Statistical inference in high-dimension & application to brain imaging

Bertrand Thirion,

bertrand.thirion@inria.fr

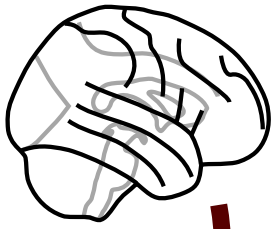
Cognitive neuroscience

How are cognitive activities affected or controlled by neural circuits in the brain ?

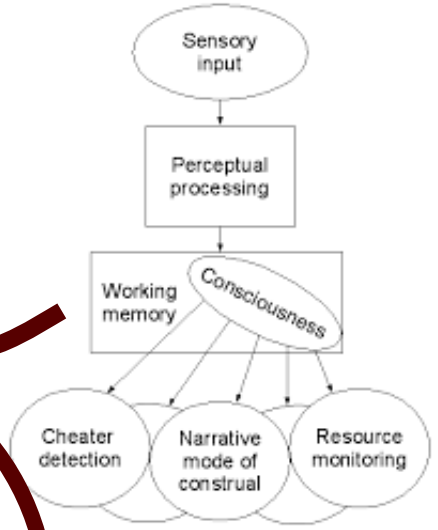
The brain, the mind and the scanner

Cognitive theories

Brain



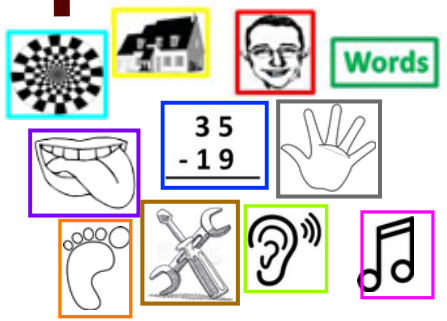
Experimental paradigm



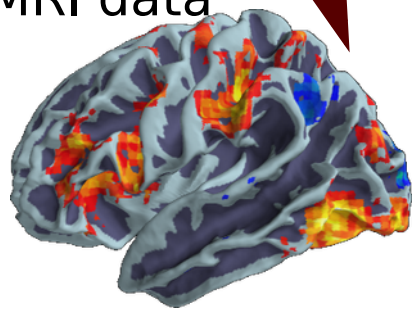
Scanner

Brain mapping

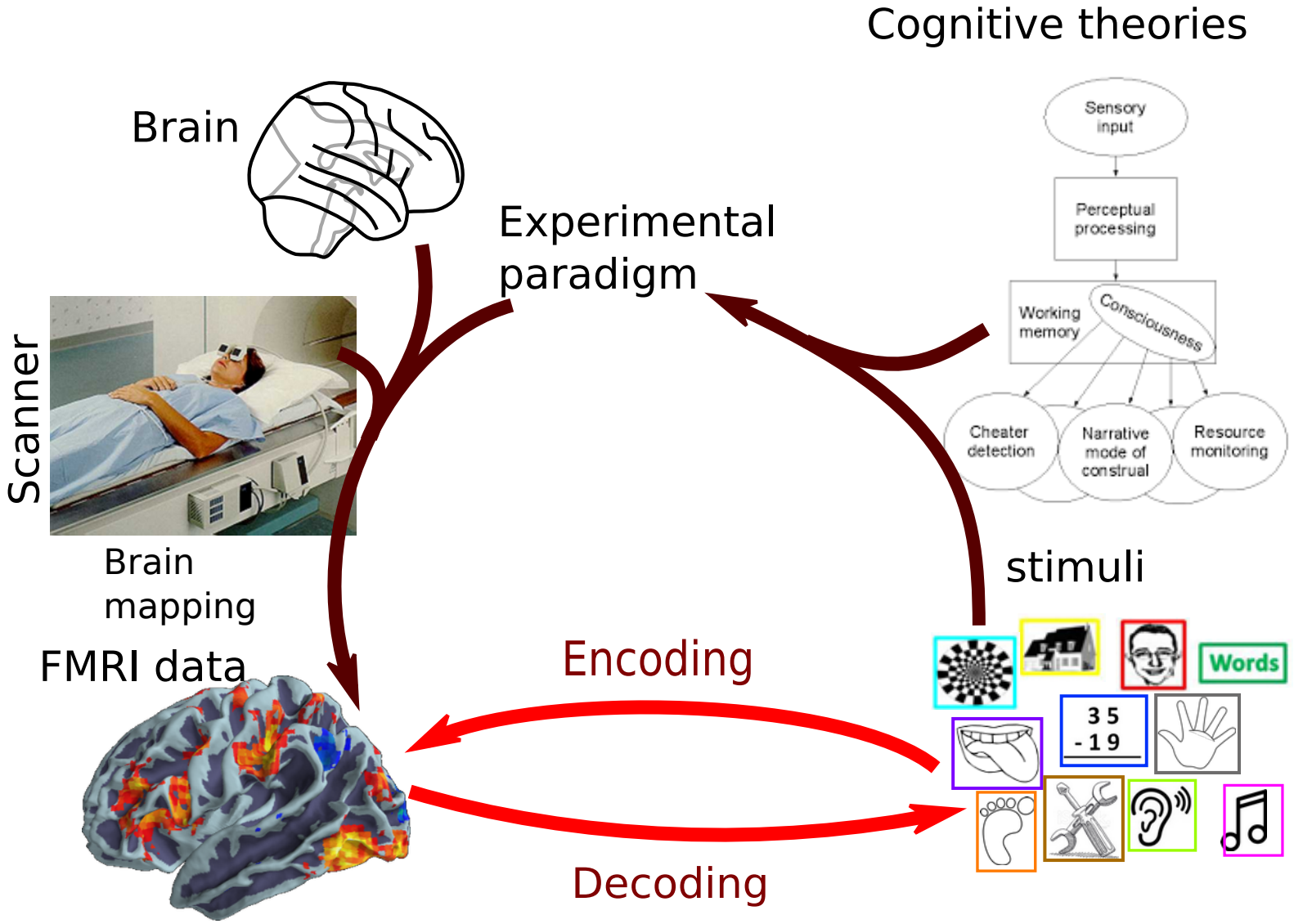
stimuli



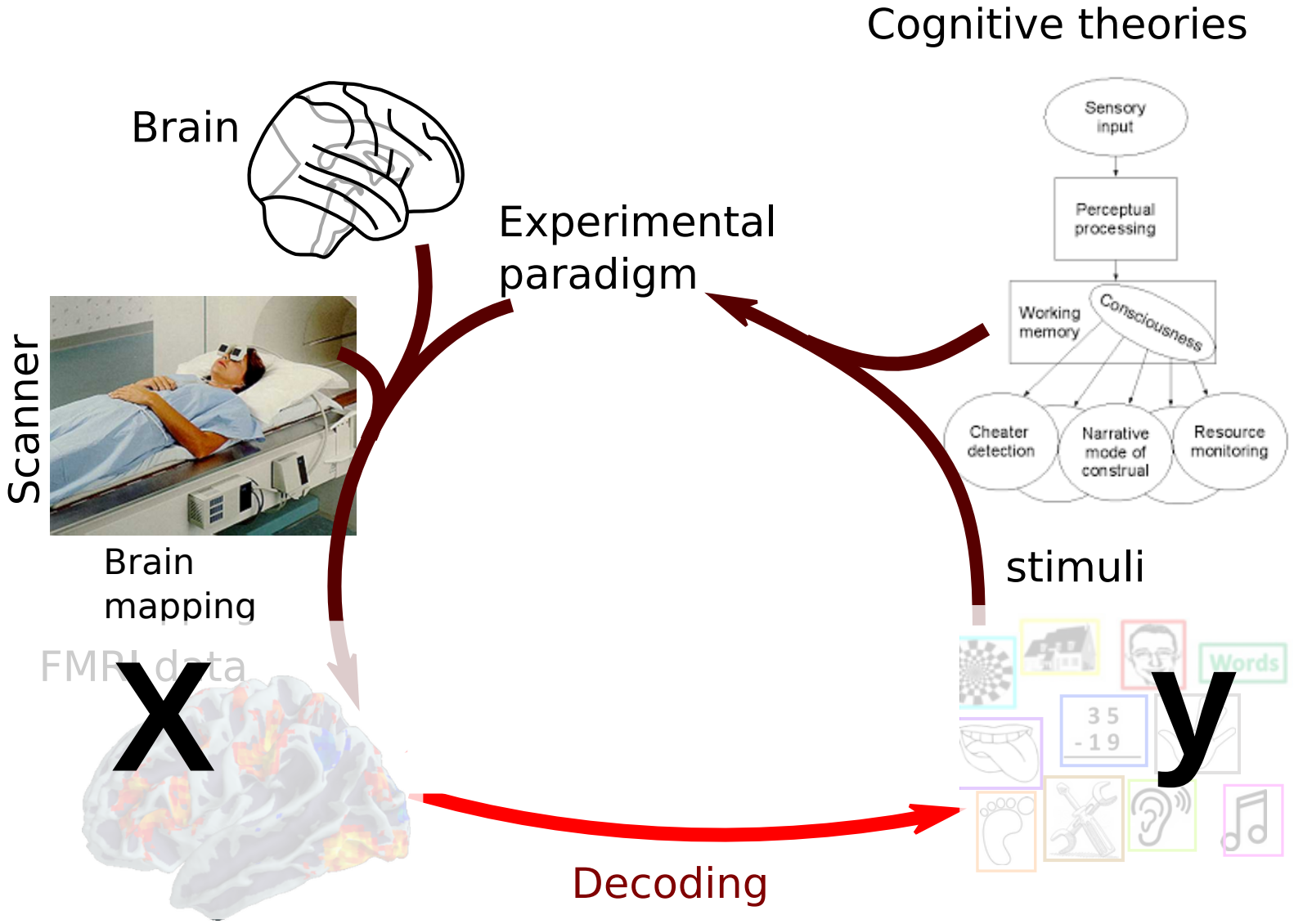
FMRI data



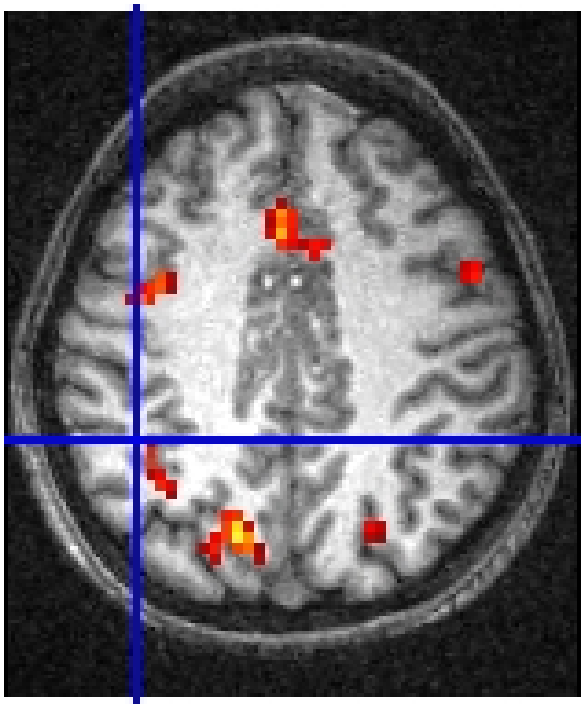
The brain, the mind and the scanner

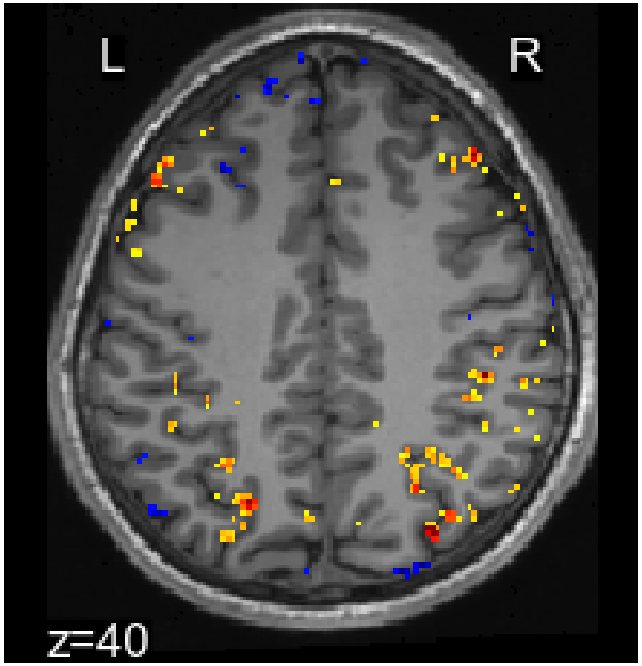


The brain, the mind and the scanner



Resolution increases





2007:
3 mm

2014:
1.5 mm

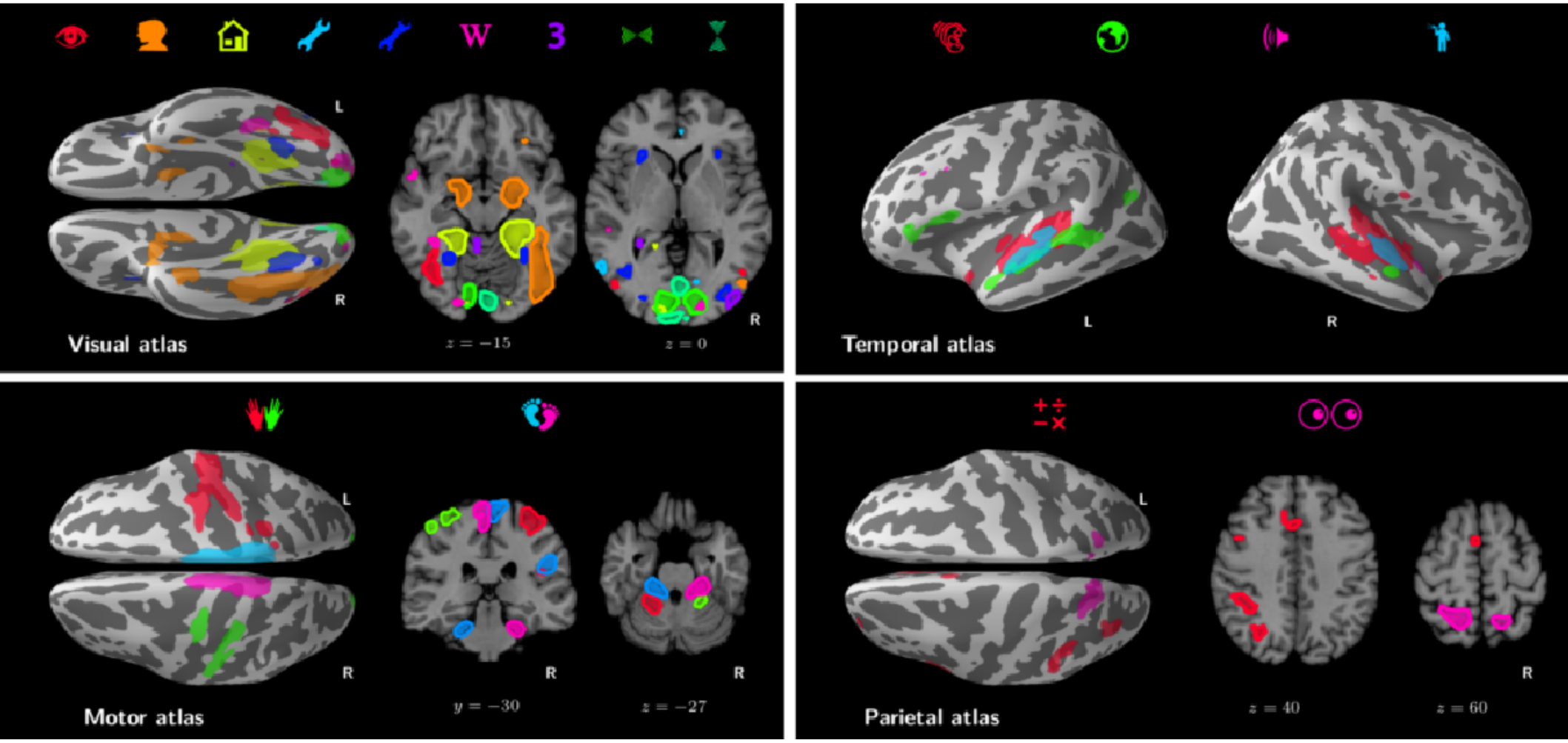
2021:
0.5 mm ?

p = 50,000

p = 400,000

p = 10^7

Joint encoding and decoding



[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018]

Use of linear models

- Brain activity decoding = small n, large p
- Low SNR
- No translation equivariance
- Linear (possibly multi-layer) models still outperform neural nets [He et al. Nimg 2019, Mensch et al. Subm.]

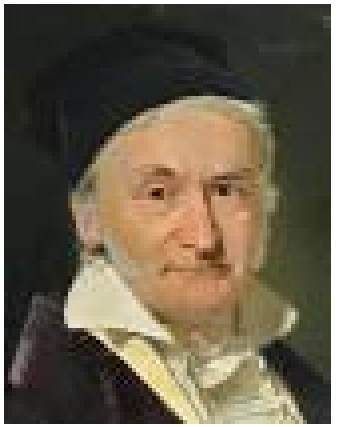
Inference in linear models

