Toward a rigorous causal framework for brain mapping

Bertrand Thirion, bertrand.thirion@inria.fr

INVENTEURS DU MONDE NUMÉRIQUE

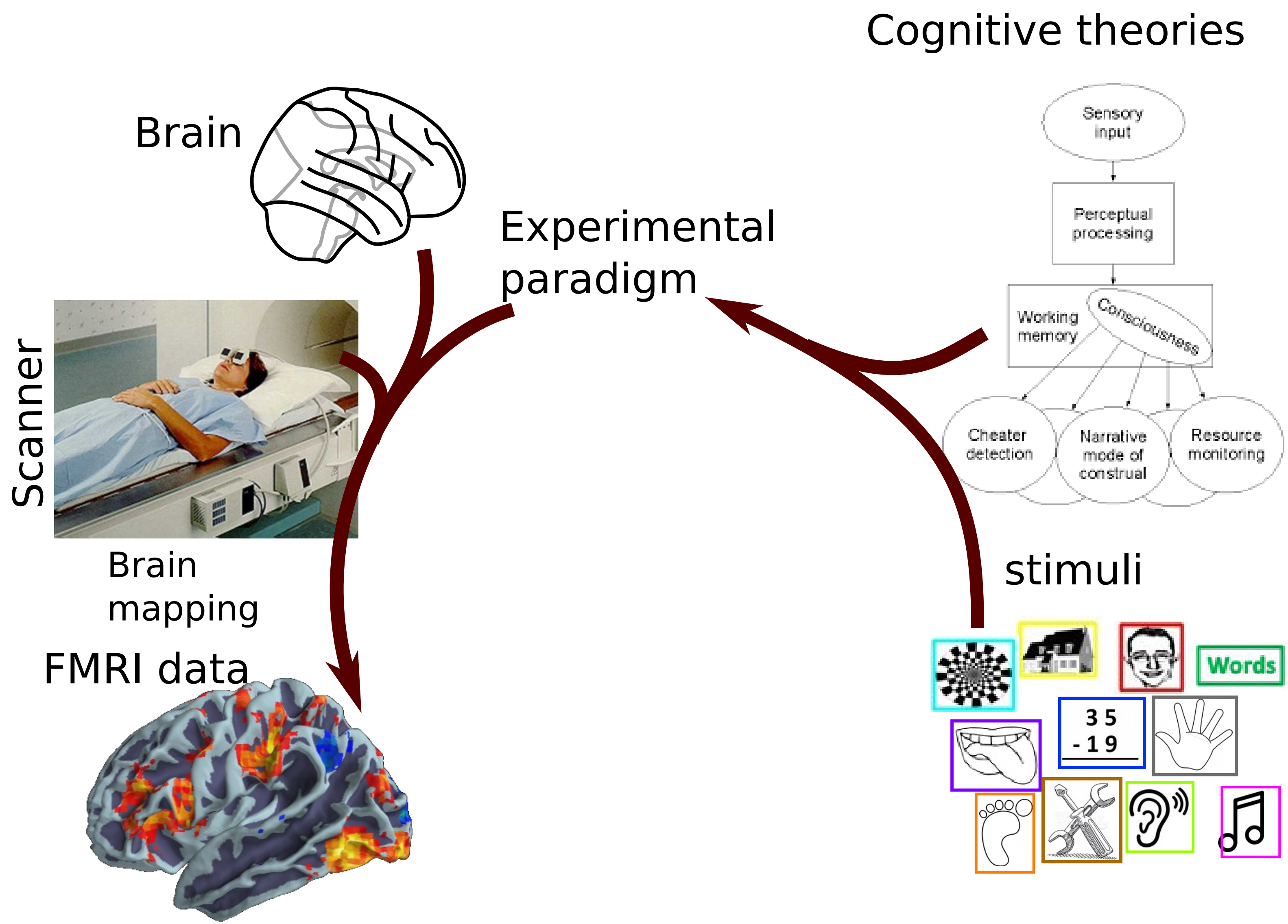
PARIETAL

DE LA RECHERCHE À L'INDUSTRIE

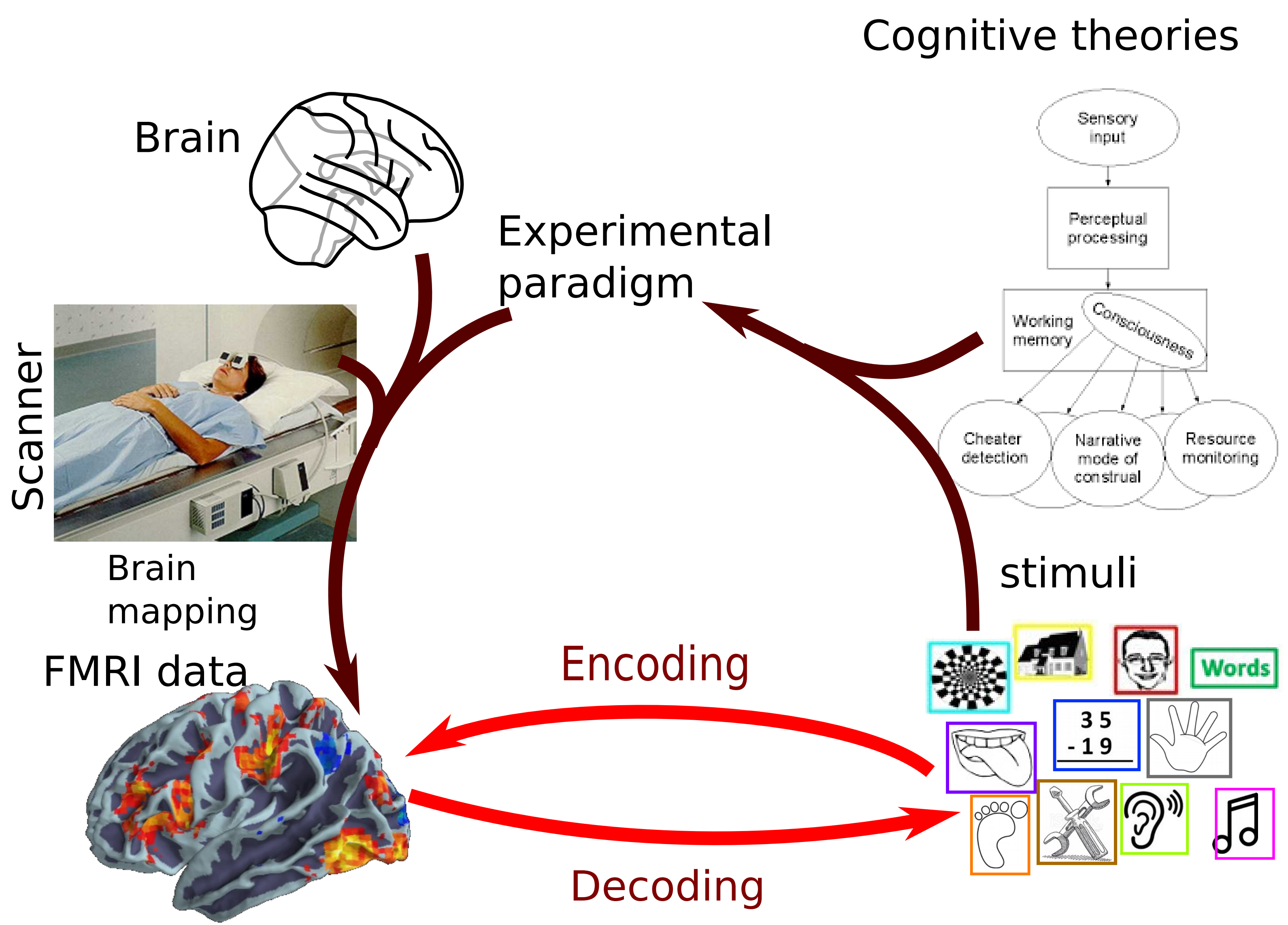
Cognitive neuroscience

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

