log_{10}P_{uncorr}$
Thalamus	0.029	0	L	–0.0239	0.7
			R	–0.0187	0.5
Caudate	0.046	0	L	–0.0789	4.5
			R	–0.0746	4.1
Putamen	0.044	0	L	–0.0570	2.6
			R	–0.0649	3.2
Pallidum	–0.002	0	L	–0.0661	3.2
			R	–0.0772	4.1
Hippocampus	0.135	>4.0	L	–0.1390	14.2
			R	–0.1343	13.7
			Mean	–0.1490	17.0
Amygdala	0.232	2.0	L	–0.0786	4.6
			R	–0.0478	2.0
			Mean	–0.0701	3.9
Accumbens	–0.048	0	L	–0.0117	0.2
			R	–0.0040	0.1



Machine learning approaches

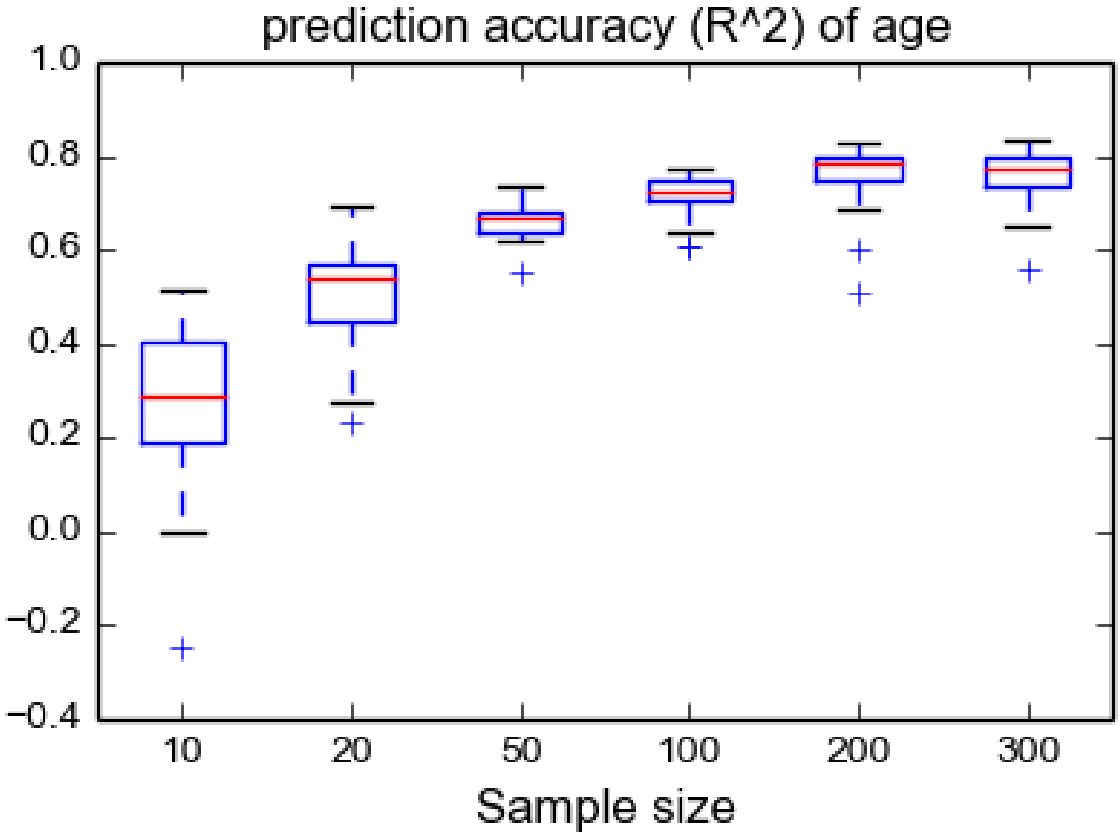
- Can one **predict** a behavioral feature based on brain images
  - Age of the subject
  - Handedness, IQ...
  - Disease status (psychiatry)
- **Diagnosis purpose**
- Multivariate approach: use signals from the brain image to predict the signal of interest

Issues

- Low SNR of the data
- Small sample / features ratio
- Consequences
  - Tendency to overfit
  - Robust rather than sophisticated models
  - Difficulty to make inference on the “pattern”

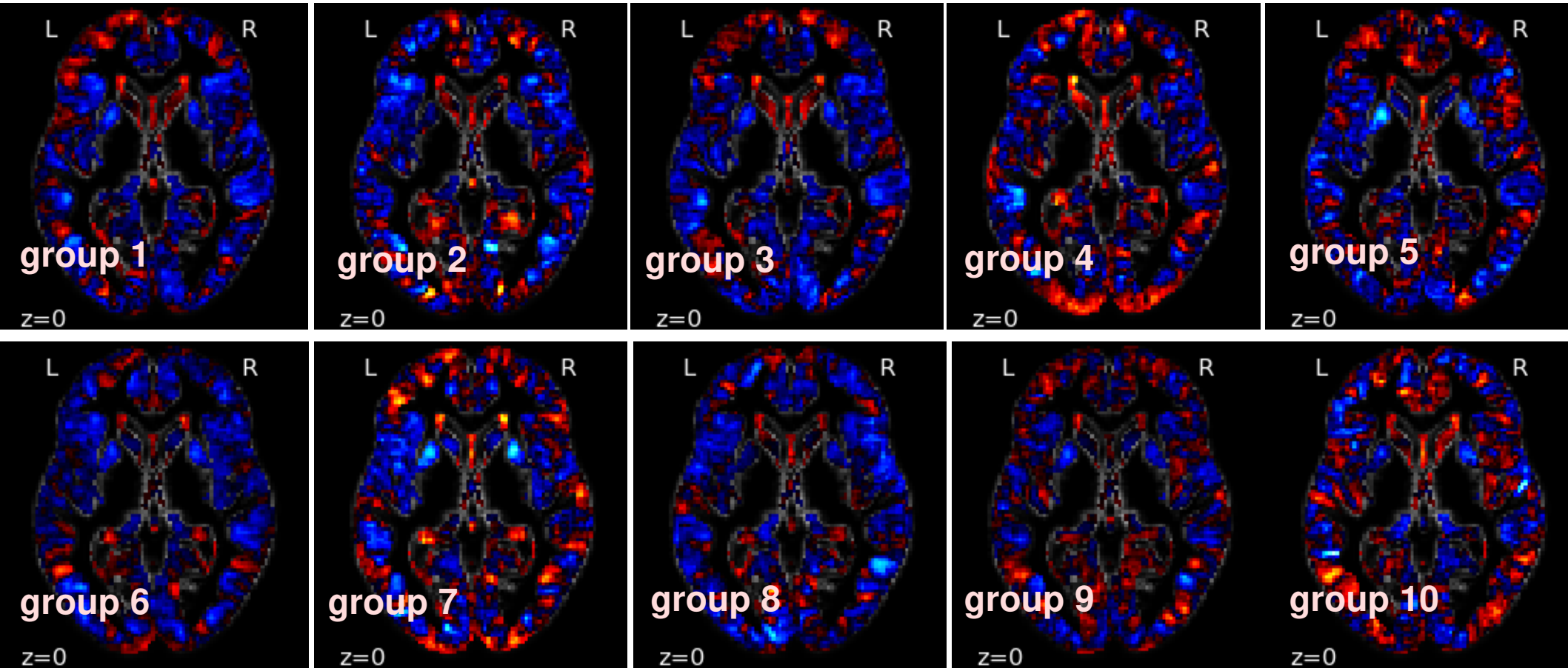
Multivariate analysis: example

- Predict the age of a subject given gray matter density maps (OASIS dataset, n=403)