- **n > p** → OLS

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

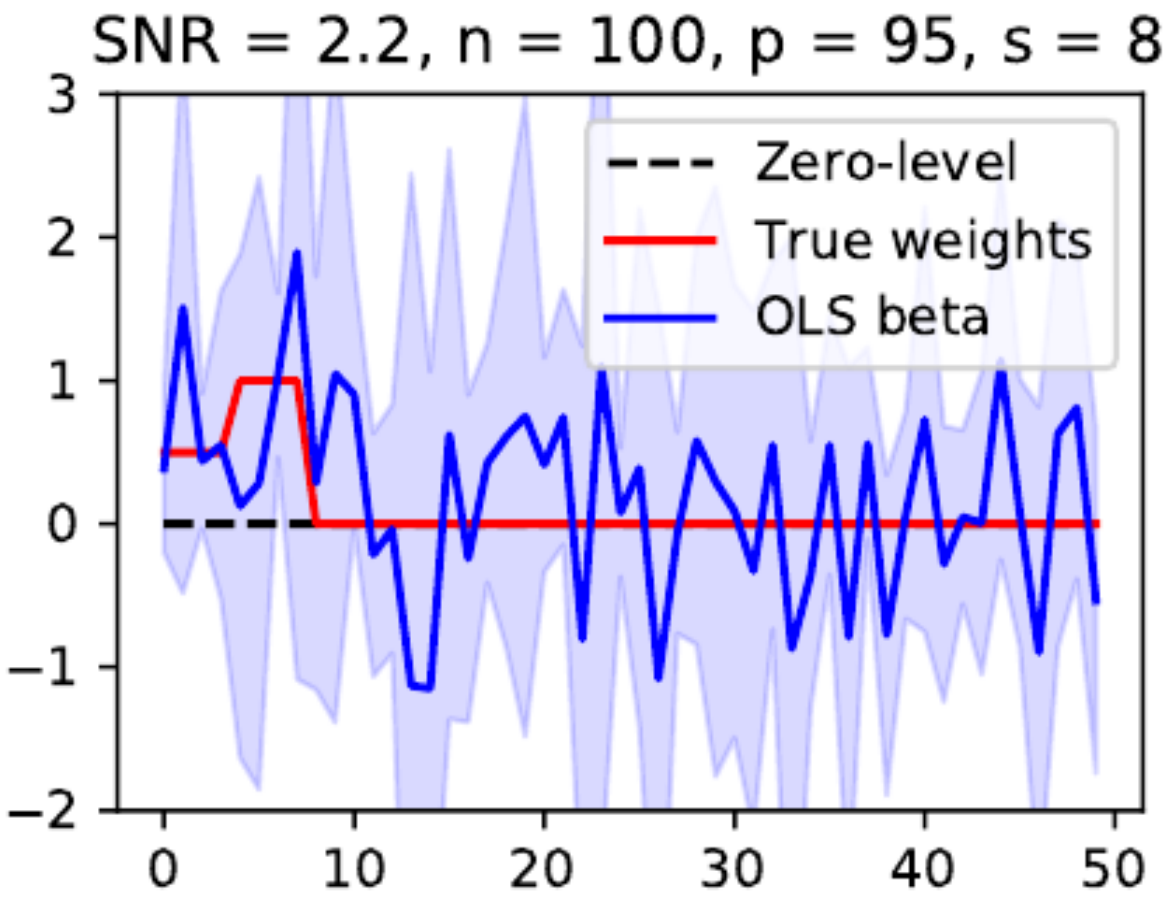
$$\hat{\mathrm{Cov}}(\mathbf{w}) = (\mathbf{X}^T\mathbf{X})^{-1}\frac{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2}{n}$$

$$t = \frac{\hat{\mathbf{w}}_i}{\sqrt{\hat{\mathrm{Cov}}(\mathbf{w})_{ii}}}$$



- As long as X is well-conditioned
- Multiple testing problem: non-independent stats

Classical inference breaks for p≅n



OLS regression when $p \approx n$

- Variance explosion,

  p>n non-defined solution
- Estimator classically needs regularization (Ridge, Lasso)
- But what about statistical inference ?

Statistical inference on w

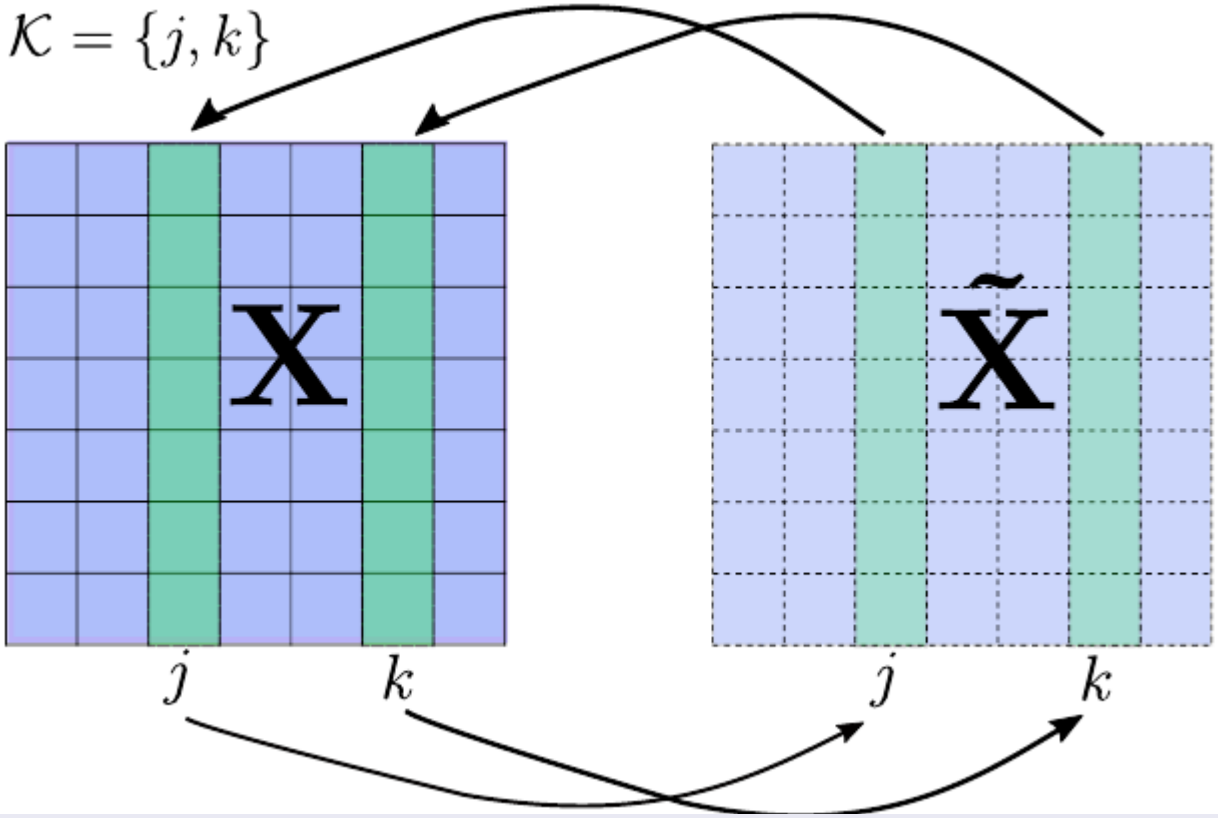
- Inference: find $\{j: \mathbf{w}_j > 0\}$ with some statistical guarantees

- Standard solutions for high-dimensional linear models ($p \cong n$)
  - knockoffs [Candès 2015+]
  - Desparsified Lasso [Zhang & Zhang 2014, Montanari 2014]
  - Multi-split [Meinshausen 2009], Corrected ridge [Bühlmann 2013]

Outline

- p~n
  - knockoff filters (KO)
    - Improvement by aggregation
  - Desparsified lasso (DL)
- p ≫ n
  - Clustered version: CKO, CDL
  - With aggregation (ensembles)

Definition of Knockoff



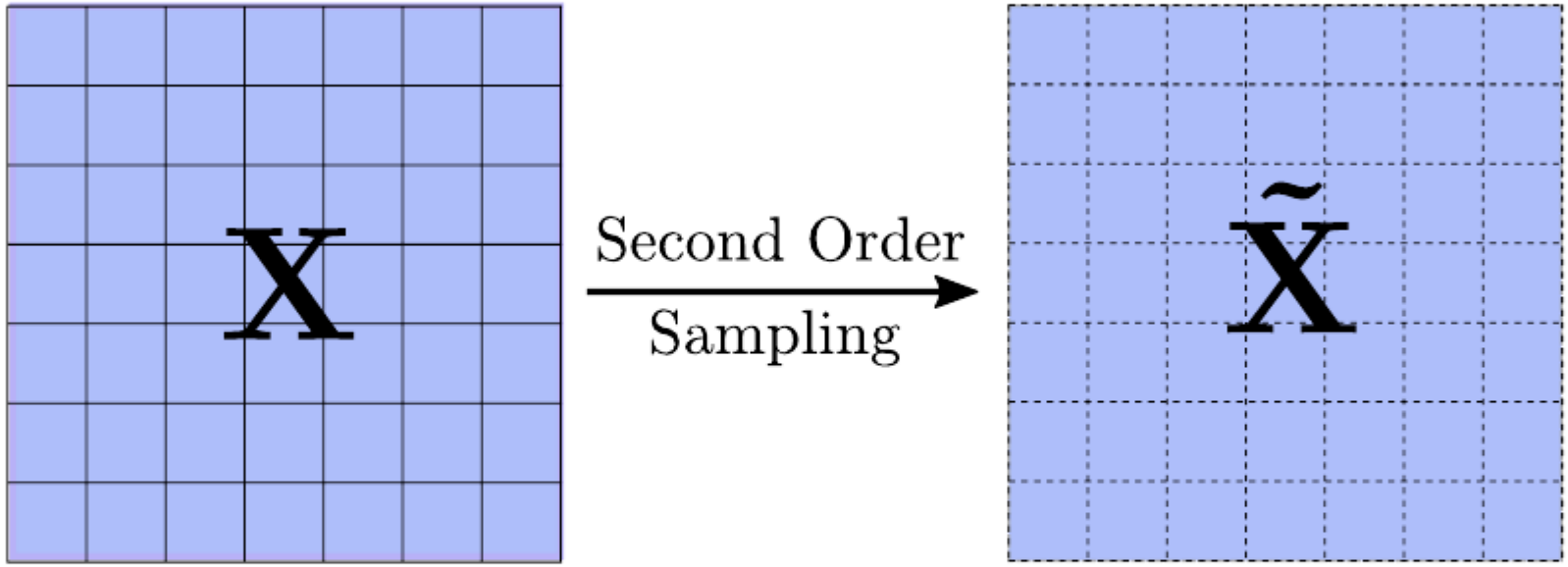
[Barber & Candès 2015]