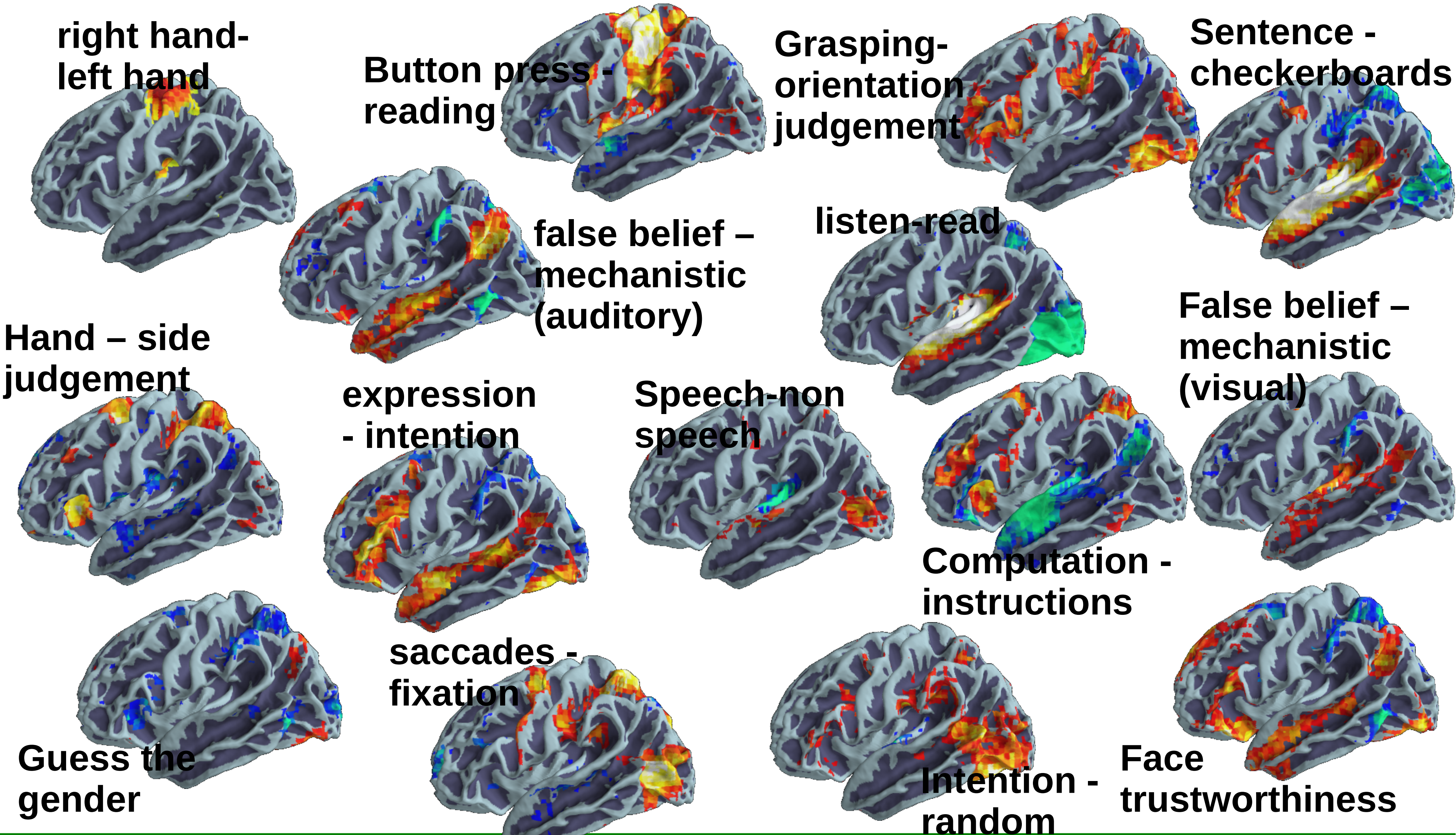
The brain, the mind and the scanner



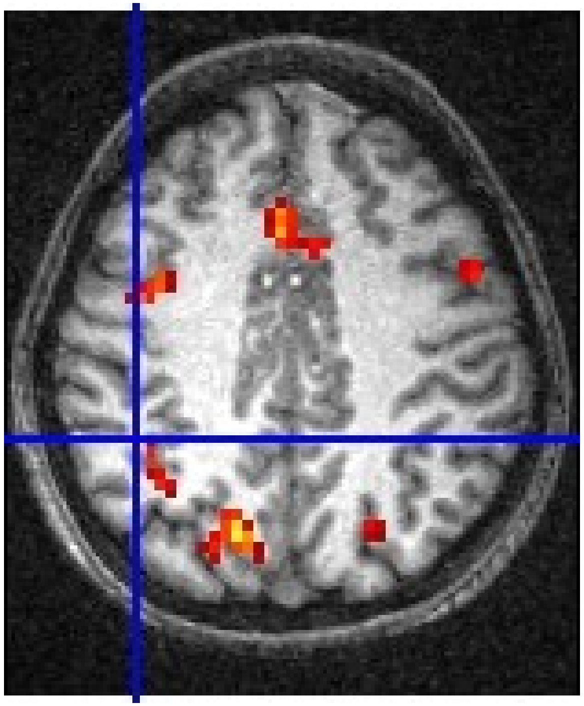
The brain, the mind and the scanner

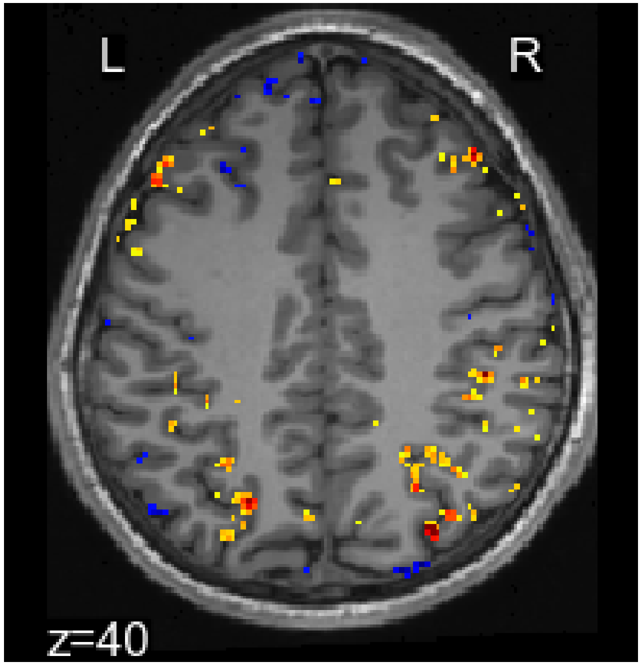


Encoding: mapping cognitive functions to brain activity



Resolution increases





2007:
3 mm

2014:
1.5 mm

2020:
0.5 mm ?