Weight maps for age prediction / OASIS

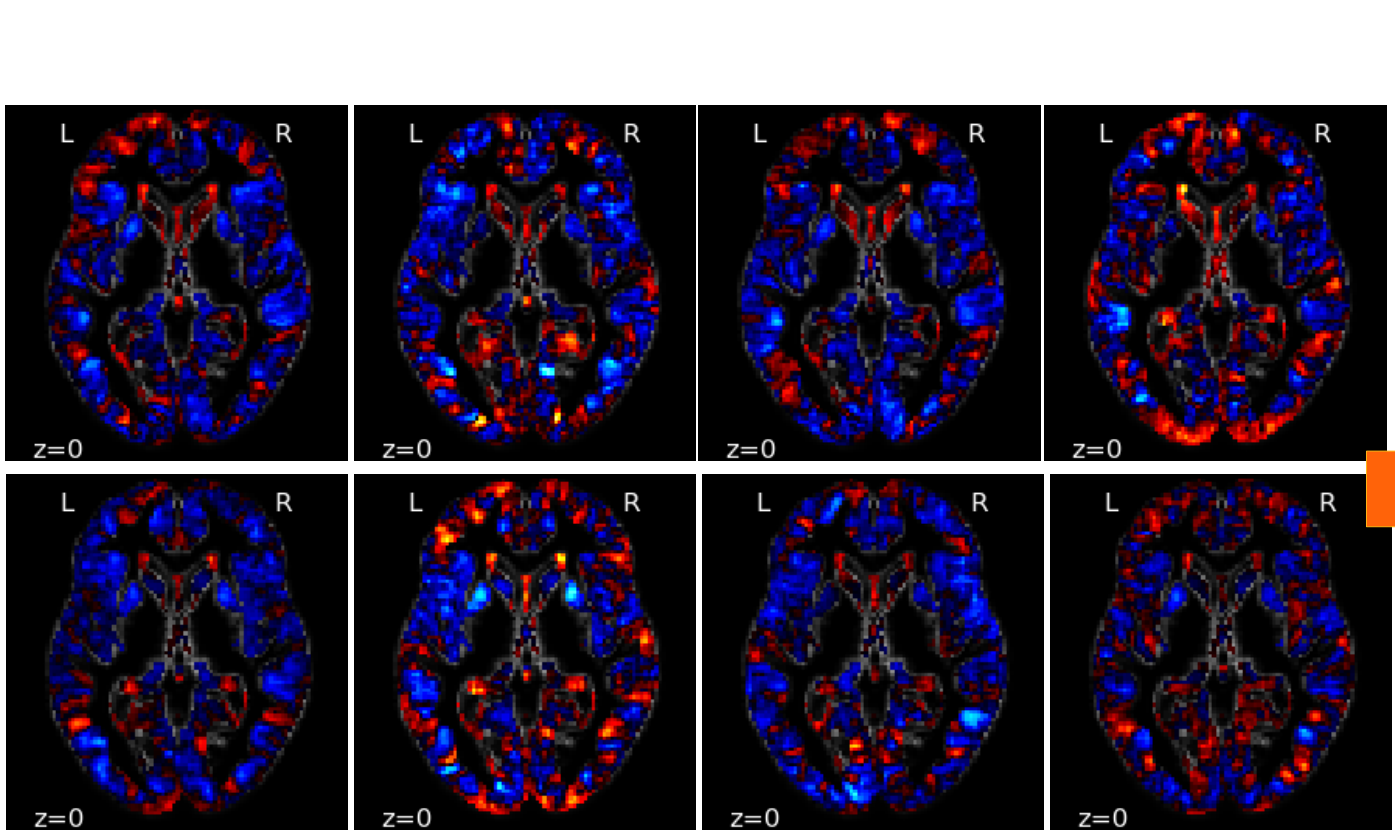
Weight maps depends on the batch of subjects considered (bootstrap):
One question, different datasets, different answers



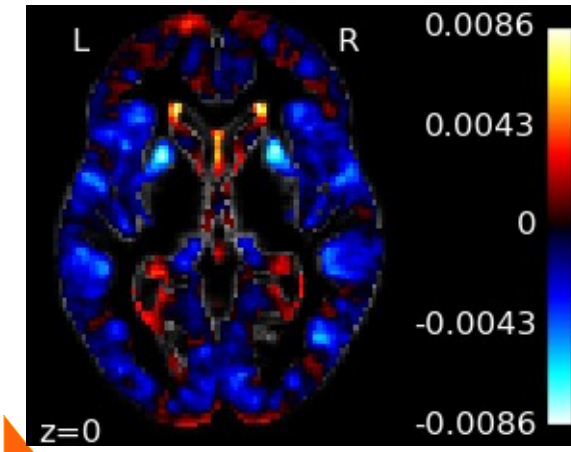
Variability actually worse than for univariate analysis !

Weight maps for age prediction / OASIS

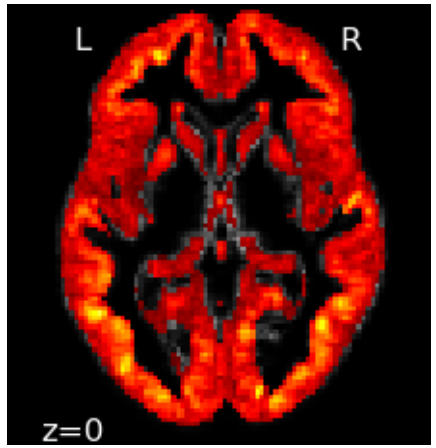
The weight map depends on the batch of subject considered (bootstrap):
One question, different dataset, different answers



mean

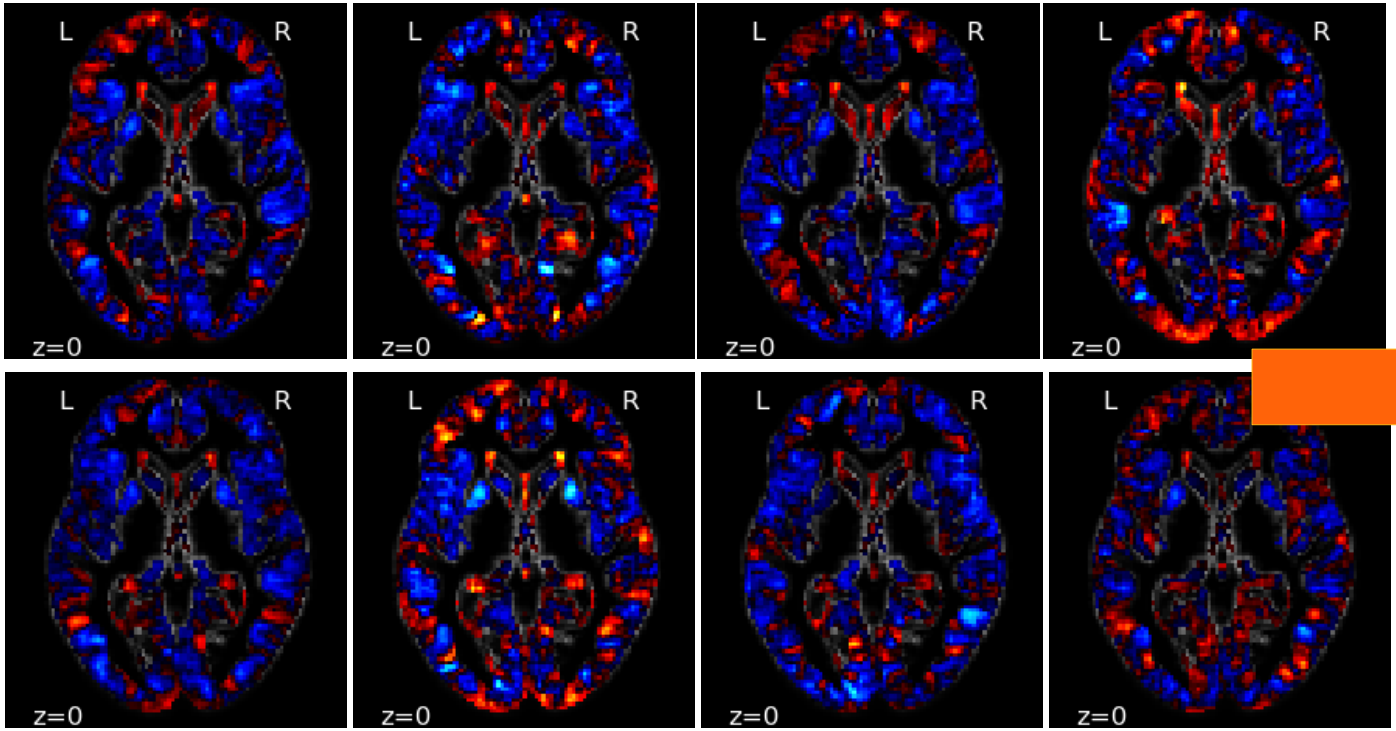


variance

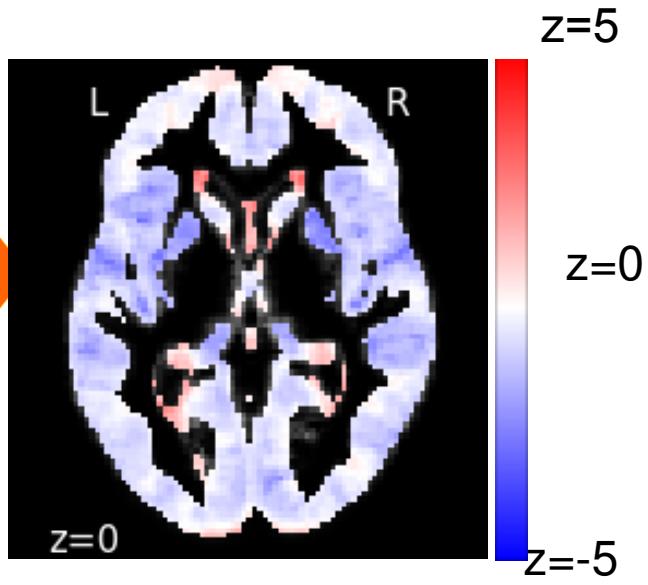


Weight maps for age prediction / OASIS

The weight map depends on the batch of subject considered (bootstrap):
One question, different dataset, different answers