$\tilde{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is model-X knockoffs of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ if and only if:

1. $\forall$ subset $\mathcal{K} \subset \{1, \ldots, p\}$: $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{K})} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$
2. $\tilde{\mathbf{X}} \perp \mathbf{y} \mid \mathbf{X}$

Sampling Knockoffs

$$\mathrm{cov}(\mathbf{X}, \tilde{\mathbf{X}}) = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diags}\{s\} \\ \boldsymbol{\Sigma} - \mathrm{diags}\{s\} & \boldsymbol{\Sigma} \end{bmatrix}$$



Shares the same first 2 moments - mean and covariance:

$$\mathbb{E}[\tilde{\mathbf{X}}] = \mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}, \quad \mathbb{E}[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}] = \boldsymbol{\Sigma} \quad \text{and} \quad \mathbb{E}[\tilde{\mathbf{X}}^T\mathbf{X}] = \boldsymbol{\Sigma} - \mathrm{diag}\{\mathbf{s}\}$$

Knockoff statistic

Step 1

Construct knockoff variables, concatenate $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$

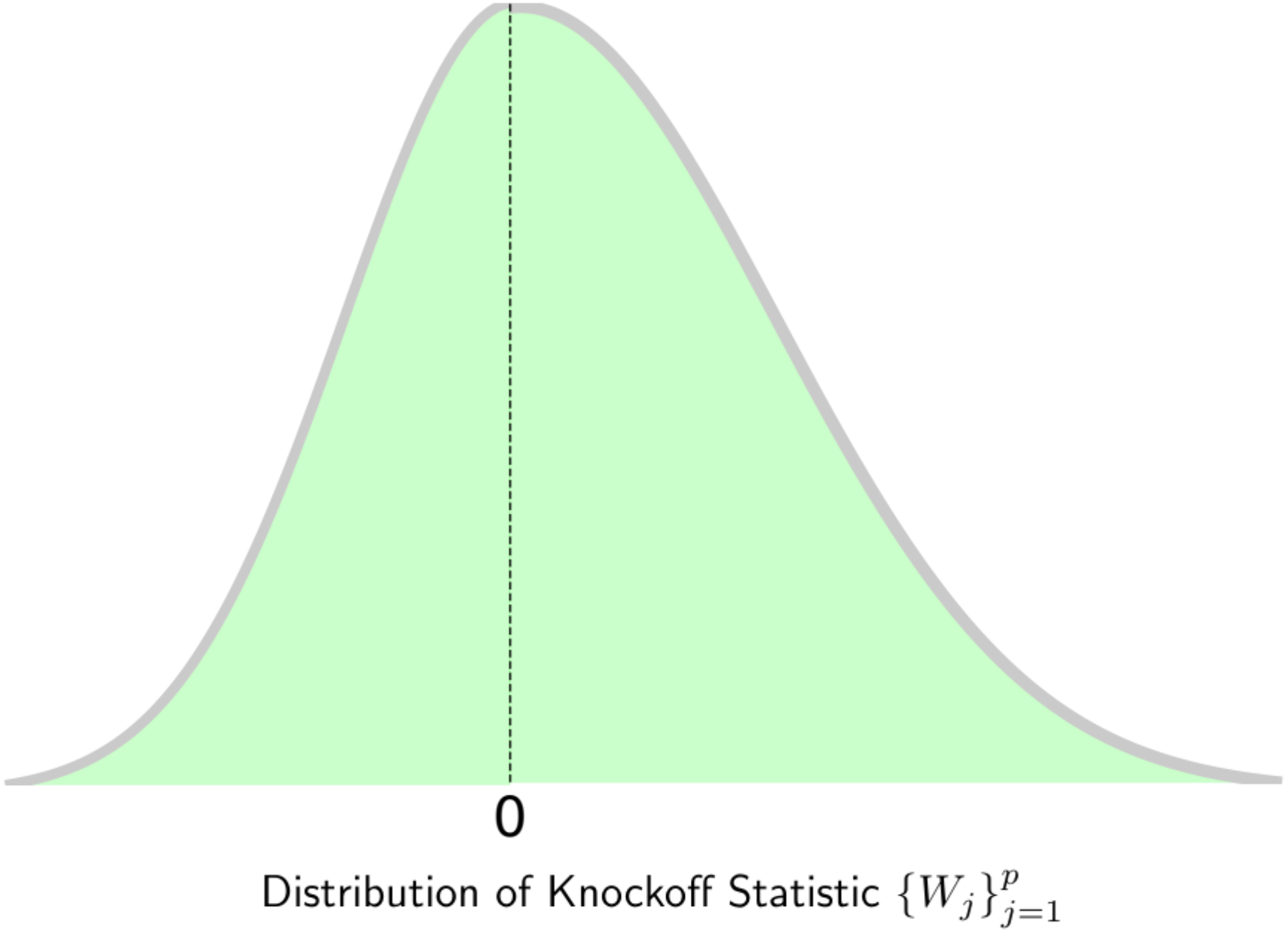
Step 2

Calculate knockoff test-statistics: *Lasso coefficient-difference*, obtain

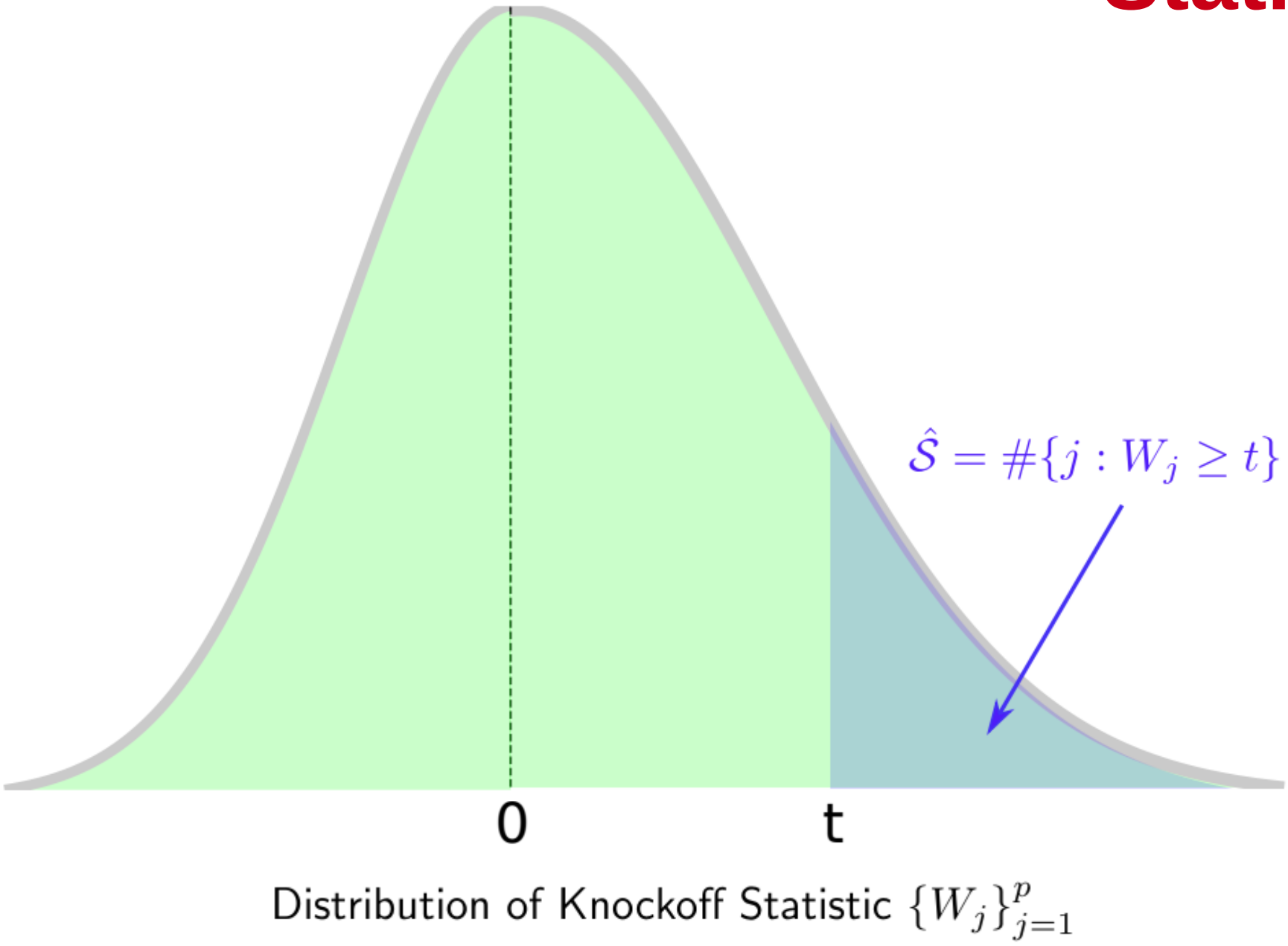
$$\hat{\boldsymbol{\beta}} = \min_{\mathbf{w} \in \mathbb{R}^{2p}} \frac{1}{2} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

then take the difference: $W_j = \left|\hat{\beta}_j(\lambda)\right| - \left|\hat{\beta}_{j+p}(\lambda)\right|$ for each j