p = 50,000

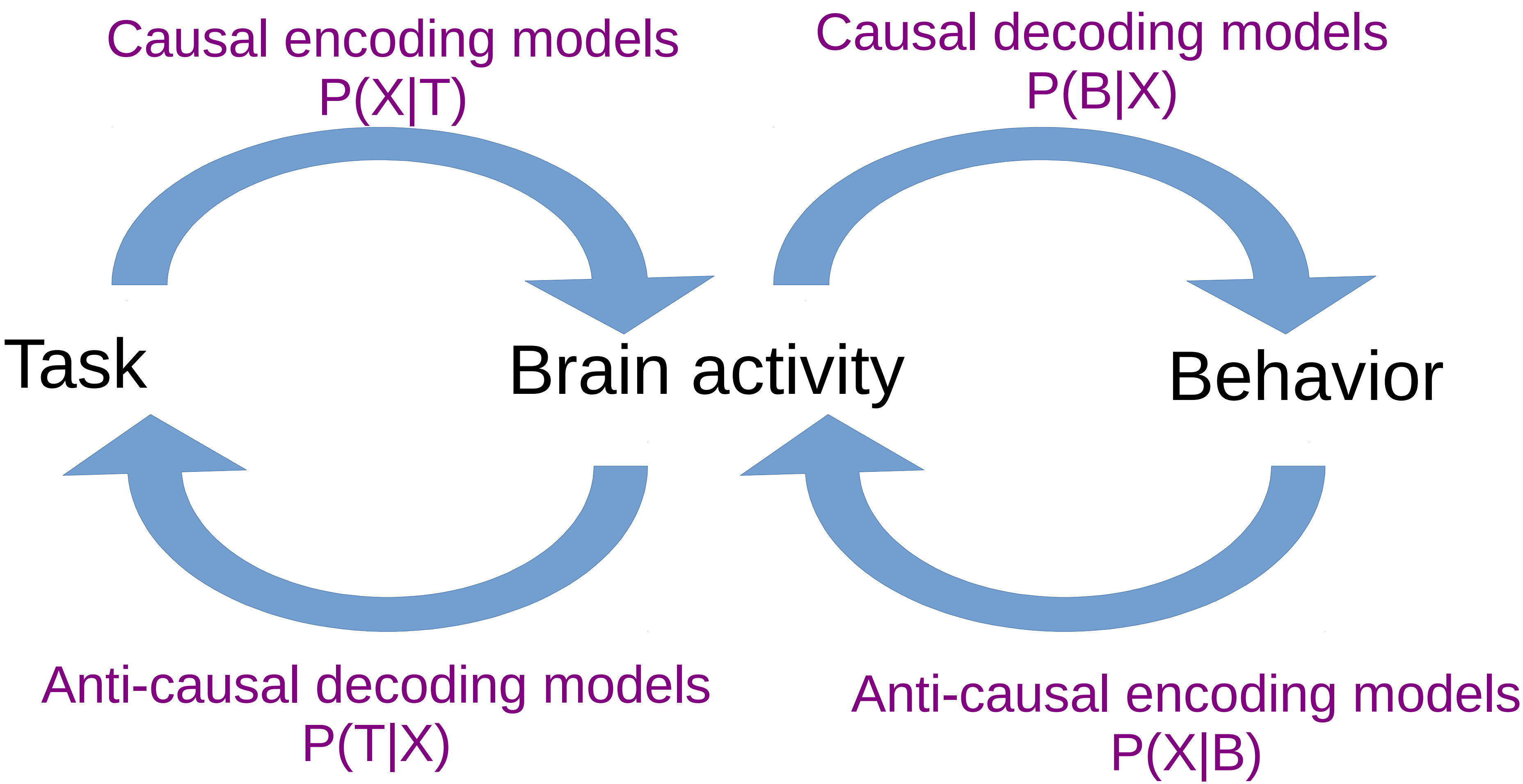
p = 400,000

p = 10^7

better estimators for large-scale brain imaging

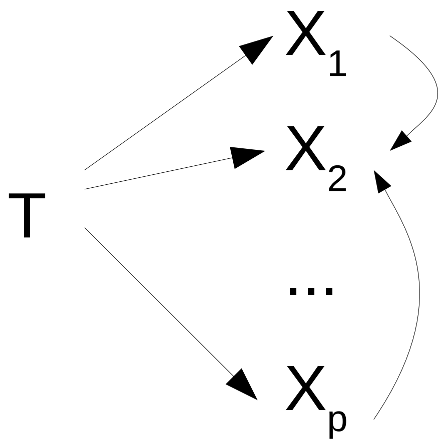
- A causal framework for brain activity decoding
- Dimension reduction for images
- Fast regularized ensembles of models
- Statistical inference for high-dimensional models

Causal reasoning on encoding/decoding

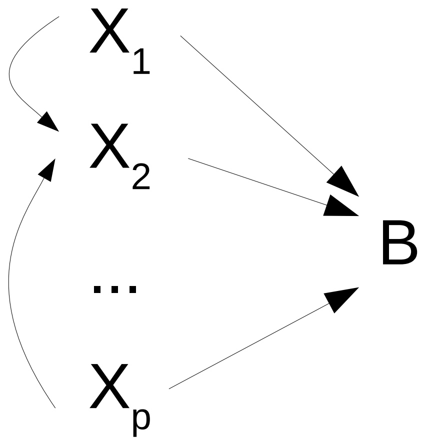


[Weichwald et al Nimg 2015]

Causal interpretation



Encoding: causal
Decoding: anti-causal



Encoding: anti-causal
Decoding: causal

Simple causal models

The Chain

$$X_1 \rightarrow X_2 \rightarrow X_3$$

$$X_1 \not\perp\!\!\!\perp X_3$$

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

Simple causal models

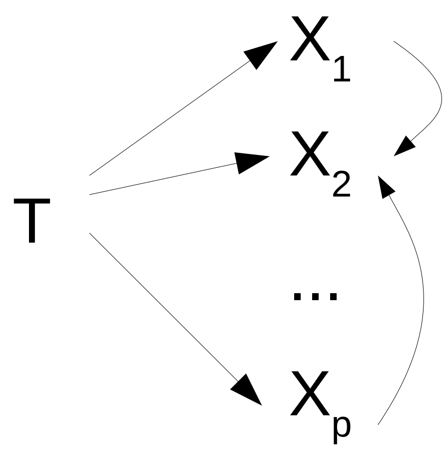
The Chain	The Fork
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$

Simple causal models

The Chain	The Fork	The Collider
$X_1 \rightarrow X_2 \rightarrow X_3$	$X_1 \leftarrow X_2 \rightarrow X_3$	$X_1 \rightarrow X_2 \leftarrow X_3$
$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \not\perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3$
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	$X_1 \not\perp\!\!\!\perp X_3 \mid X_2$

Causal reasoning on encoding/decoding

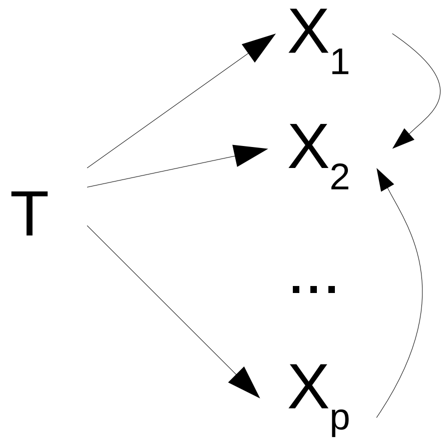
		Feature X_i relevant?		
		Encoding	Decoding	Causal interpretation
Experimental setting	Task	×		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
	Task	$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
	Behaviour			