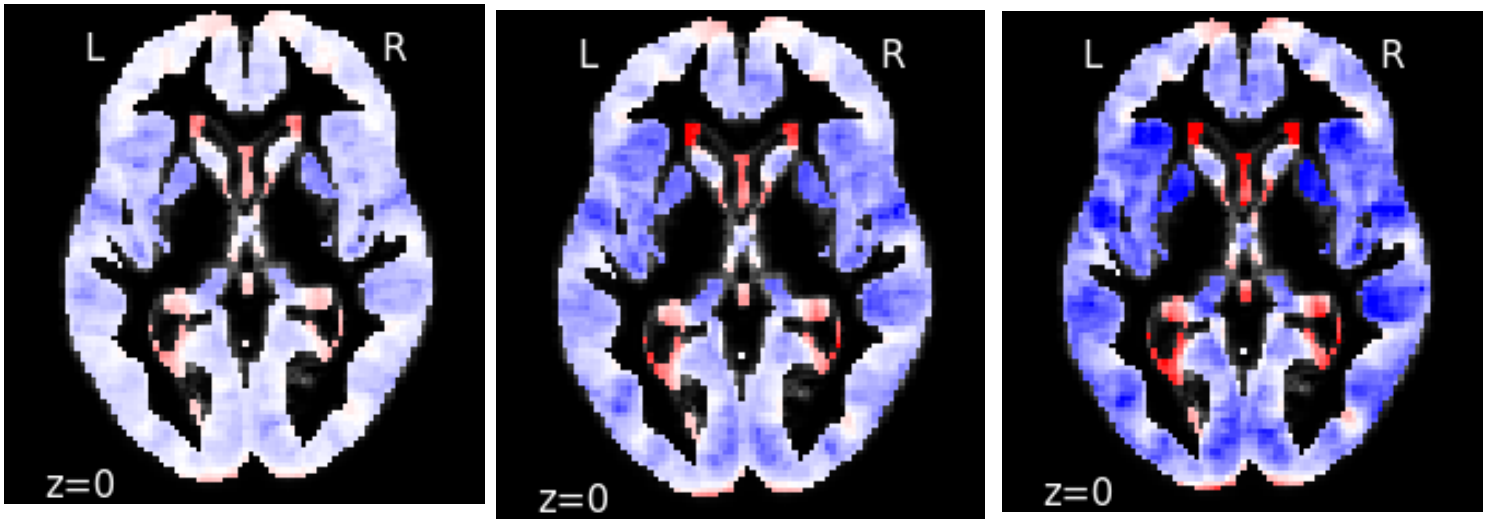


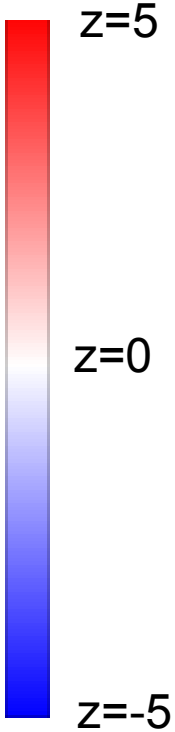
Summarized into a z image:
(effect size) / (effect std)



Weight maps for age prediction / OASIS

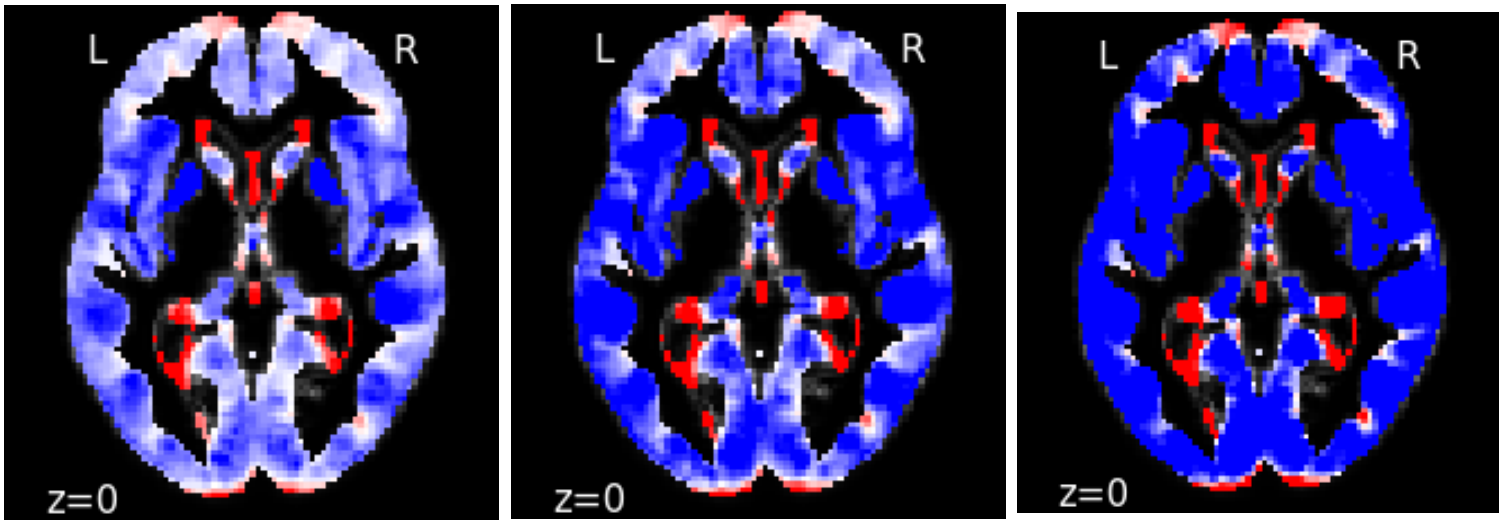
n=10 n=20 n=50





(effect size estimated by bootstrap)