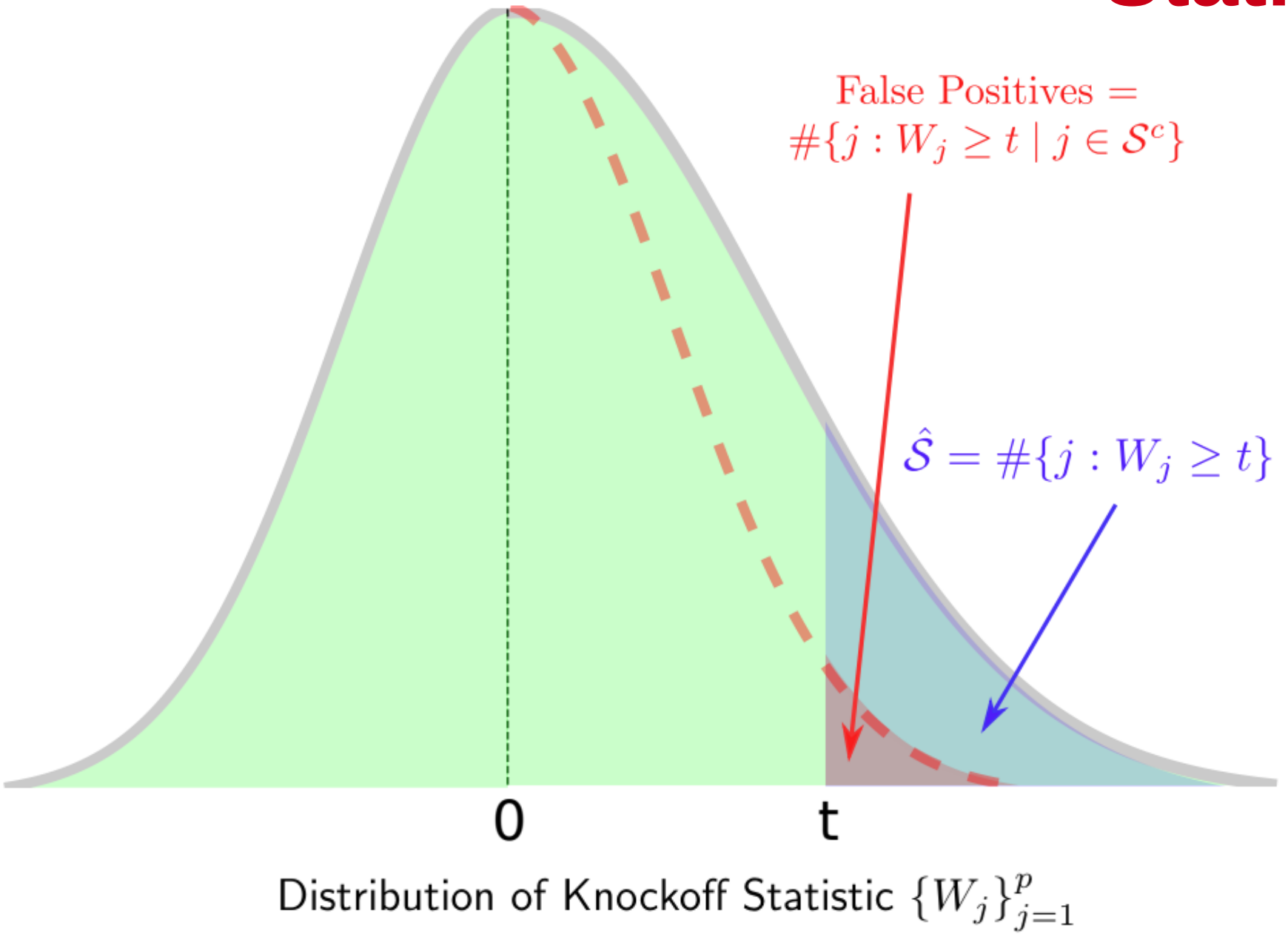
FDP estimation with Knockoff Statistic



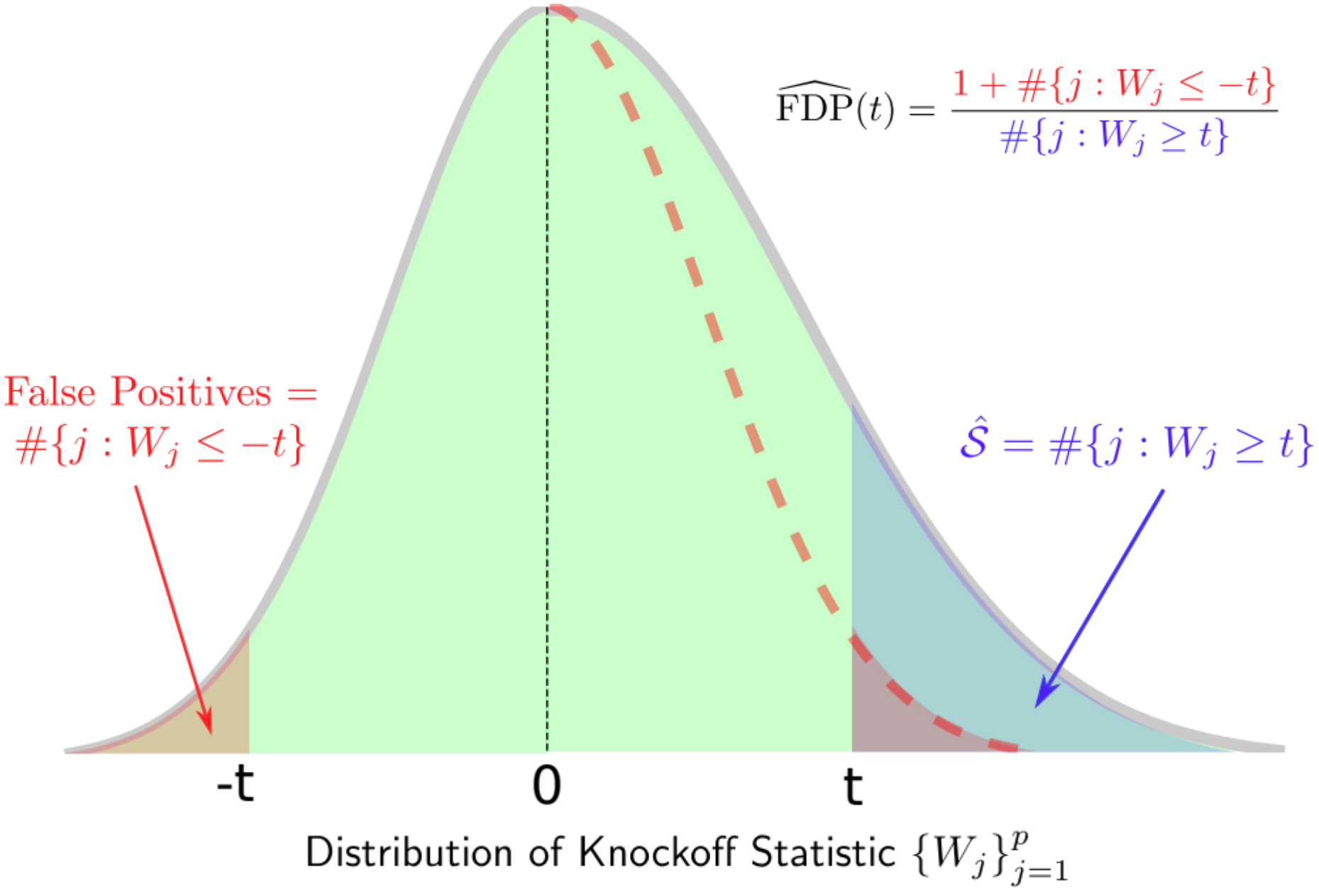
FDP estimation with Knockoff Statistic



FDP estimation with Knockoff Statistic



FDP estimation with Knockoff Statistic



Problem:
Instability of knockoff estimates

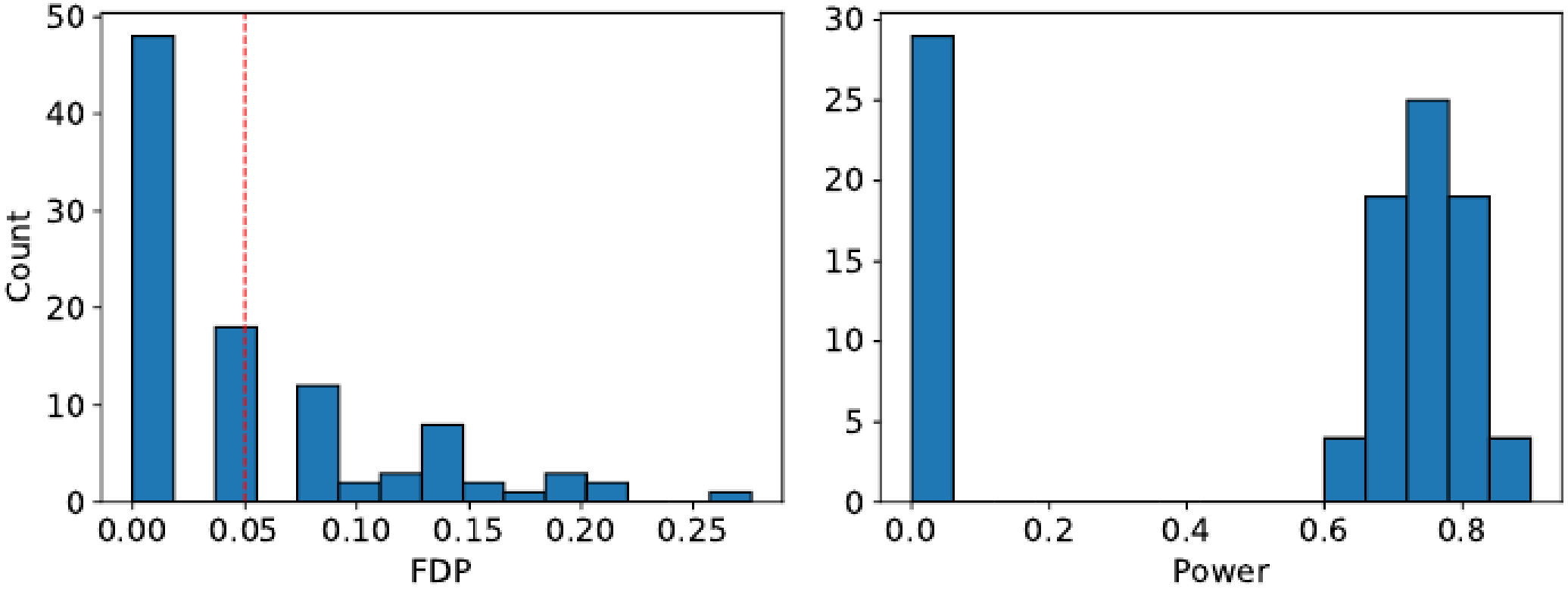
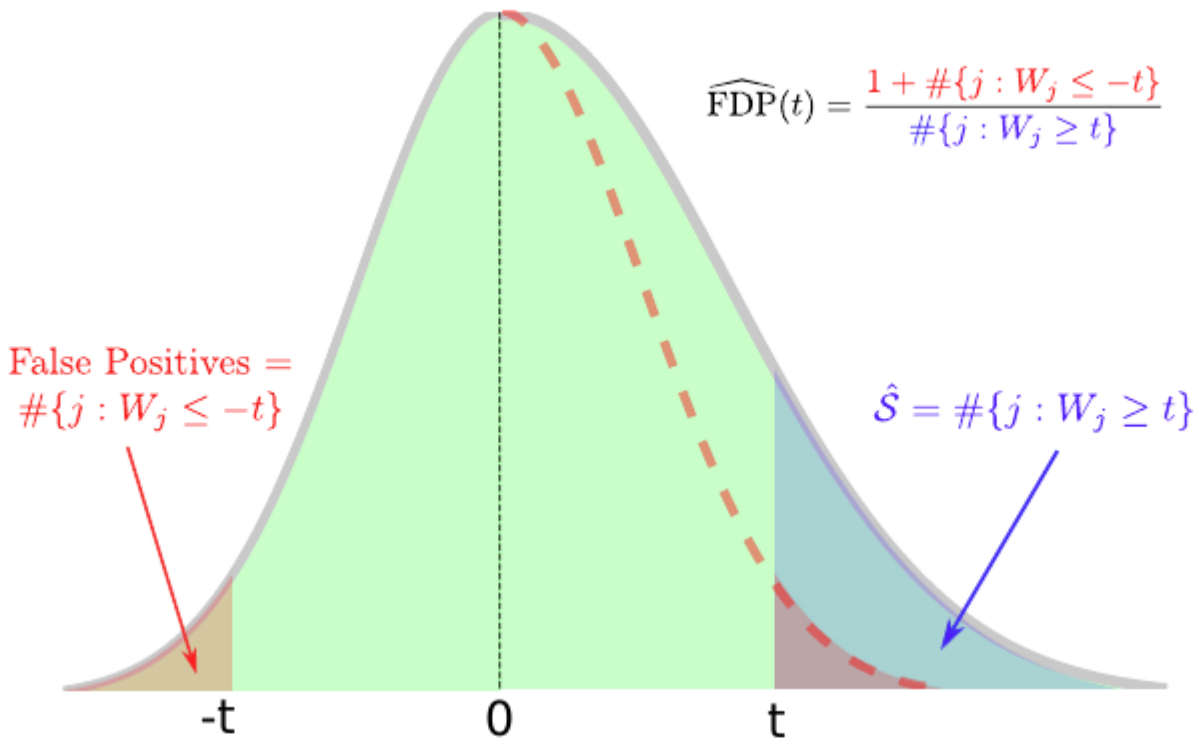


Figure: 100 runs of knockoff inference on the **same simulation**
n=500, p=1000, snr=3.0, $\rho = 0.7$, sparsity $= 0.06$

Intermediate p-values



Introduce the intermediate p-values: convert Knockoff statistic W_j to π_j:

$$\pi_j = \begin{cases} \dfrac{1+\#\{k : W_k \le -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \le 0 \end{cases}$$

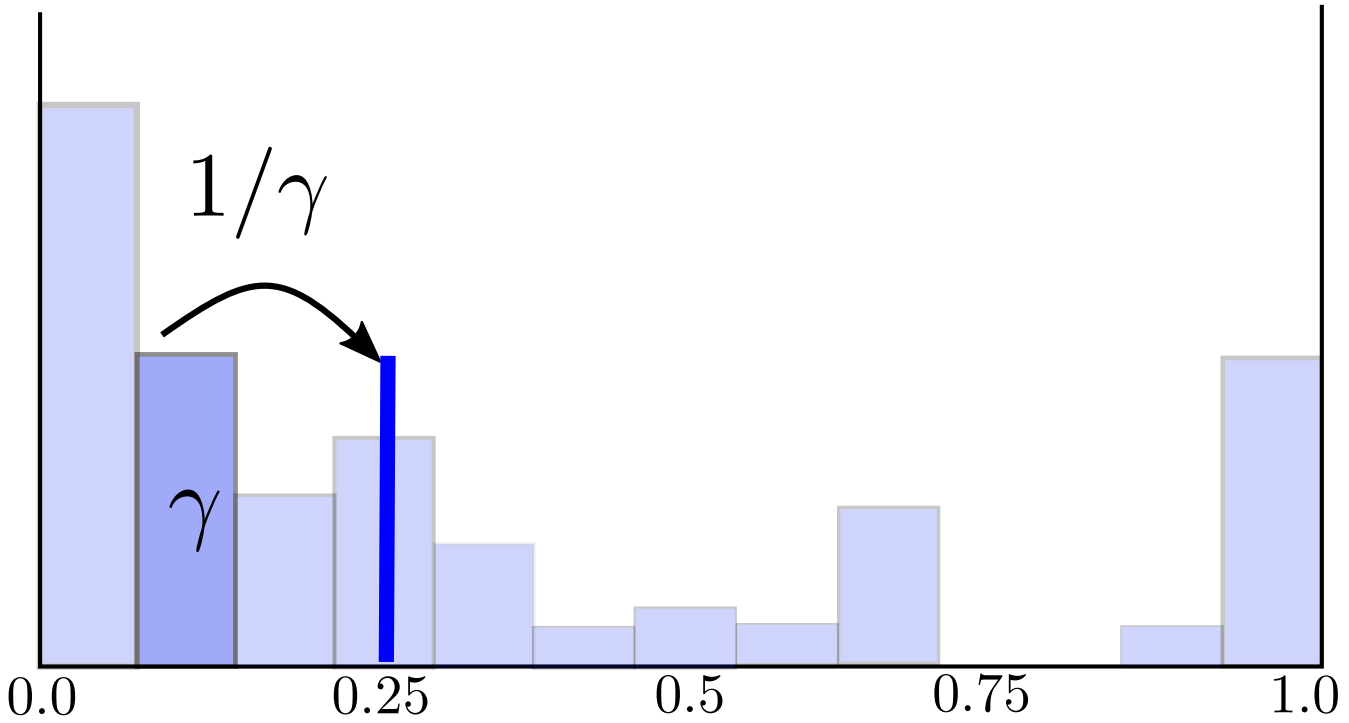
Knockoff aggregation

Step 1: For $b = 1, 2, \ldots, B$:

- Run knockoff sampling, calculate test statistic $\left\{W_j^{(b)}\right\}_{j \in [p]}$
- Convert the test statistic $W_j^{(b)}$ to $\pi_j^{(b)}$:

$$\pi_j^{(b)} = \begin{cases} \dfrac{1 + \#\{k : W_k^{(b)} \leq -W_j^{(b)}\}}{p} & \text{if} \quad W_j^{(b)} > 0 \\ 1 \quad \text{if} \quad W_j \leq 0 \end{cases}$$

Knockoff aggregation



Step 2 – Quantile Aggregation of p-values (Meinshausen et al., 2009)

$$\bar{\pi}_j = \min\left\{ \frac{q_\gamma(\pi_j^{(b)})}{\gamma}, 1 \right\} \quad \forall j \in [p]$$

For $\gamma \in (0, 1)$ with $q_\gamma(\cdot)$ the empirical γ-quantile function.