[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

Experimental setting		Feature X_i relevant?		
		Encoding	Decoding	Causal interpretation
Experimental setting	Task	×		$T \perp X_i \Rightarrow X_i$ is no effect of T
	Task	$\surd$		$T \not\perp X_i \Rightarrow X_i$ is an effect of T
	Task		×	$T \perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive
	Task		$\surd$	$T \not\perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive
	Behaviour			



[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

Experimental setting		Feature X_i relevant? Encoding	Decoding	Causal interpretation
	Task	$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
	Task	$\surd$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
	Task		$\times$	$T \perp\!\!\!\perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive
	Task		$\surd$	$T \not\perp\!\!\!\perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive
	Behaviour	$\times$		$B \perp\!\!\!\perp X_i \Rightarrow X_i$ is no cause of B
	Behaviour	$\surd$		$B \not\perp\!\!\!\perp X_i \Rightarrow$ inconclusive
	Behaviour		$\times$	$B \perp\!\!\!\perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive
	Behaviour		$\surd$	$B \not\perp\!\!\!\perp X_i \mid \boldsymbol{X} \setminus X_i \Rightarrow$ inconclusive

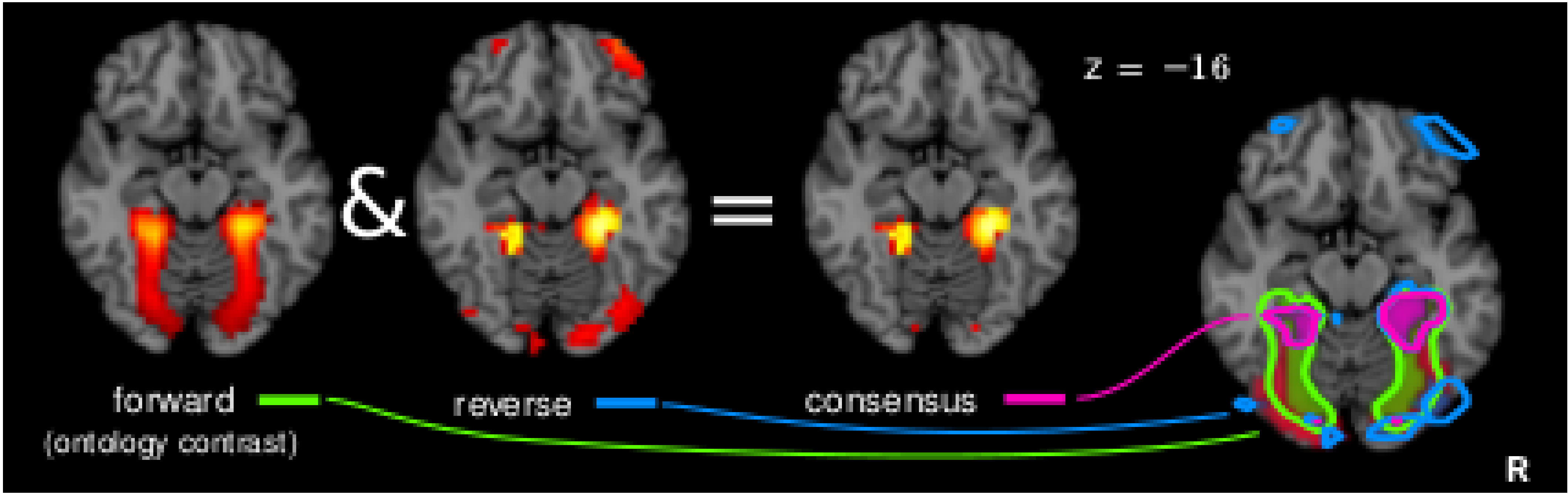
[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

Experimental paradigm		Feature X_i relevant? Encoding	Decoding	Causal interpretation
	Task	$\times$	$\times$	X_i is no effect of T
	Task	$\surd$	$\times$	X_i is an indirect effect of T
	Task	$\times$	$\surd$	X_i provides context
	Task	$\surd$	$\surd$	X_i is an effect of T
	Behaviour	$\times$	$\times$	X_i is no cause of B
	Behaviour	$\surd$	$\times$	X_i is no direct cause of B
	Behaviour	$\times$	$\surd$	X_i provides context
	Behaviour	$\surd$	$\surd$	inconclusive

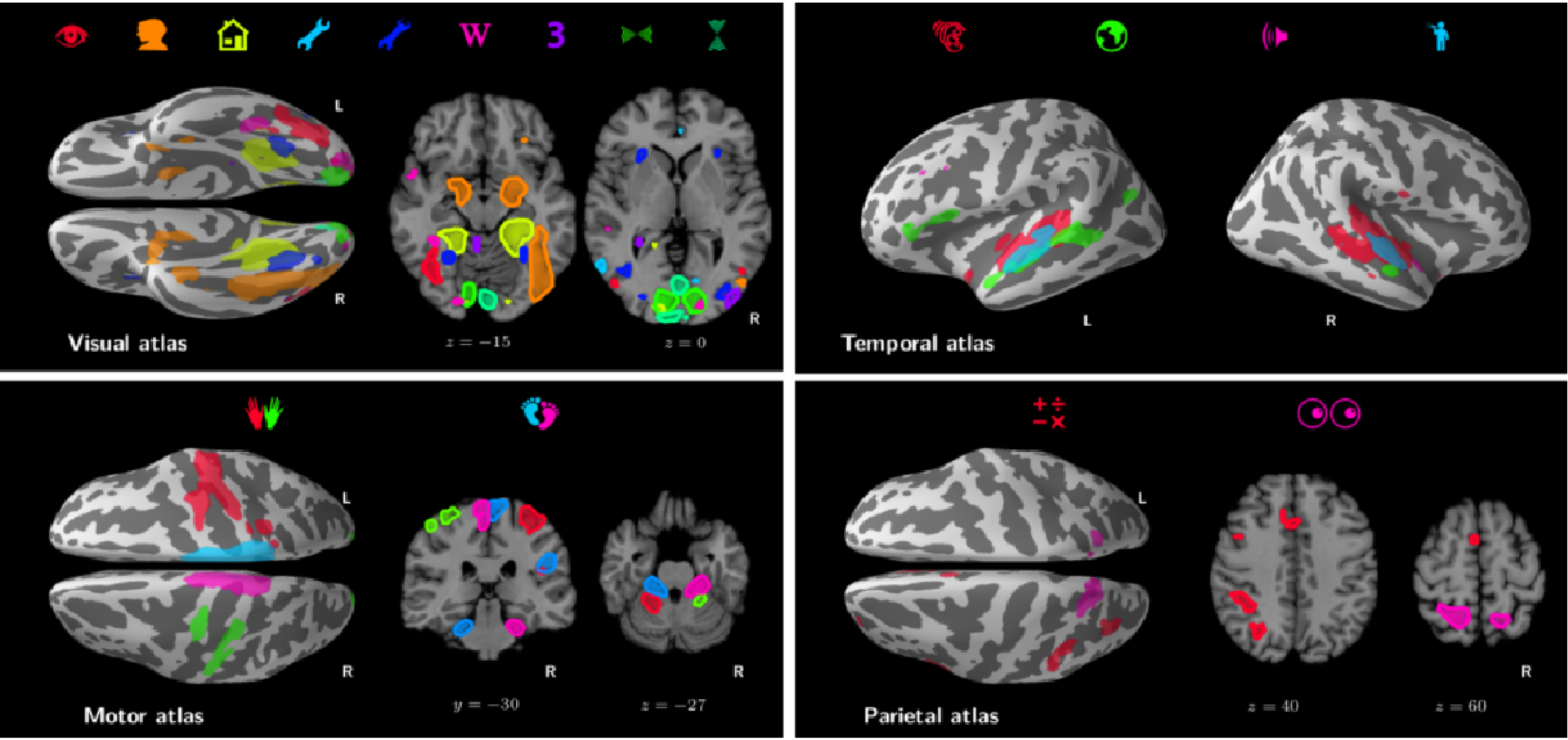
[Weichwald et al. NIMG 2015]