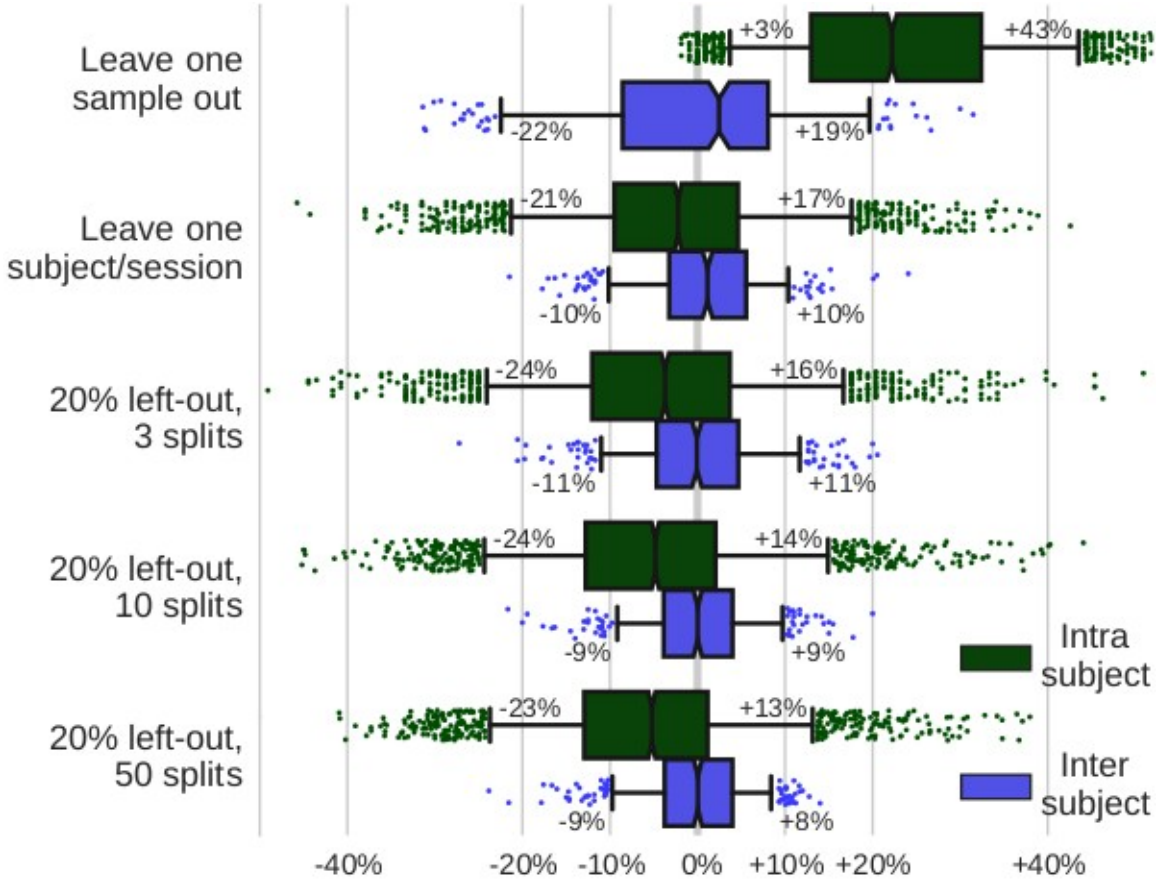
n=100 n=200 n=300



Sample size & cross-validation

Cross-validation strategy

Difference in accuracy measured by cross-validation and on validation set



cross-validation < validation set

cross-validation > validation set

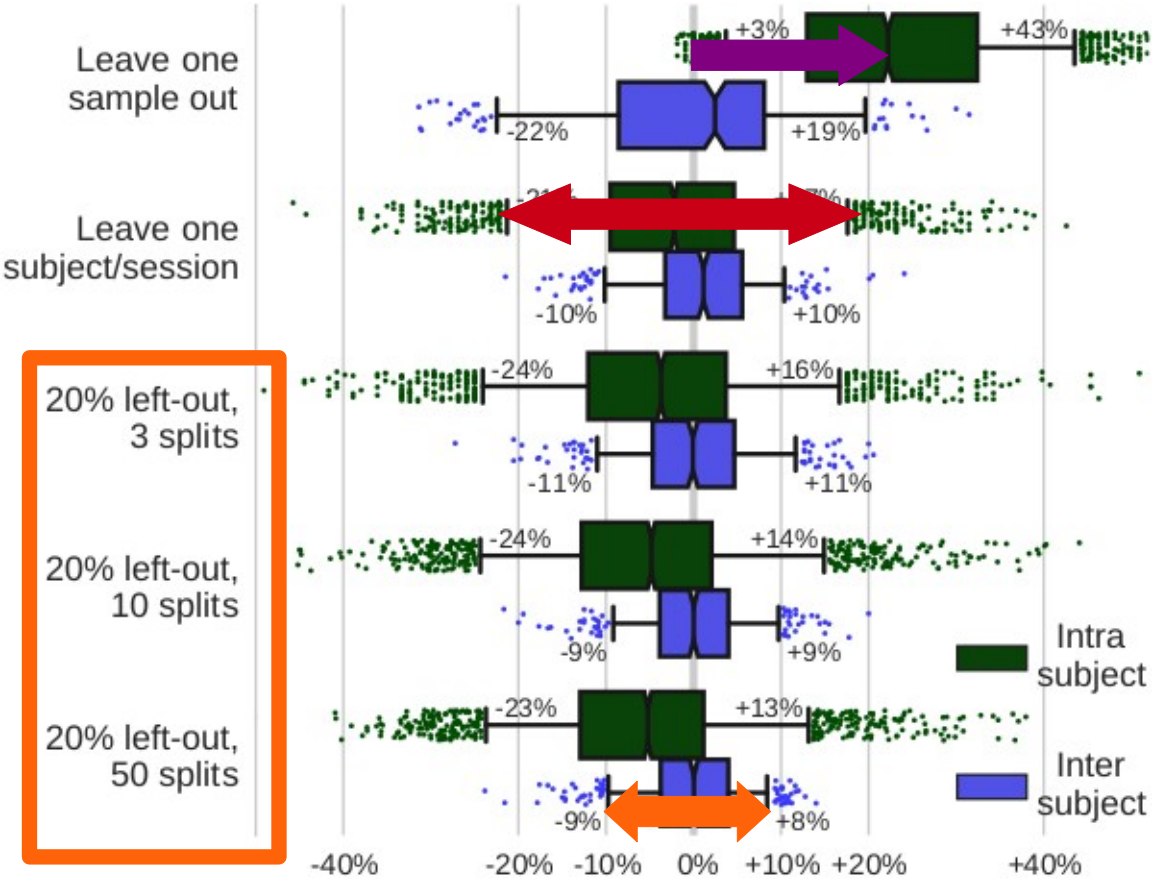
Large-scale experiment: 4 classifiers, 7 datasets, 1 anatomical dataset, 1 MEG dataset

[Varoquaux et al. NIMG 2016]

Sample size & cross-validation

Cross-validation strategy

Difference in accuracy measured by cross-validation and on validation set



- optimistic bias in LOO for non-independent samples
- higher variance in LOO
- variance large overall → use shuffle-split with many splits

cross-validation < validation set

cross-validation > validation set

[Varoquaux et al. NIMG 2016]

Why not use leave-one-out CV ?

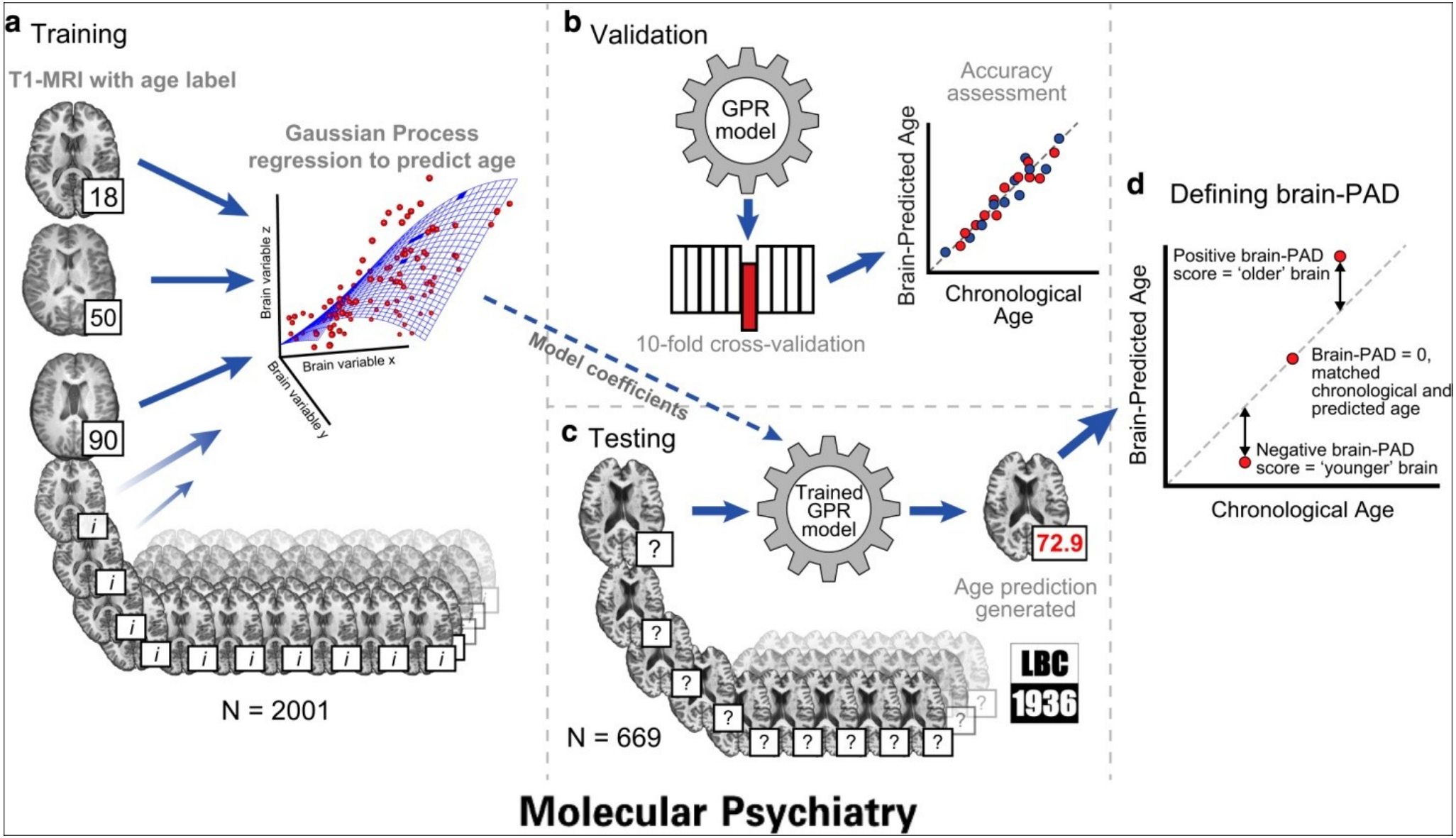
- Not a good estimator: compare with bootstrap
- Larger variance of the results
- More susceptible to artifacts
  - e.g. when the intercept is not fit properly, trivially good results in regressions

Predicting age using neuroimaging: innovative brain ageing biomarkers

"Furthermore, a growing body of research is demonstrating that so-called ***brain age*** has both clinical and broader scientific relevance. This paradigm has provided a new way to explore how the brain changes during ageing and how brain diseases interact with 'normal' brain ageing. Potentially, *brain age* could be used as a personalised biomarker of brain health during ageing, and this individual-specific nature is particularly important. The extensive study of group-mean differences in case-control studies of brain diseases has yielded few clinical applications. Conversely, *brain age* locates an individual within a normative ageing distribution. If this location can be shown to be relevant for health outcomes, then *brain age* presents a framework for applying neuroimaging clinically to characterise brain health."