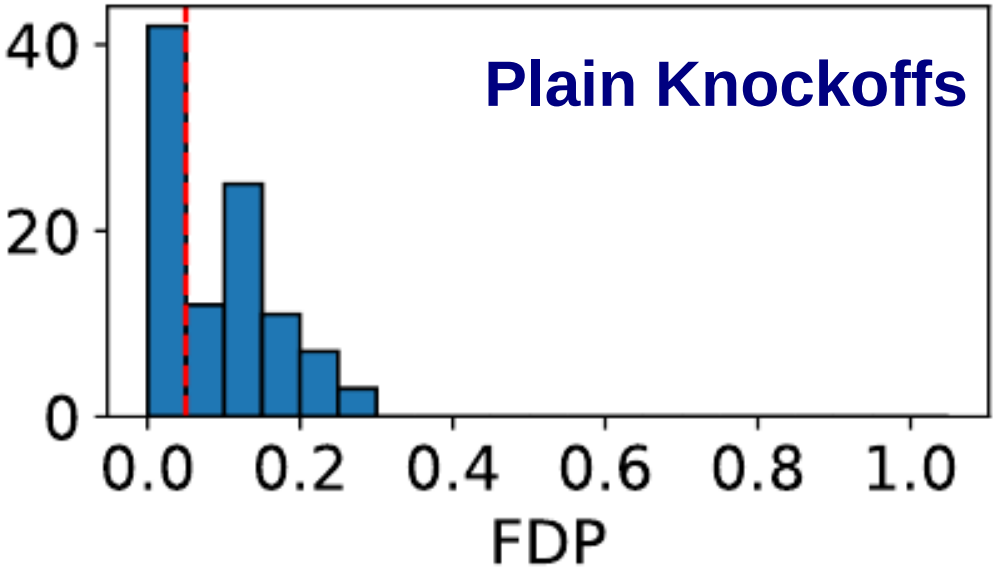
Knockoff aggregation

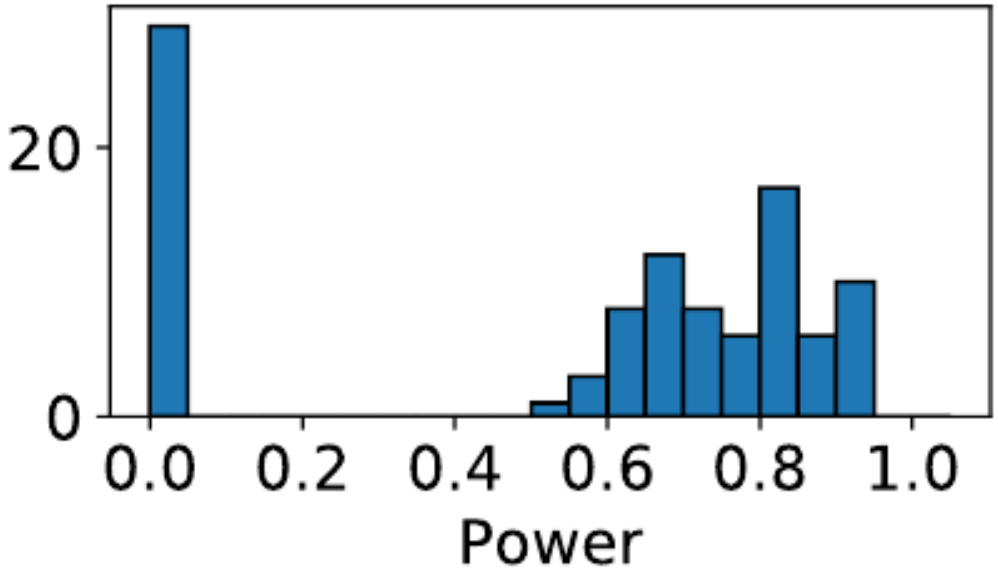
Step 3 – FDR control with $\bar{\pi}$

- Order $\bar{\pi}_j$ ascendingly: $\bar{\pi}_{(1)} < \bar{\pi}_{(2)} \cdots < \bar{\pi}_{(p)}$
- Given FDR control level $\alpha \in (0, 1)$, find largest k such that:
  - $\bar{\pi}_{(k)} \leq k\alpha/p$ (Benjamini and Hochberg, 1995), or
  - $\bar{\pi}_{(k)} \leq \dfrac{k\alpha}{p \sum_{i=1}^{p} 1/i}$ (Benjamini and Yekutieli, 2001)

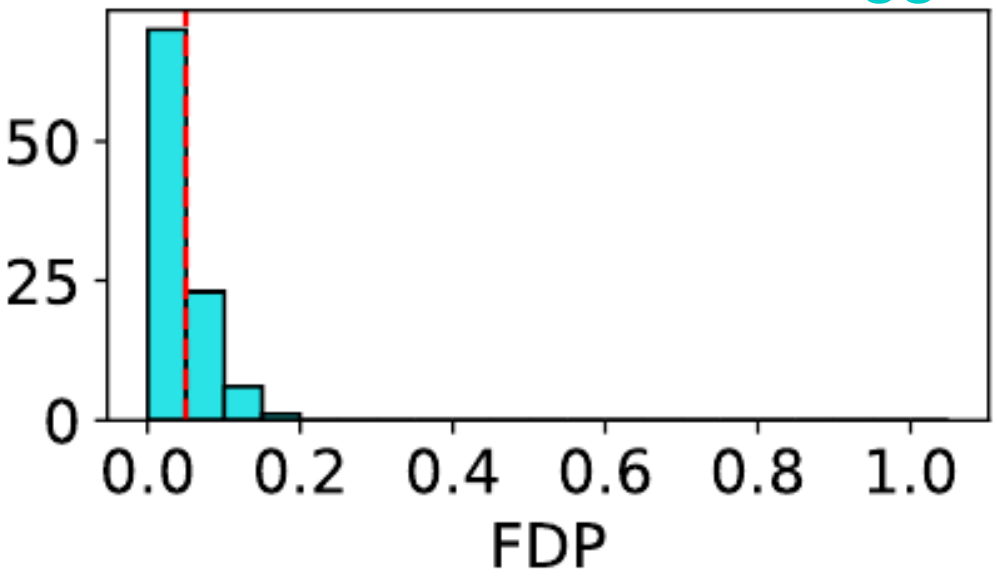
  $\longrightarrow$ FDR threshold: $\tau = \bar{\pi}_{(k)}$
- $\hat{\mathcal{S}}_{AKO} = \{j : \bar{\pi}_j \leq \tau \mid j \in [p]\}$

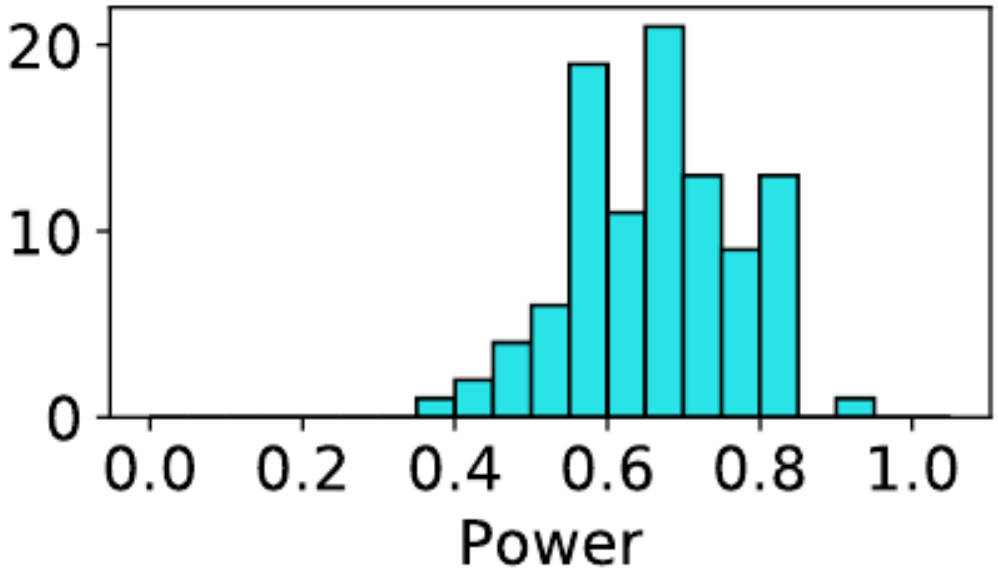
Empirical results: more stability





Aggregated Knockoffs





Brain activity decoding example

- Data: Human Connectome Project
- Objective: predict the experimental condition per task given brain activity
- $n = 900$ subjects, $p \approx 212000$
- Preprocessing: dimension reduction by clustering

$$p = 212000 \longrightarrow p = 1000$$

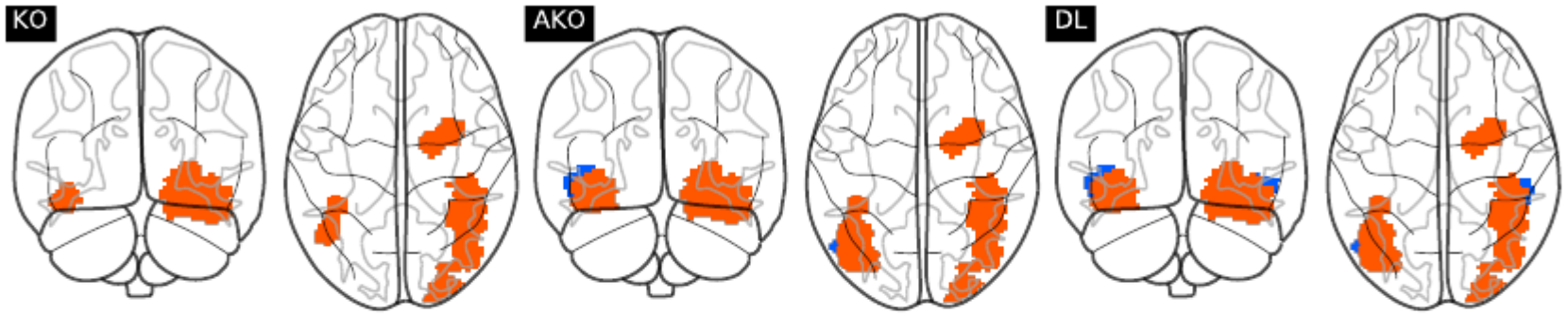
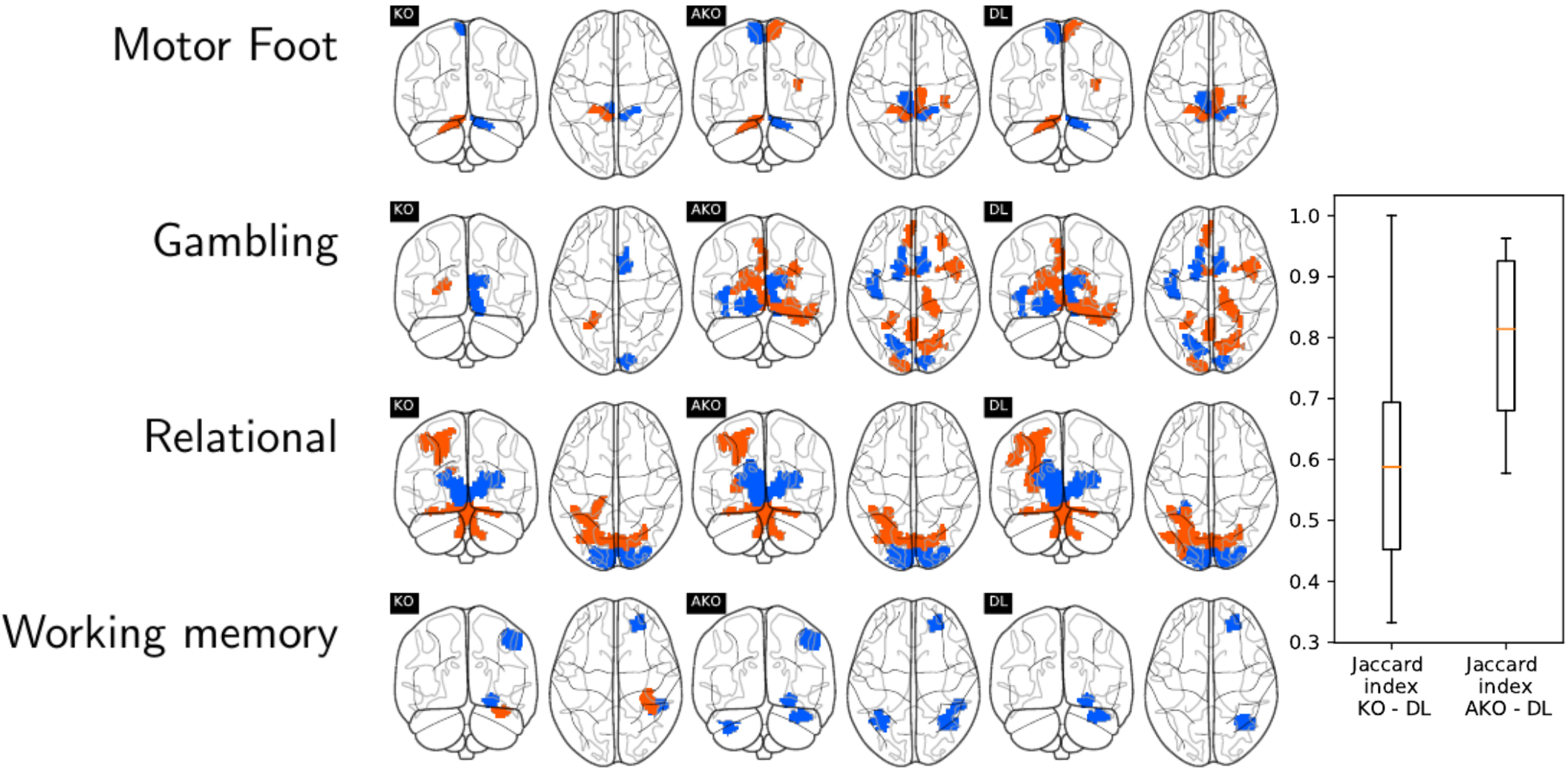