Joint encoding and decoding



[Schwartz et al. NIPS 2013, Varoquaux et al. Submitted to PCB]

Joint encoding and decoding



[Schwartz et al. NIPS 2013, Varoquaux et al. Submitted to PCB]

Statistical associations and causal reasoning

Definition: X_i is a cause of X_j ($X_i \rightarrow X_j$), iff there exist values of X_i and X_j such that $p(x_j|\mathrm{do}\{x_i\}) \neq p(x_j)$.

- Problems:
  - How do you establish $p(x_j|\mathrm{do}\{x_i\}) \neq p(x_j)$ based on finite datasets ?
  - **Large number of conditioning variables**
  - Encoding models: **Multiple comparison issues**
  - Decoding problem: **statistical tests in multiple regression**

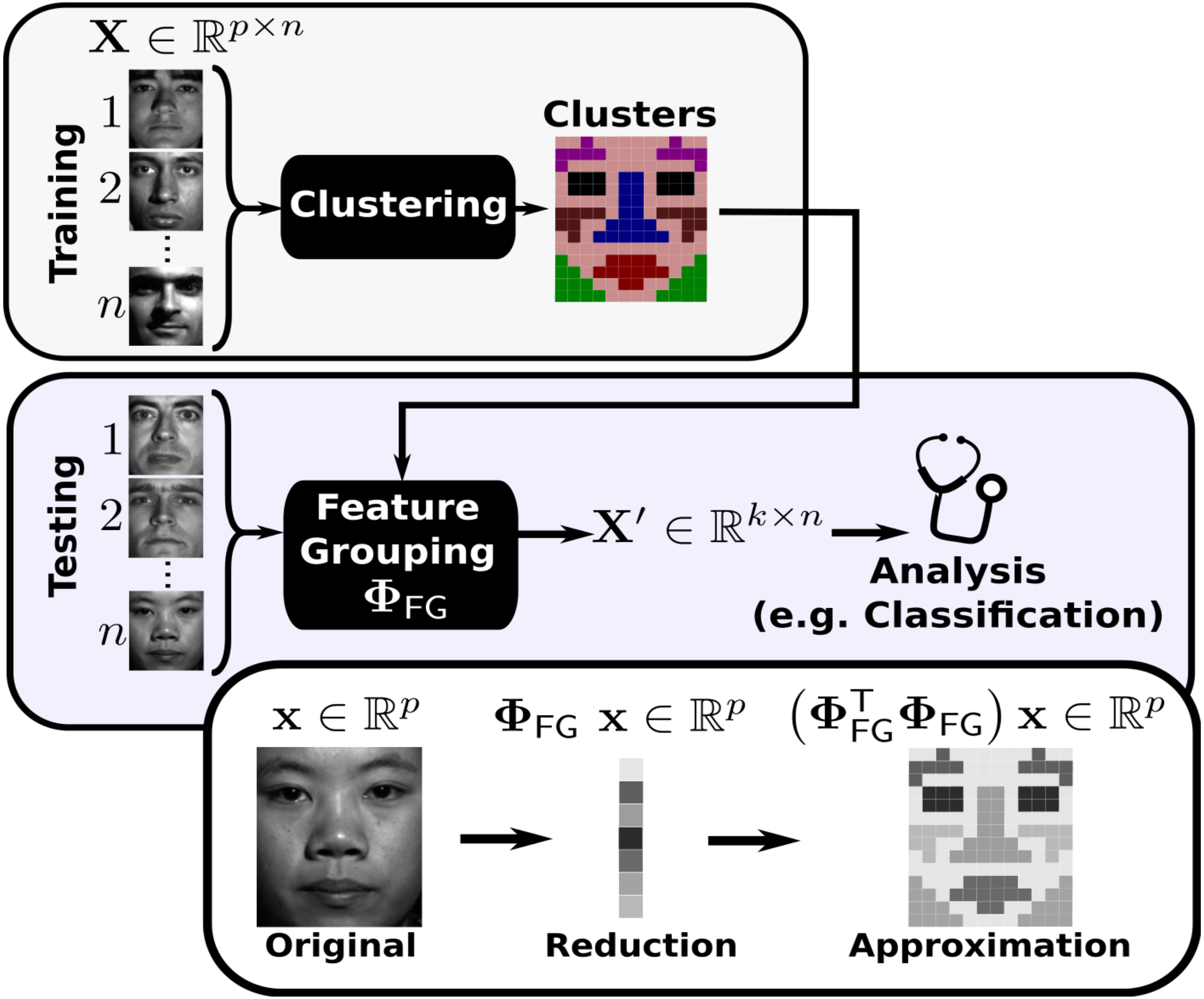
Outline

- A causal framework for brain activity decoding
- Dimension reduction for images
- Fast regularized ensembles of Models
- Statistical inference for high-dimensional models

Compression in the image domain

- Reduce the complexity of learning algorithms: $p \rightarrow k \ll p$
- Random projections = fast generic solution, but
  - Sub-optimal for structured signals
  - Not invertible when p and k are large
- Local redundancy → feature grouping strategies / clustering: “super-pixels”
  - Fast clustering procedures needed (large k regime)

Compression by feature grouping



Crafting good image compression

- Key assumption: signal of interest L-Lipschitz

$$|\mathbf{x}_i - \mathbf{x}_j| \leq L \,\mathrm{dist}_{\mathcal{G}}(v_i, v_j), \quad \forall (i,j) \in [p]^2$$

- Feature grouping matrix $\mathbf{\Phi}_{\mathsf{FG}} \in \mathbb{R}^{k \times p}$
- almost trivially: $\|\mathbf{x}\|^2 - L^2 \sum_{q=1}^{k} |\mathcal{C}_q|^3 \leq \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|^2 \leq \|\mathbf{x}\|^2$
- Worst case $\|\mathbf{x}\|_2^2 - kL^2 \max_{q \in [k]} \{|\mathcal{C}_q|^3\} \leq \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_2^2$

Need a fast method to learn balanced clusters

Denoising properties

- Noisy signal model $\mathbf{x} = \mathbf{s} + \mathbf{n}$

$$\text{MSE}_{\text{approx}} \leq L^2 \sum_{q=1}^{k} |\mathcal{C}_q| \, \text{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 + \frac{k}{p} \text{MSE}_{\text{orig}}$$