[Cole, Franke *Trends In Neurosciences*, 2017]

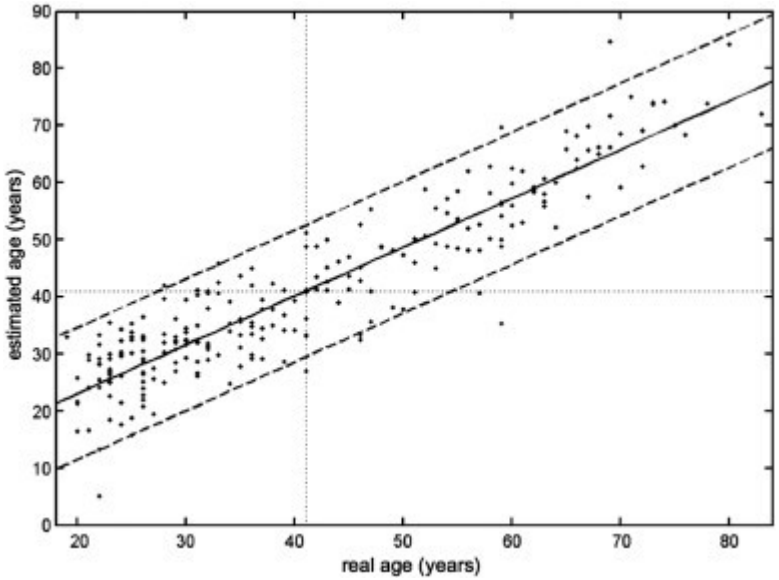
Predicting age using neuroimaging: innovative brain ageing biomarkers



Predict brain age

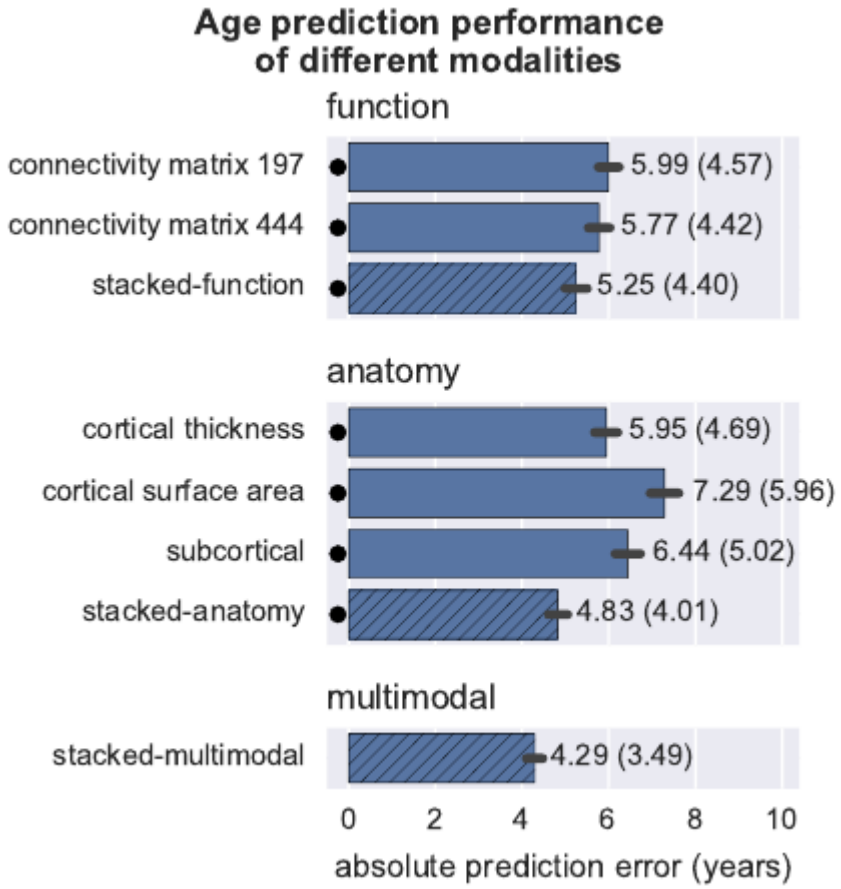
… anatomically

Thickness/shape of the cortex predict age +/-5 yrs (1 yr in adolescents)



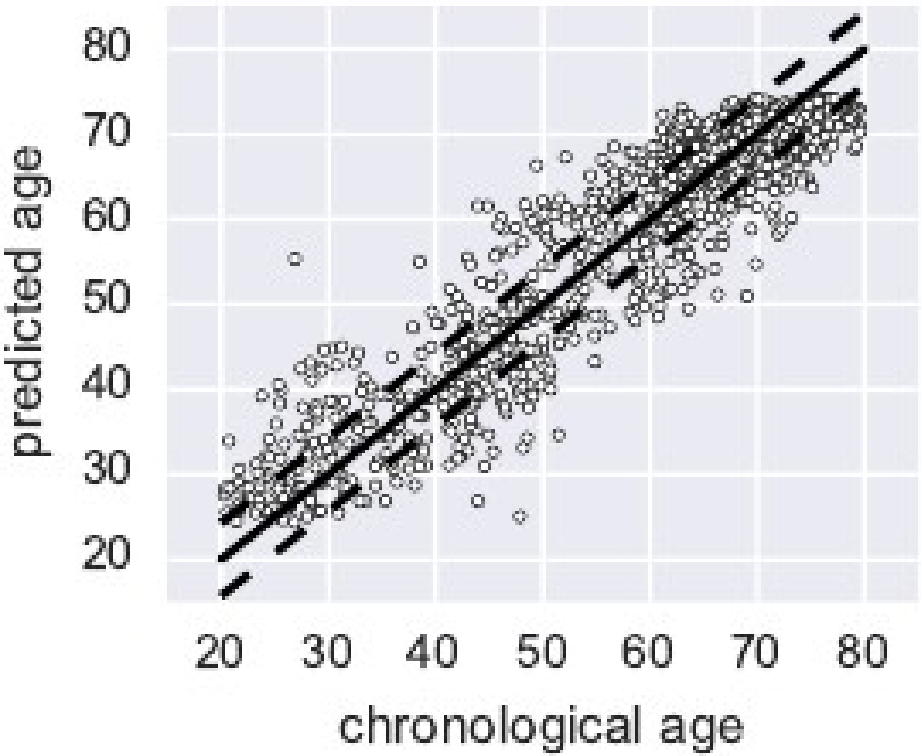
[Gaser et al. NIMG 2010]

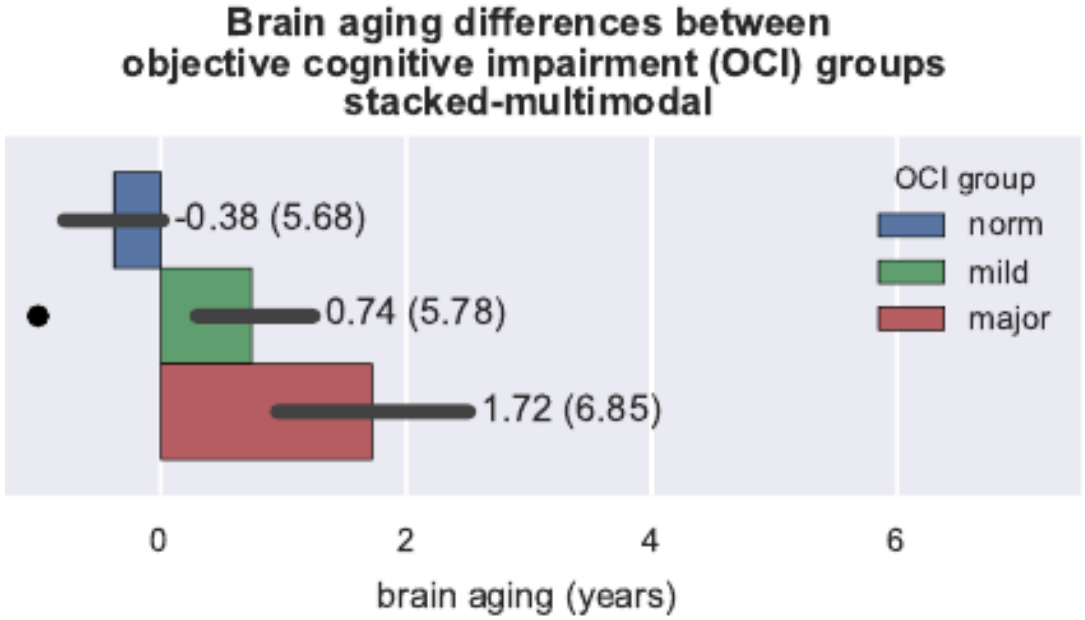
[Liem et al. NIMG 2017]



Interest : accelerated aging (AD), dementias, diabetes..