Figure: Detection of significant brain regions for HCP data (900 subjects)
Selected regions in a reaction with Emotion images task.
Orange: brain areas with positive sign activation.
Blue: brain areas with negative sign activation

Brain activity decoding example

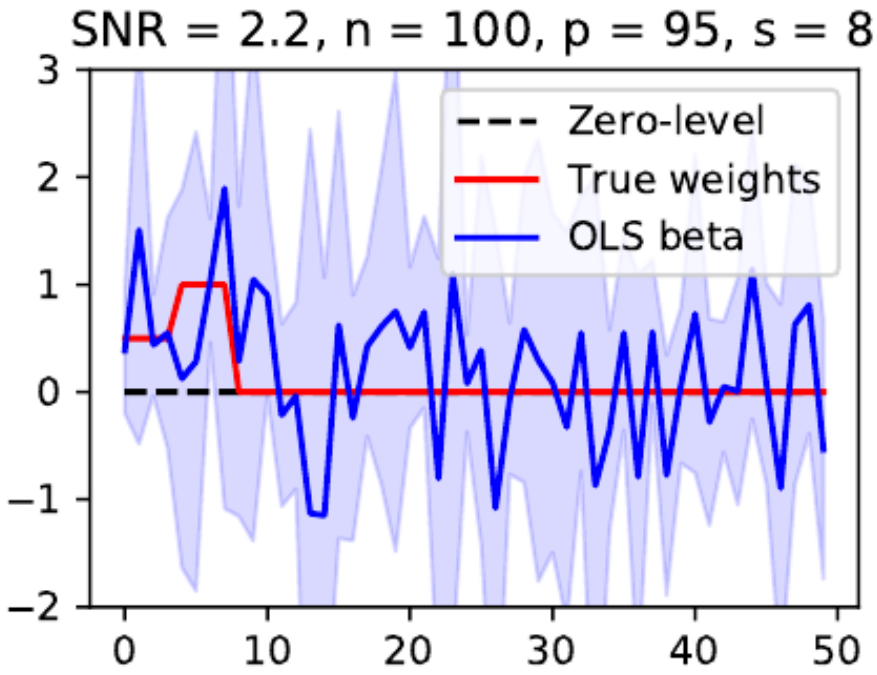


Aggregation of Multiple Knockoffs

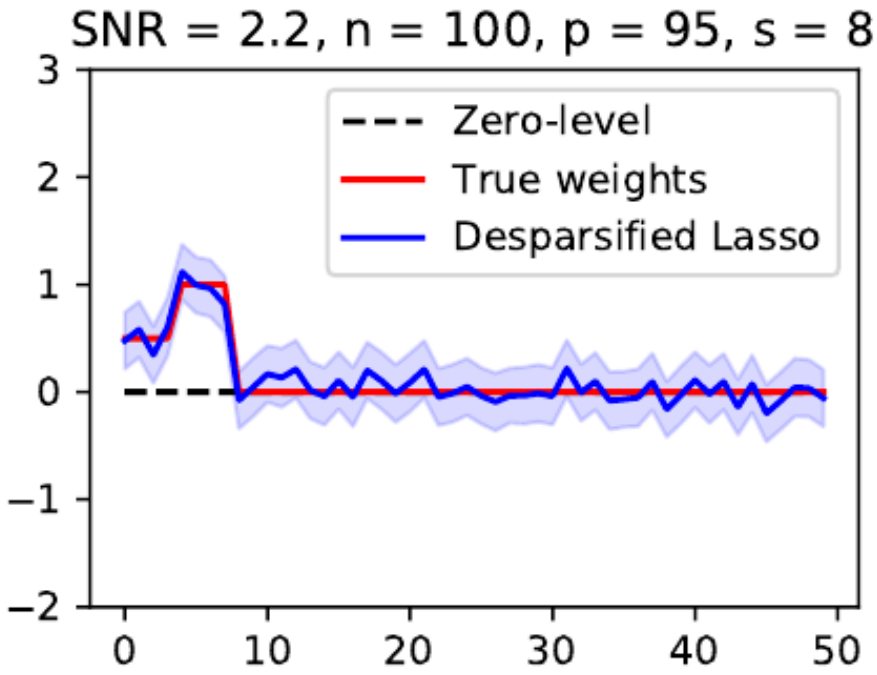
Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, Sylvain Arlot

Desparsified Lasso

- **Comparing OLS and Desparsified Lasso solutions:**



OLS regression when $p \approx n$



Desparsified Lasso when $p \approx n$

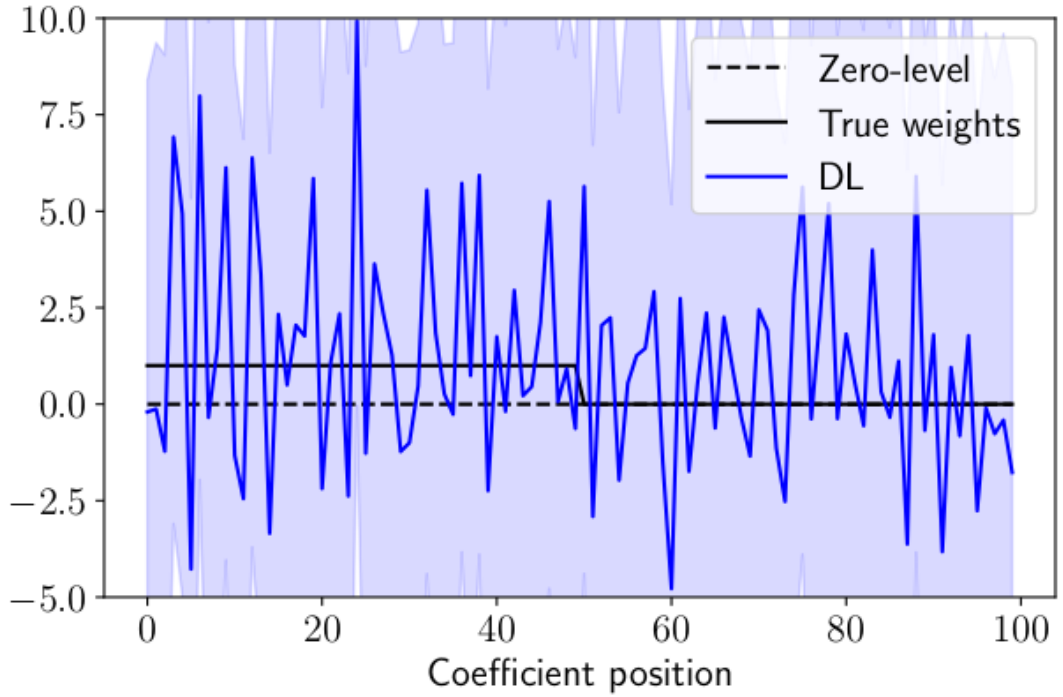
Interim conclusion

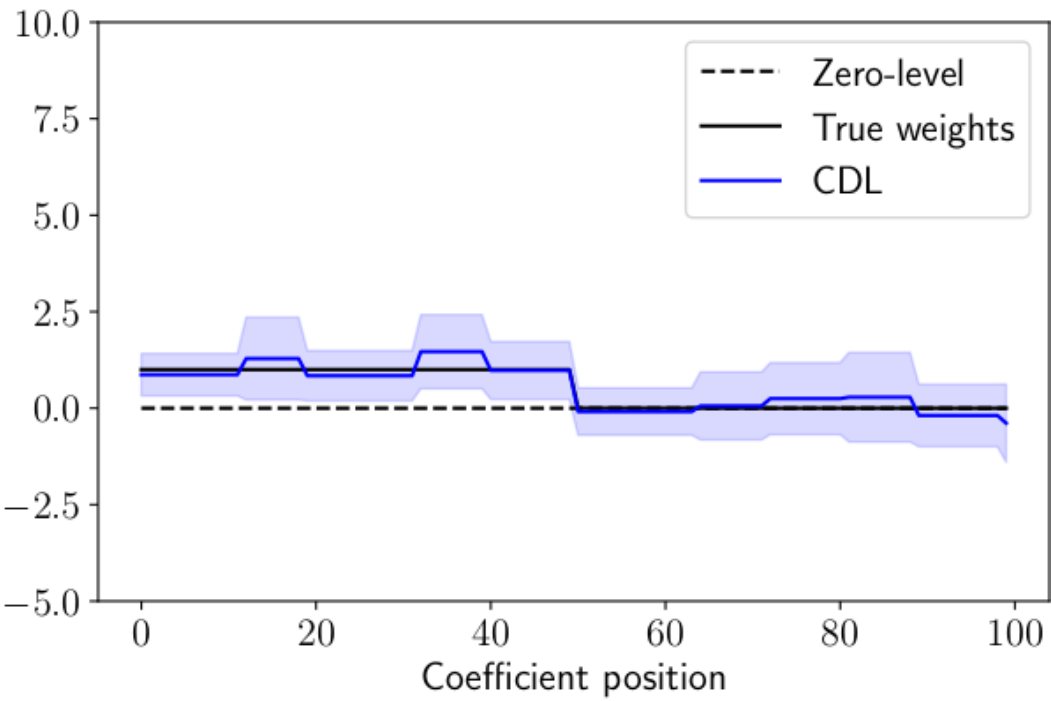
Desparsified lasso	Knockoff
+ powerful	+ versatile
+ deterministic	+ aggregation
- only linear regression	- power (FWER control)

- Aggregated knockoff and DL make inference possible with p > n
- Are we done ?
- Not exactly: clustering was essential in making this work ...

Large p → need dimension reduction

p=2000, n=100





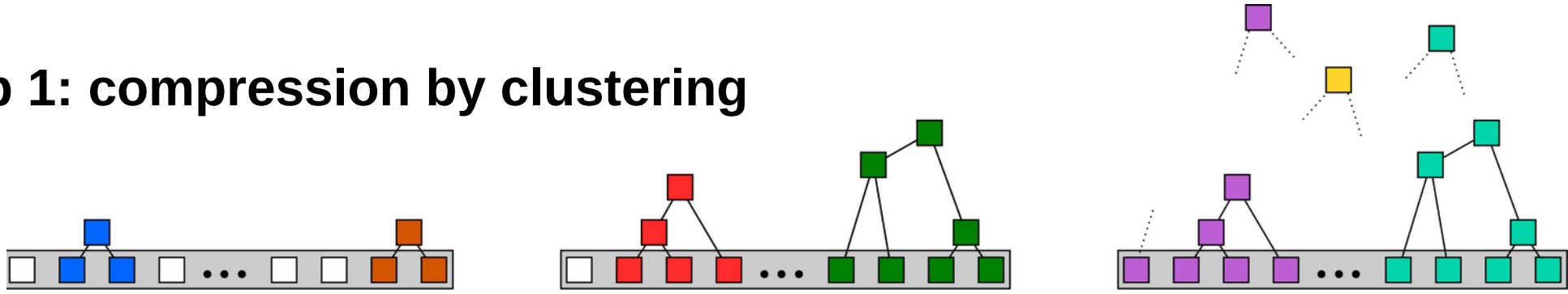
Large p kills statistical power

CDL tames variance

[Chevalier et al. MICCAI 2018]

Adaptation to brain imaging

Step 1: compression by clustering



Step 2: inference on compressed representations

$$\sigma_*^{-1}(\Omega_{jj})^{-1/2}(\hat{w}_j - w_j^*) \sim \mathcal{N}(0,1)$$

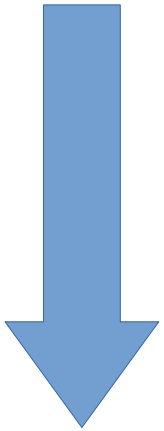
***C**lustered*
***D**esparsified*
***L**asso*