- Denoising

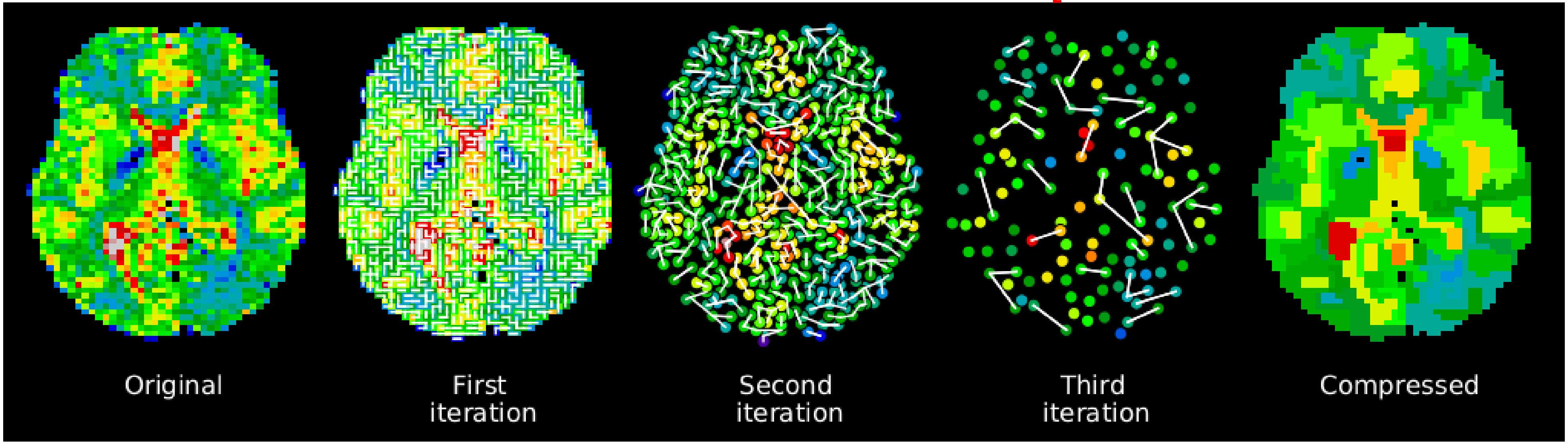
$$\text{MSE}_{\text{approx}} \leq \text{MSE}_{\text{orig}} \qquad L^2 \leq \frac{(p-k)}{\sum_{q=1}^{k} |\mathcal{C}_q| \, \text{diam}_{\mathcal{G}}(\mathcal{C}_q)^2} \sigma^2$$

- Equal-size clusters

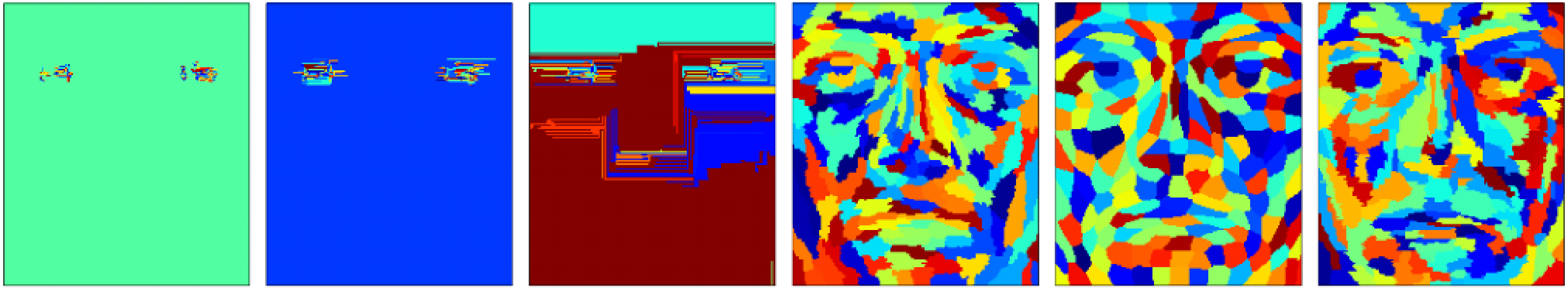
$$\text{MSE}_{\text{approx}} \leq p \left(\frac{L}{k} \right)^2 + \frac{k}{p} \text{MSE}_{\text{orig}} = O\left(\max \left\{ \frac{p}{k^2}, \frac{k}{p} \right\} \right)$$

Recursive neighbor Agglomeration

[Thirion et al. Stamlins 2015]

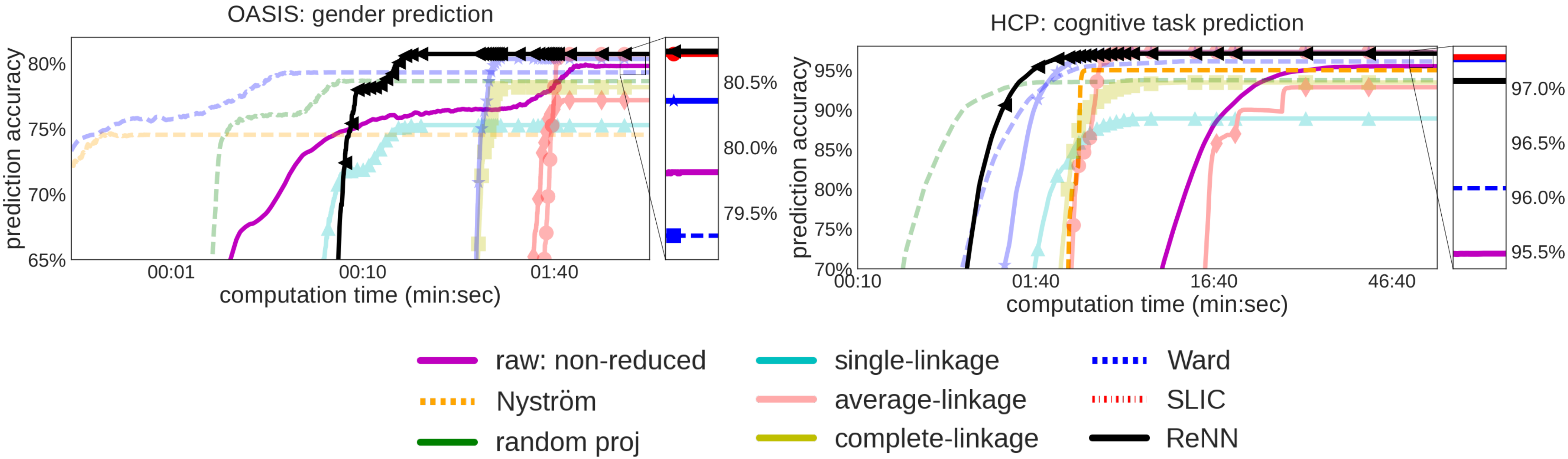


Based on local decisions = fast (linear time) – avoid percolation



(a) single-linkage (b) average-linkage (c) complete-linkage (d) Ward (e) SLIC (f) ReNN