Predict brain age

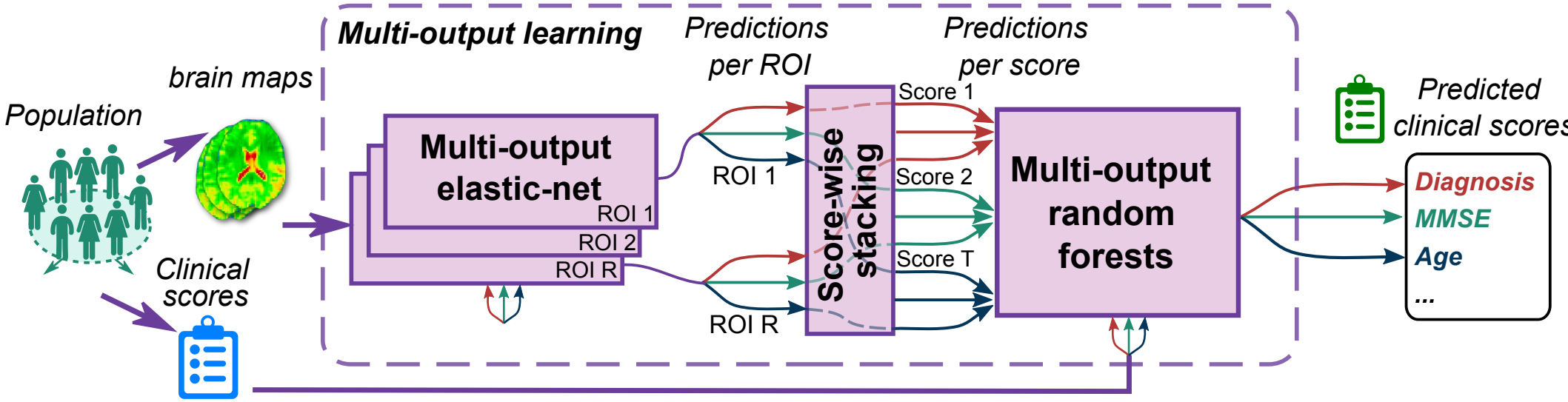




"Prediction errors" are actually informative about health: "brain age"

[Liem et al. NIMG 2017]

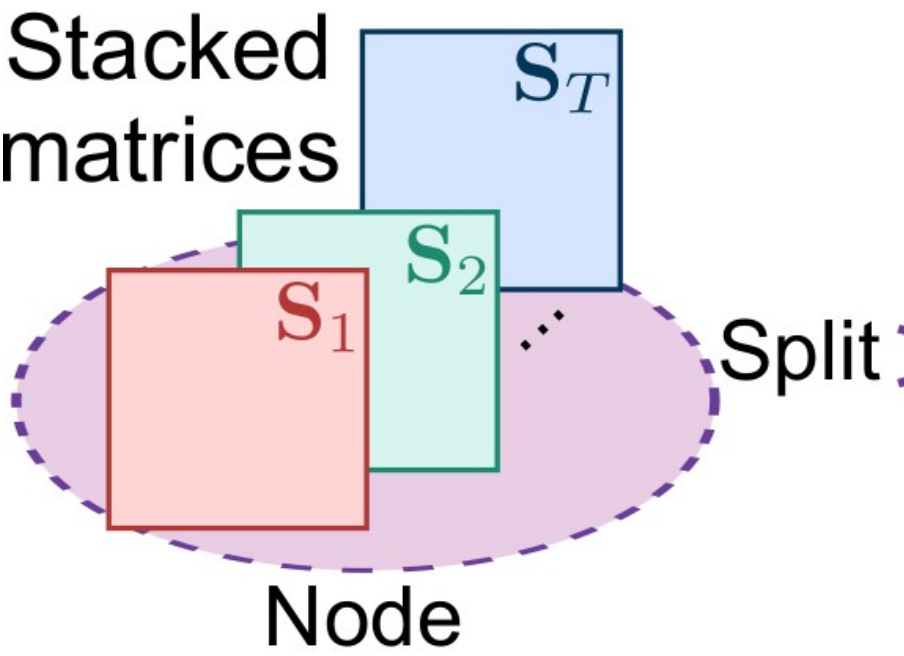
Multi-task learning



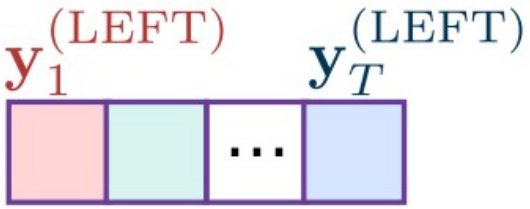
Leverage scores batteries to improve learning:

- Denoising effect
- Better generalization

Multi-output random forests

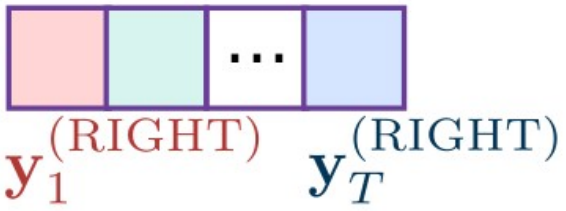


Split



Impurity criterion

$$\Gamma(\,\mathrm{node}, \mathrm{split}\,) = \sum_{j=1}^{T} \Gamma_j(\mathrm{node}, \mathrm{split})$$