Step 3: ensembling across different clusterings
→ aggregate p-values

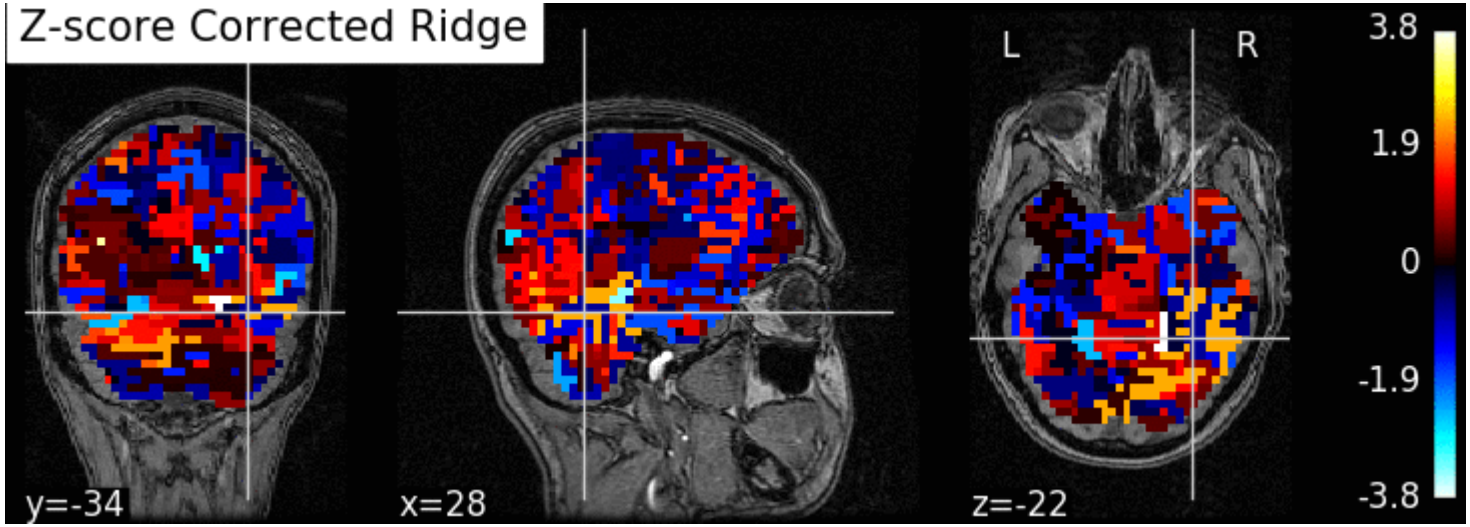
***E**nsemble of*
***C**lustered*
***D**esparsified*
***L**asso*