Effect on data analysis tasks

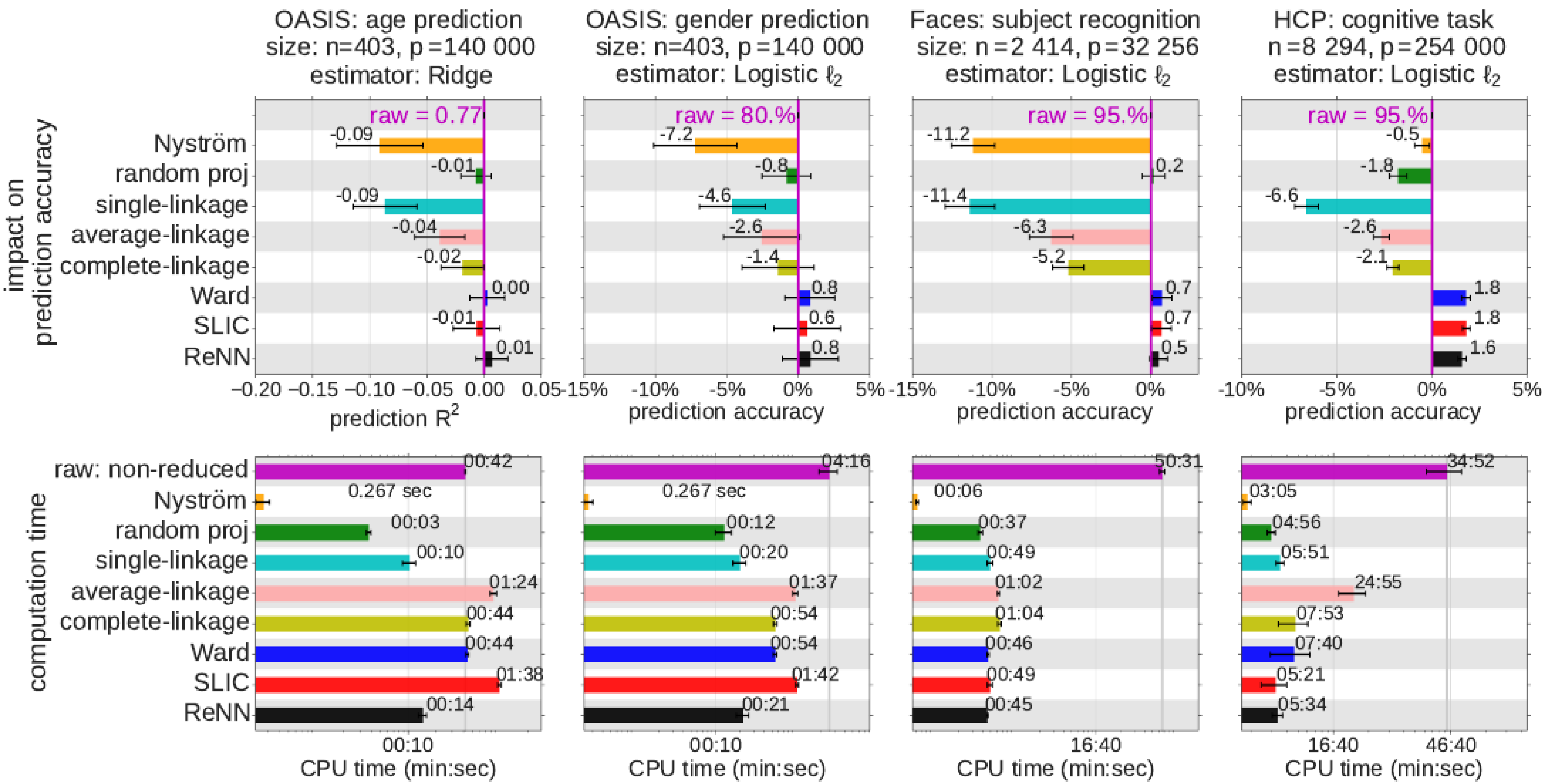


Impressive speed-up and increased accuracy with respect to non-compressed representation

– Clustering has a denoising effect

[Hoyos Idrobo IEEE PAMI in Press]

More results

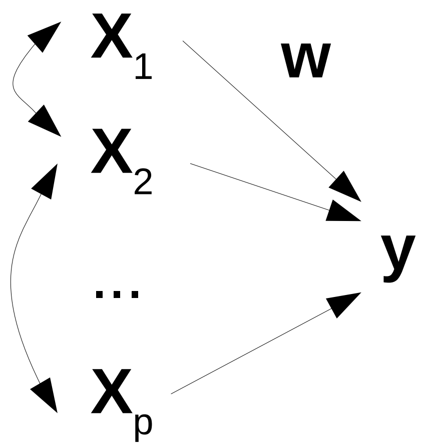


[Hoyos Idrobo IEEE PAMI in Press]

Outline

- A causal framework for brain activity decoding
- Dimension reduction for images
- Fast regularized ensembles of Models
- Statistical inference for high-dimensional models

Brain activity decoding



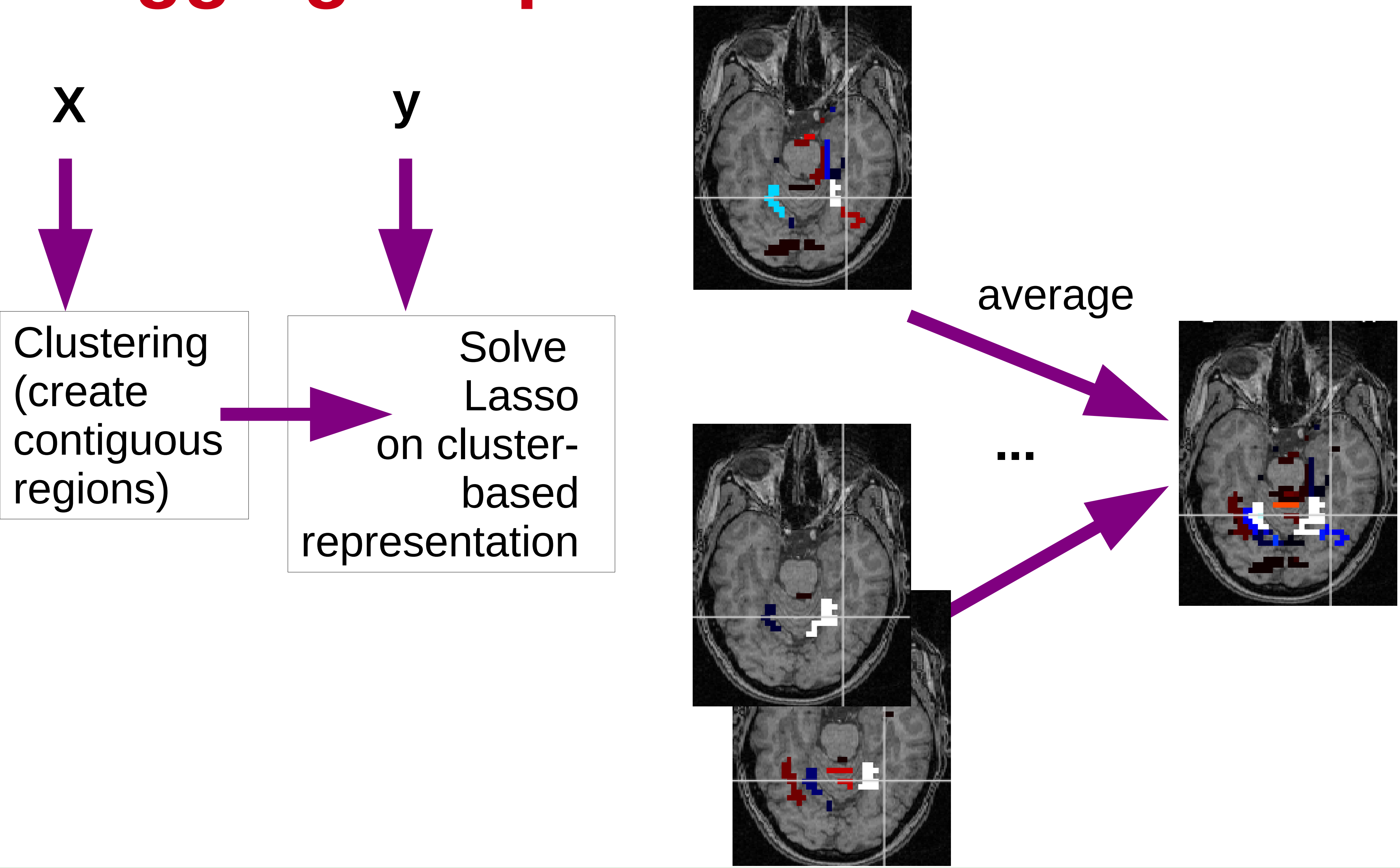
- behavior = f (brain activity)

$$\mathbf{y} = \mathbf{X}\boldsymbol{w}^* + \sigma_* \varepsilon$$

- error vector: $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
- noise magnitude: $\sigma_* > 0$