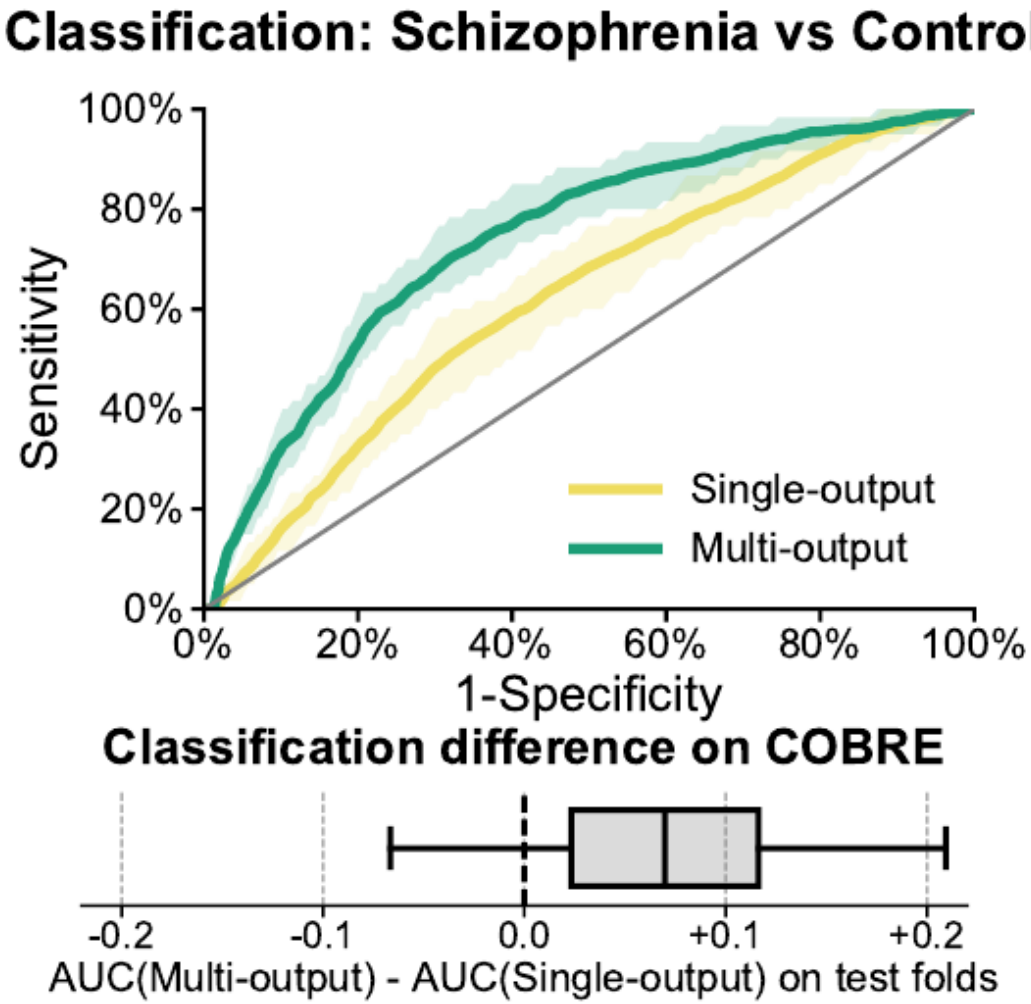


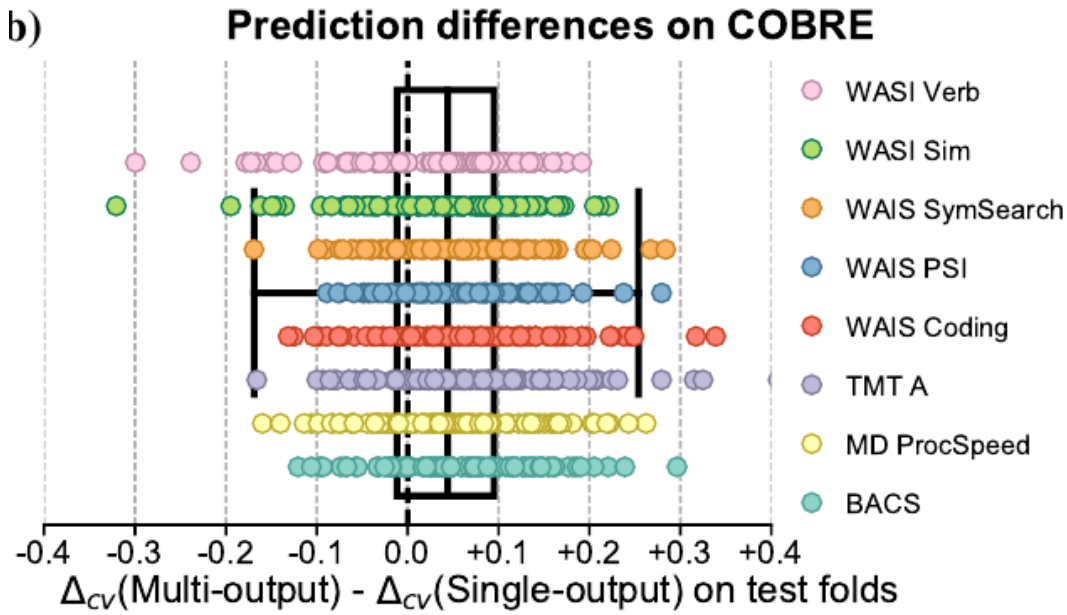
Inputs are stacked predictions for each clinical score j.

Nodes (subsets from stacking matrices S) are divided according to an impurity criterion over all outputs

[Rahim et al. Nimg 2017]

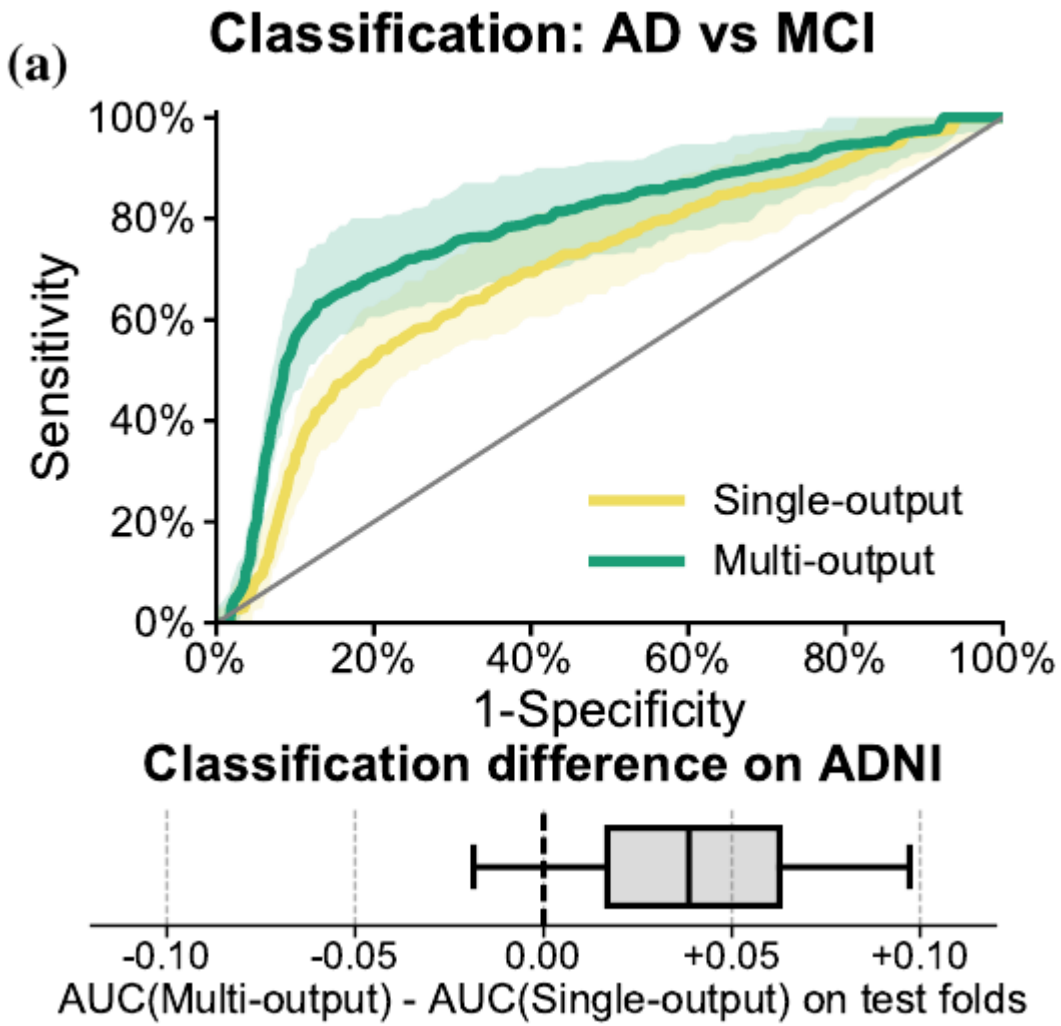
Results on schizophrenia (cobre)

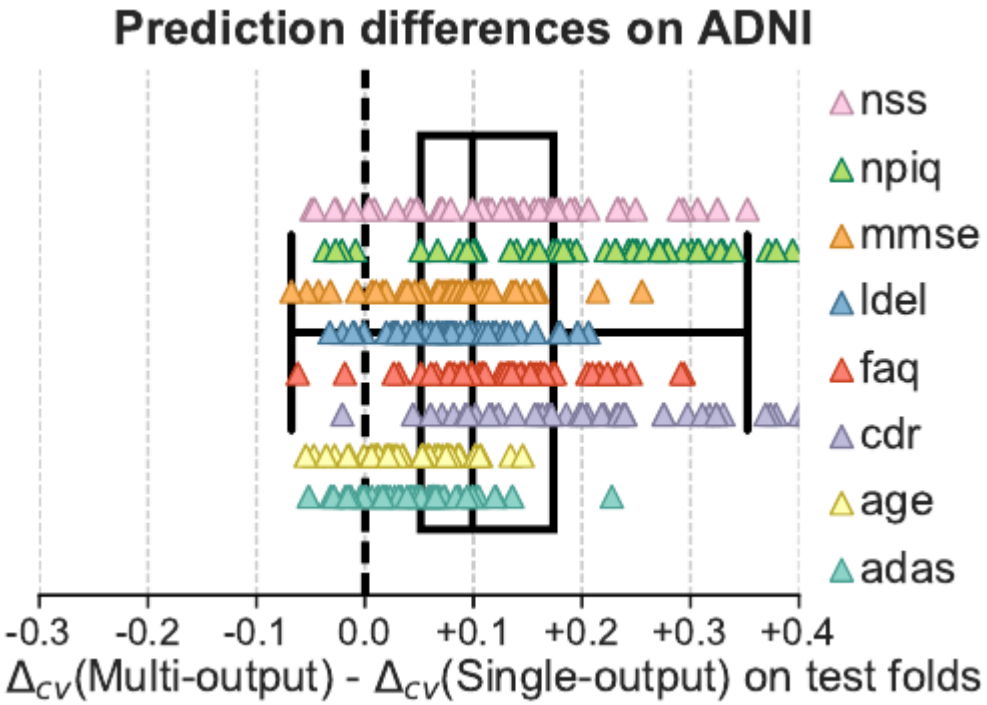




[Rahim et al. Nimg 2017]