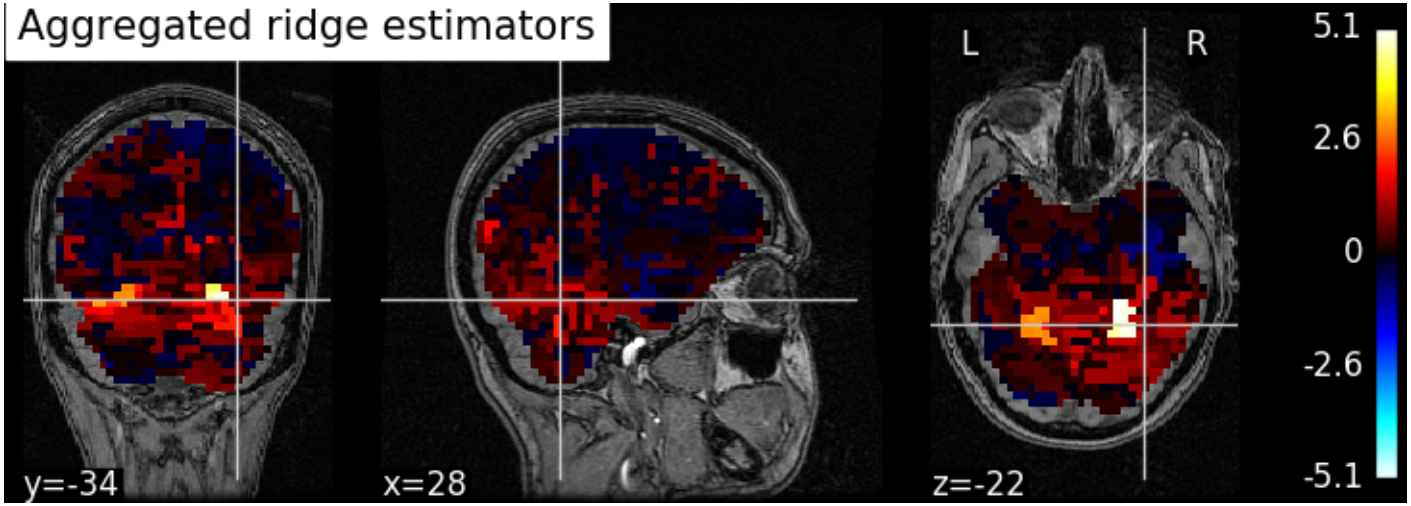
Why we need ensembling

DL p-values from different clusterings

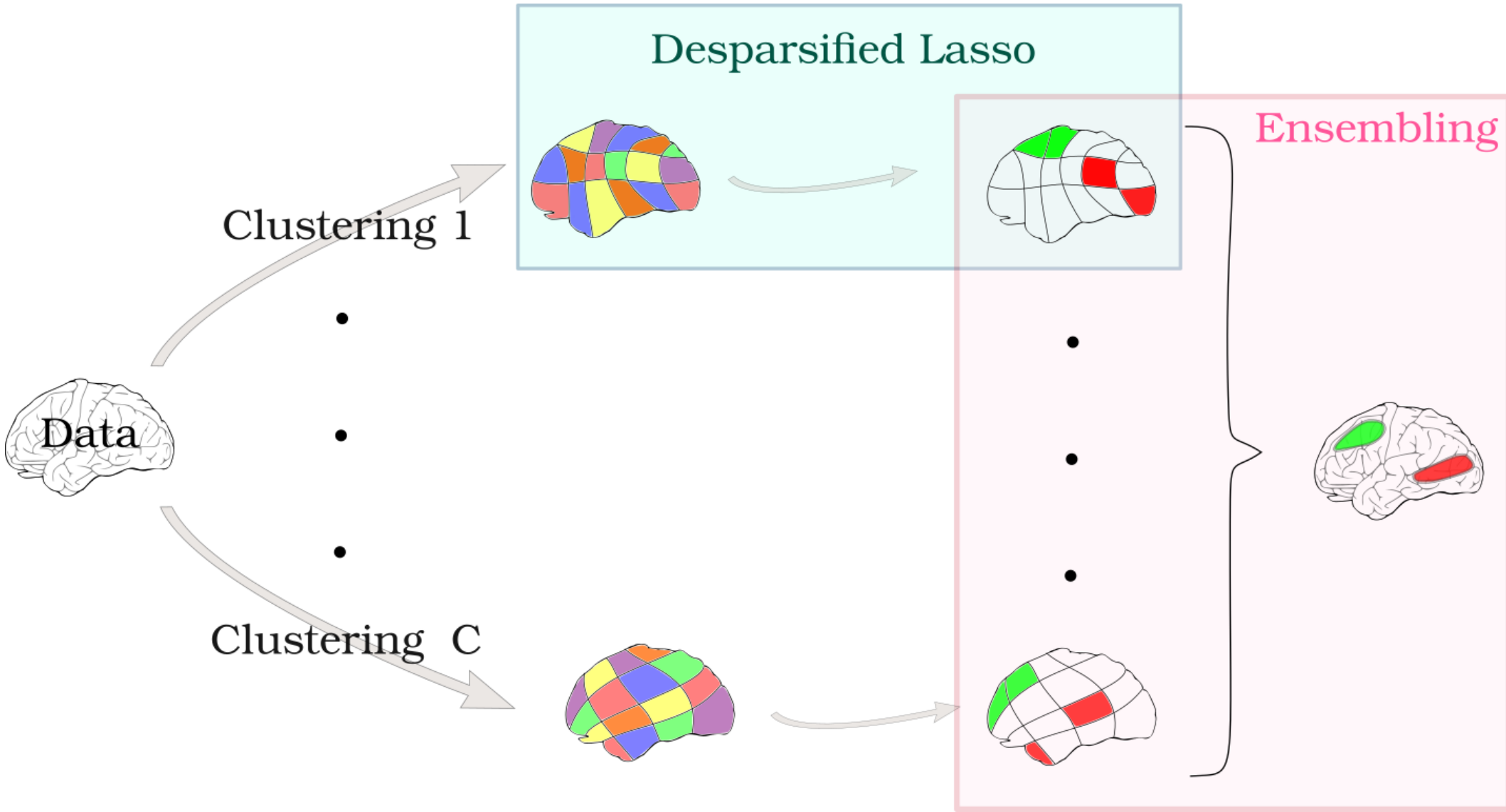


aggregation

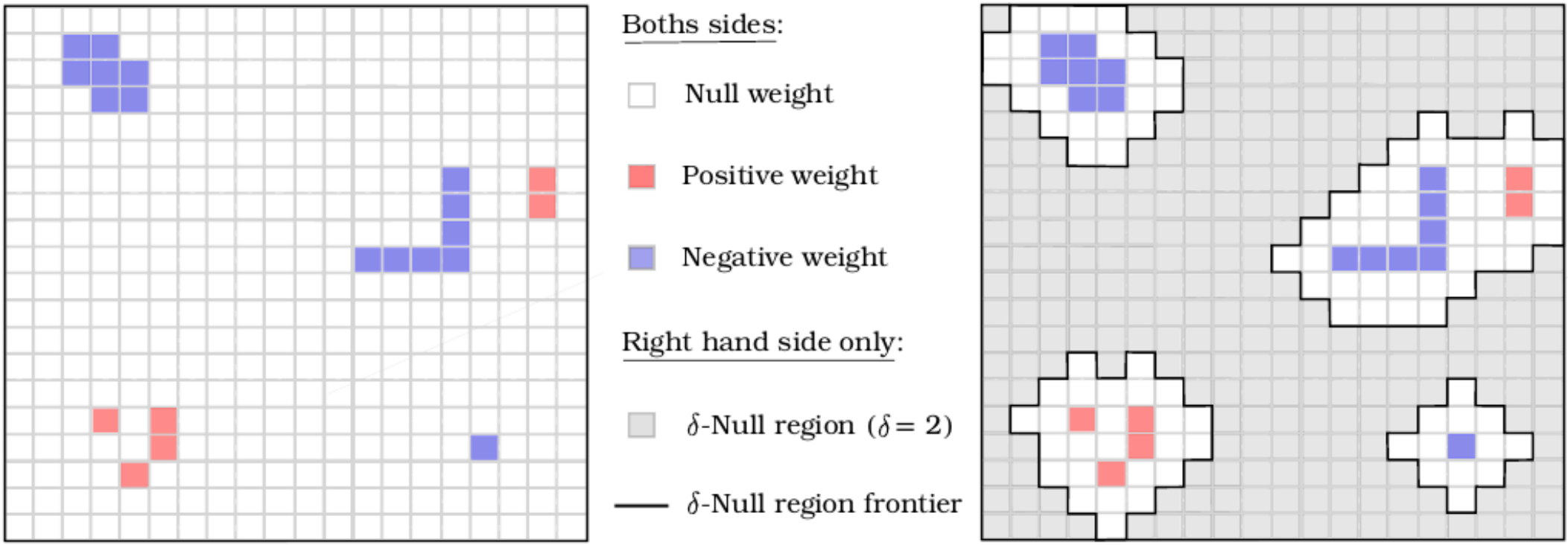




ECDL for brain imaging

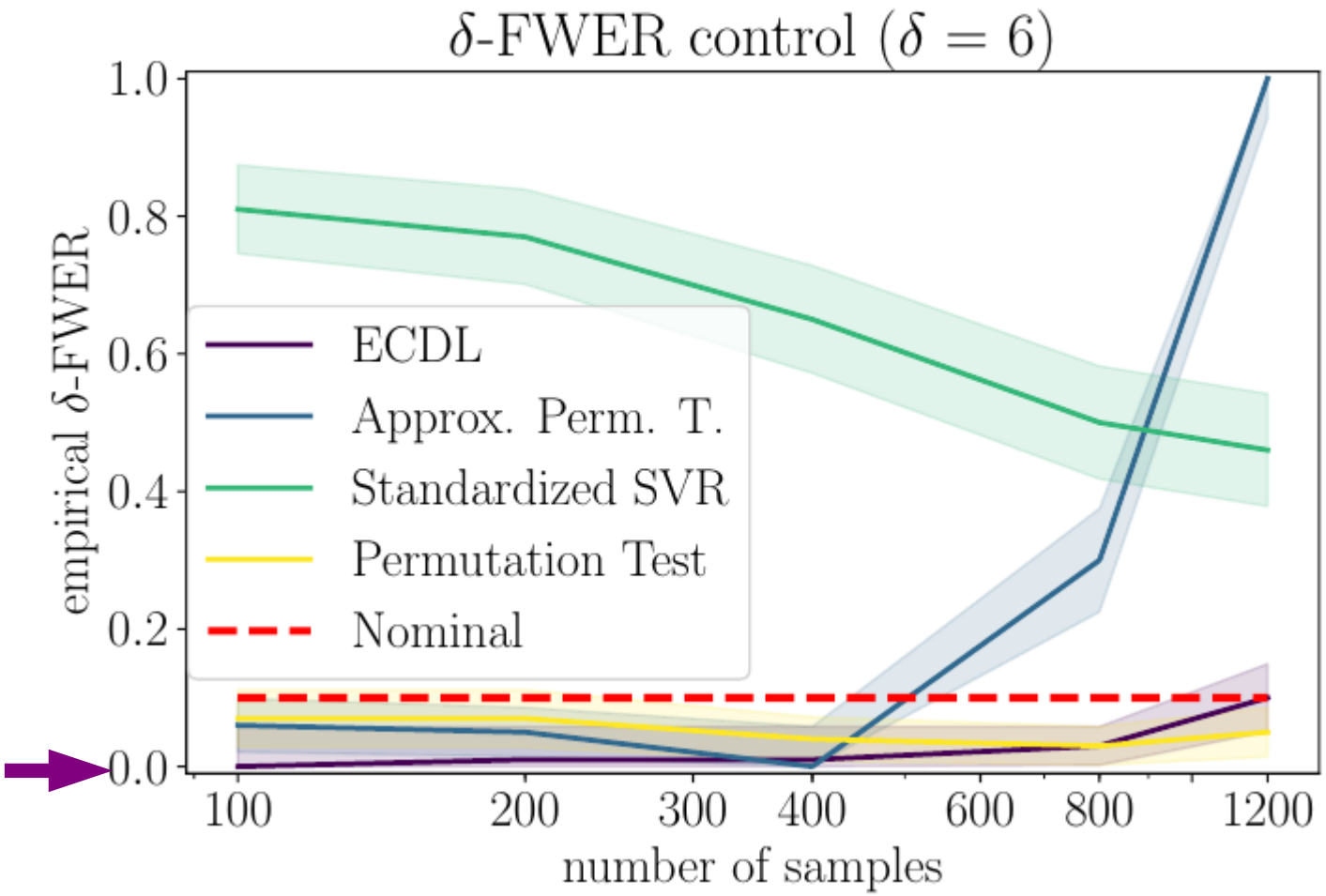


δ-FWER-control



Spatial relaxation on FWER control

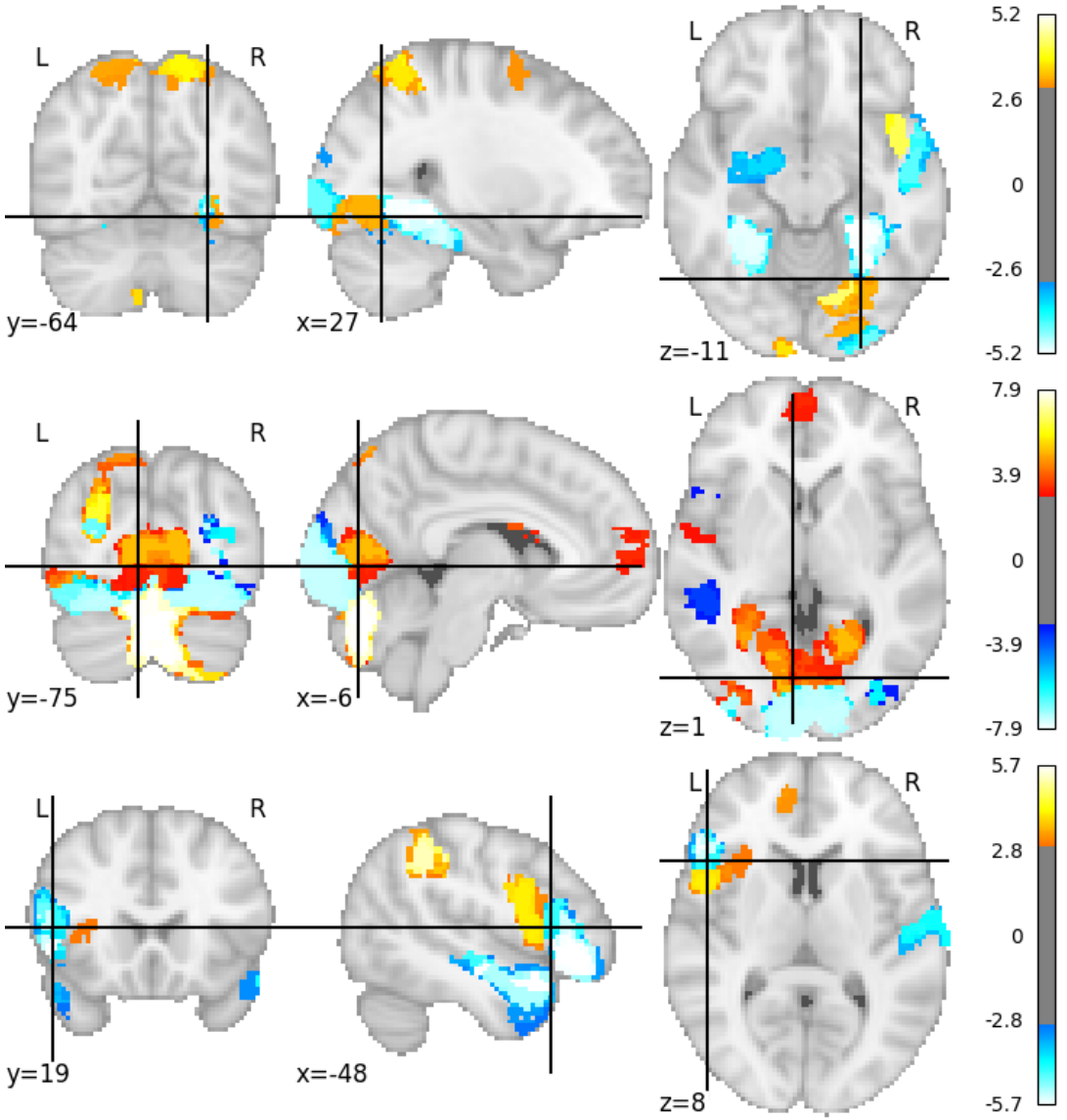
δ-FWER is controlled



δ-FWER control on semi-simulated data, obtained with 100 repetitions for every sample size.

[Chevalier et al. Nimg 2020]

Effects on real data



HCP dataset, n=900

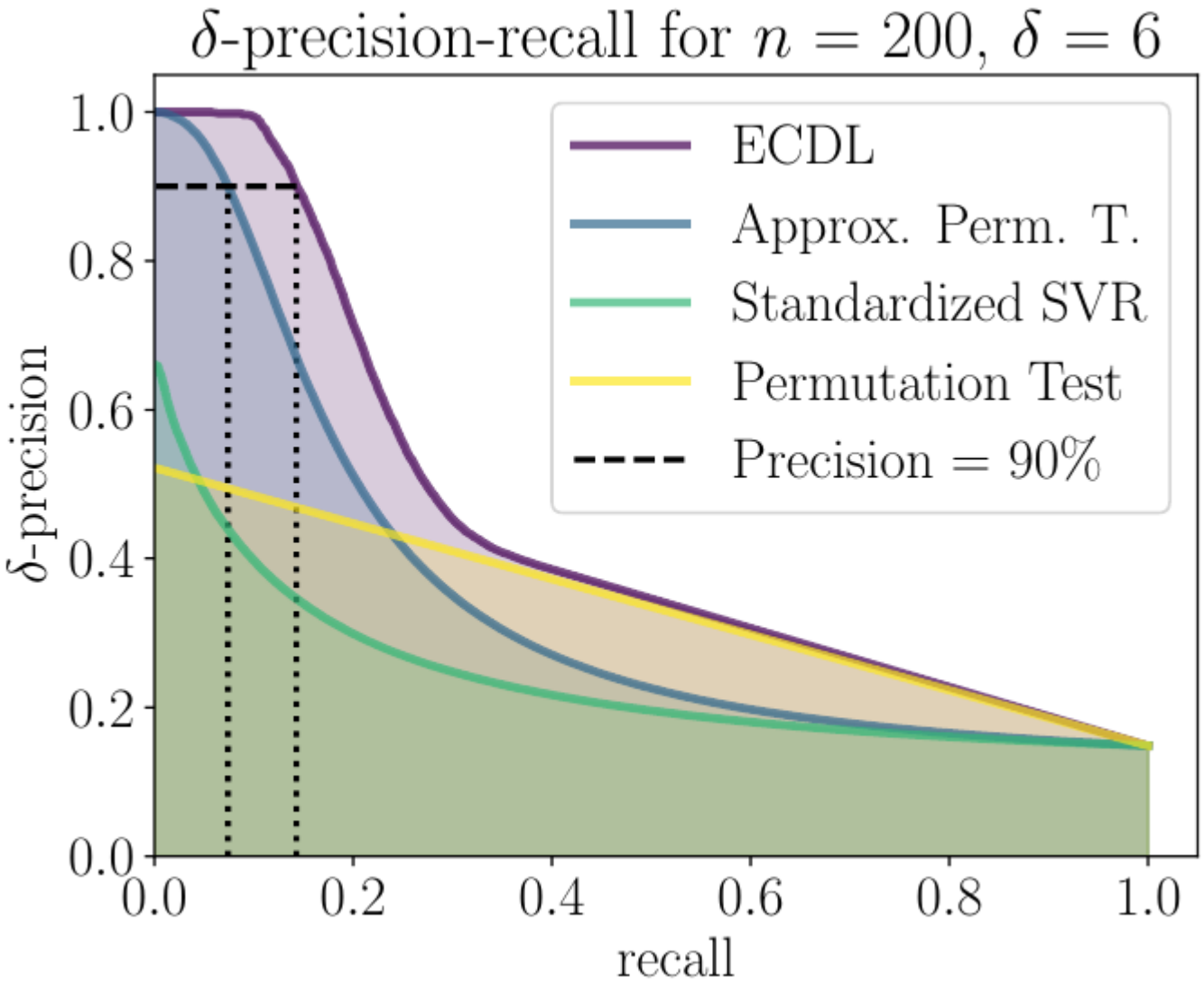
Social cognition

Visual feature discrimination

Language vs maths

[Nguyen et al. IPMI 2019, Chevalier et al. MICCAI 2018]

Results: higher PR curve



- On semi-simulated data, EDCL achieves better PR compromise

[Chevalier et al. Nimg 2020]

Conclusion

- Large-p data bring challenges:
  - Computation cost
  - Difficulty of statistical inference
- Solutions: compression, subsampling, ensembling
- Efficient stochastic regularizers
- Extension toward causal reasoning: R-learner

[Aydore et al. ICML 2019]

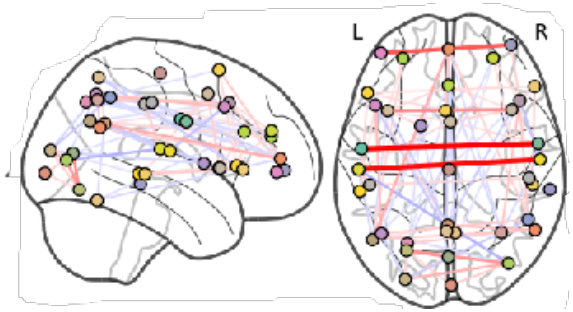
From good ideas to good practices: software

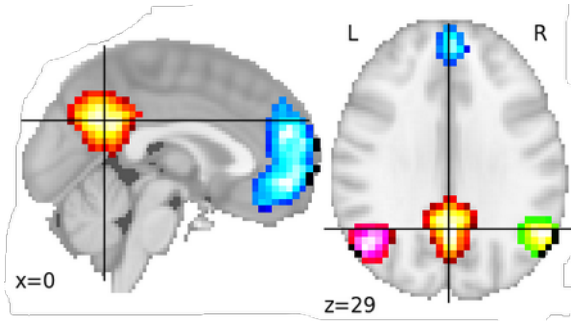
scikits learn
machine learning in Python

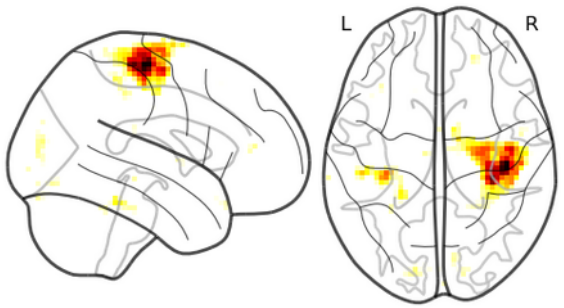
ni learn

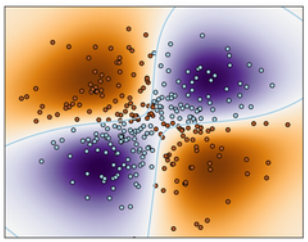
MNE
MEG + EEG ANALYSIS & VISUALIZATION

- Machine learning in Python
- Machine learning for neuroimaging
- BSD licence, Python, OSS









Acknowledgements

Parietal

G. Varoquaux,
A. Gramfort,
P. Ciuciu,
D. Wassermann,
D. Engemann,
B. Nguyen
A.L. Grilo Pinho,
E. Dohmatob,
A. Mensch,
J.A. Chevalier,
A. Hoyos idrobo,
D. Bzdok,
J. Dockès,
P. Cerda,
C. Lazarus
D. La Rocca
G. Lemaitre
L. El Gueddari
O. Grisel
M. Massias
P. Ablin
H. Janati
J. Massich
K. Dadi
H. Richard
C. Petitot

Other collaborators

R. Poldrack,
J. Haxby
C. F. Gorgolevski
J. Salmon
S. Arlot
M. Lerasle

Human Brain Project

université PARIS-SACLAY

ANR