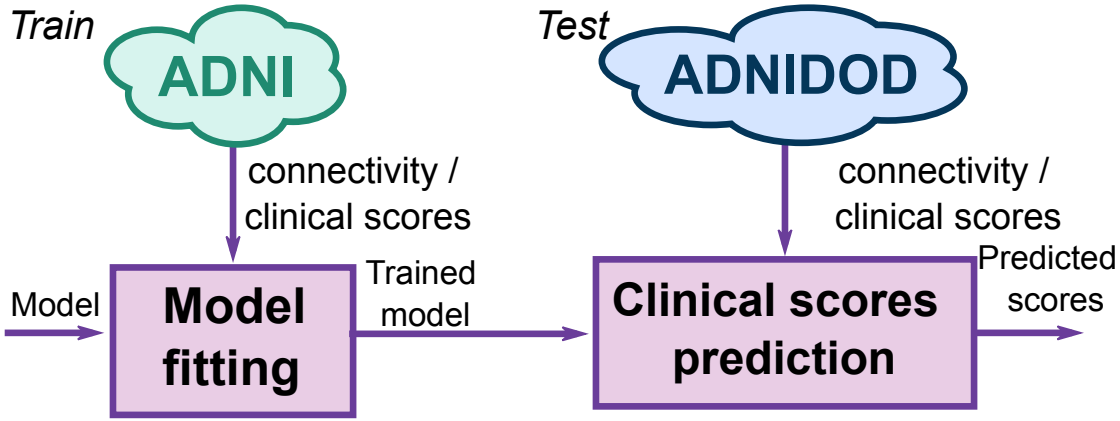
Results on adni (Alzheimer)

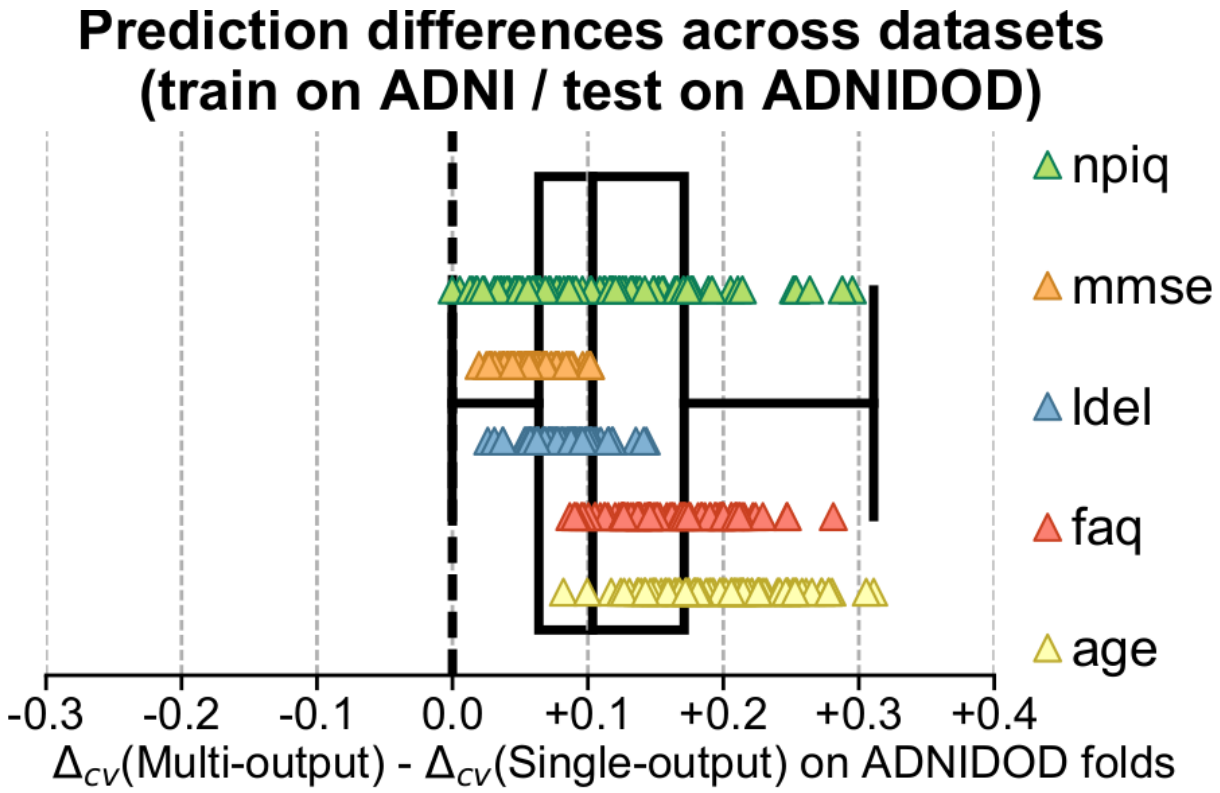




[Rahim et al. Nimg 2017]

Generalization to new datasets

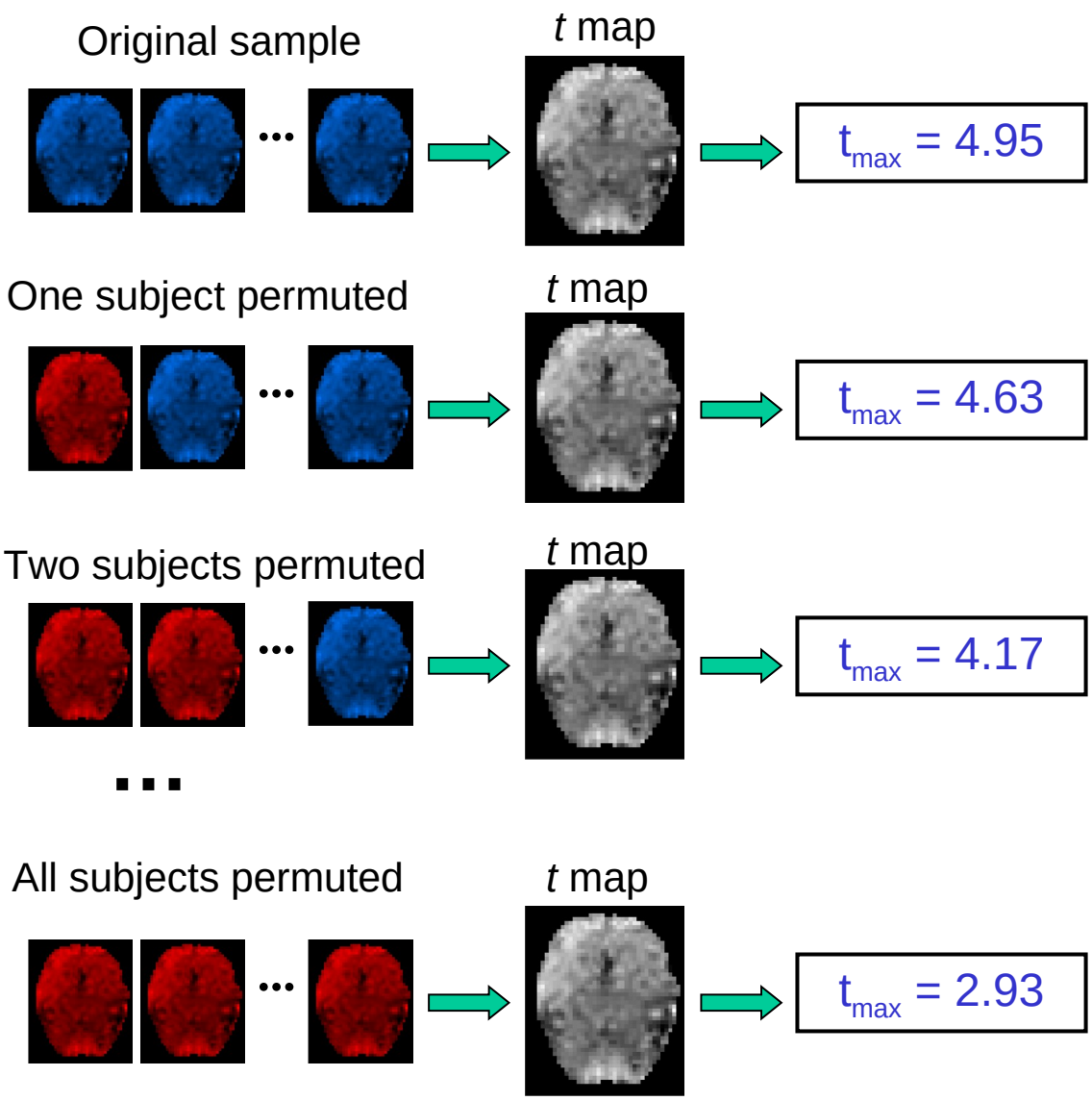




In case of transfer, multi-task learning improves generalization

[Rahim et al. Nimg 2017]

Familywise error correction



Histogram of the t_{max} statistics