- prediction: find $\hat{\boldsymbol{w}}$ that minimizes $\|\mathbf{X}\hat{\boldsymbol{w}} - \mathbf{X}\boldsymbol{w}^*\|_2$
- estimation: find $\hat{\boldsymbol{w}}$ with control on $|\hat{w}_j - w_j^*|$ for all $j \in [p]$

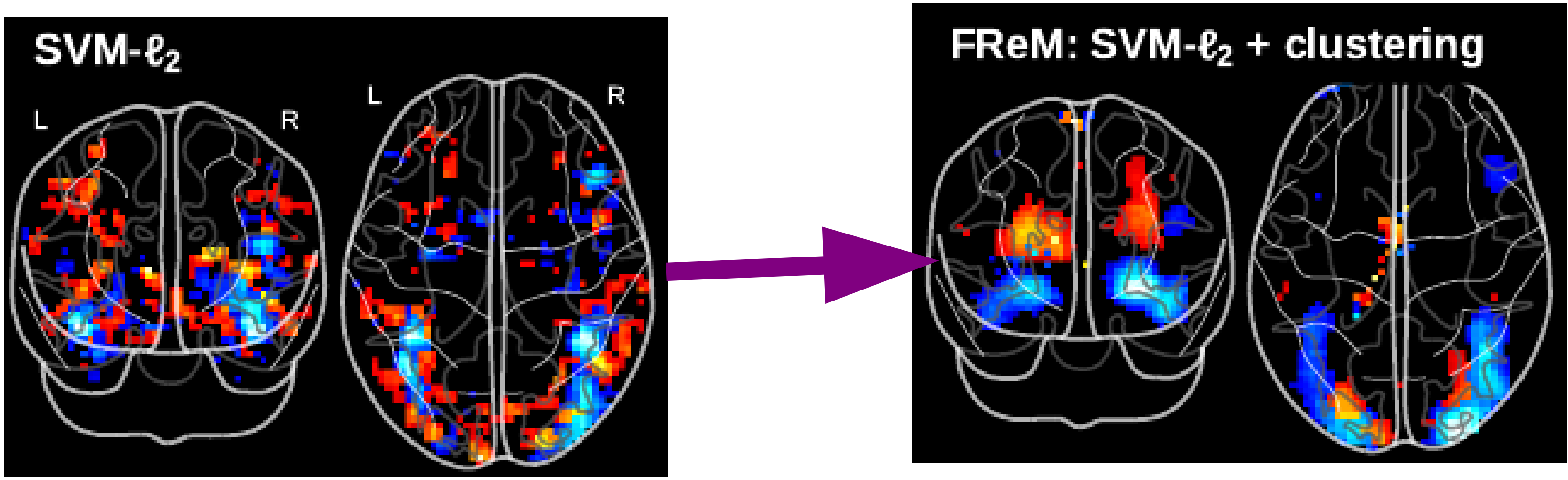
Bagging of sparse clustered models



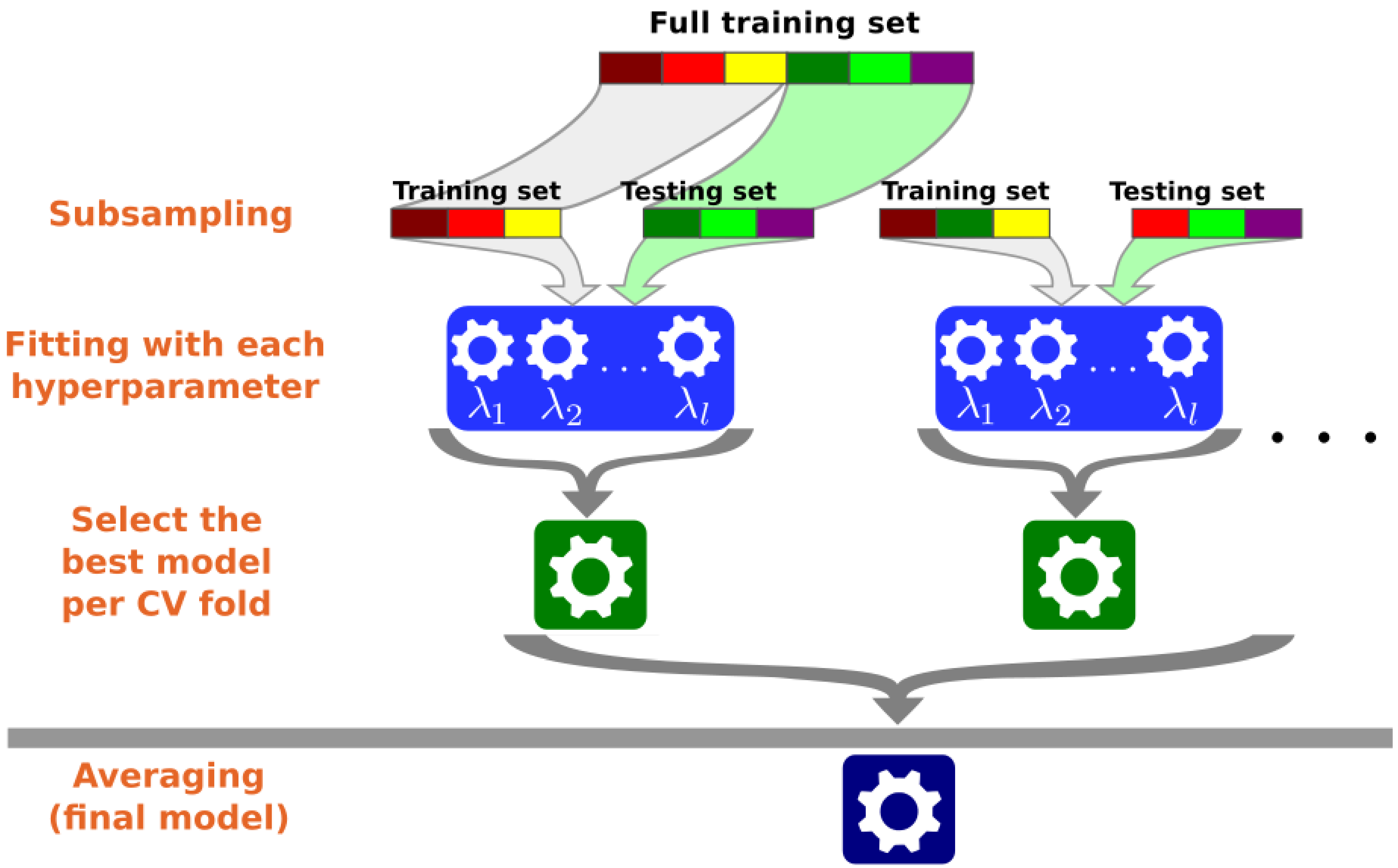
Computationally efficient structure

"fast regularized ensembles of models"

State of the art solution: not very stable, but cheap

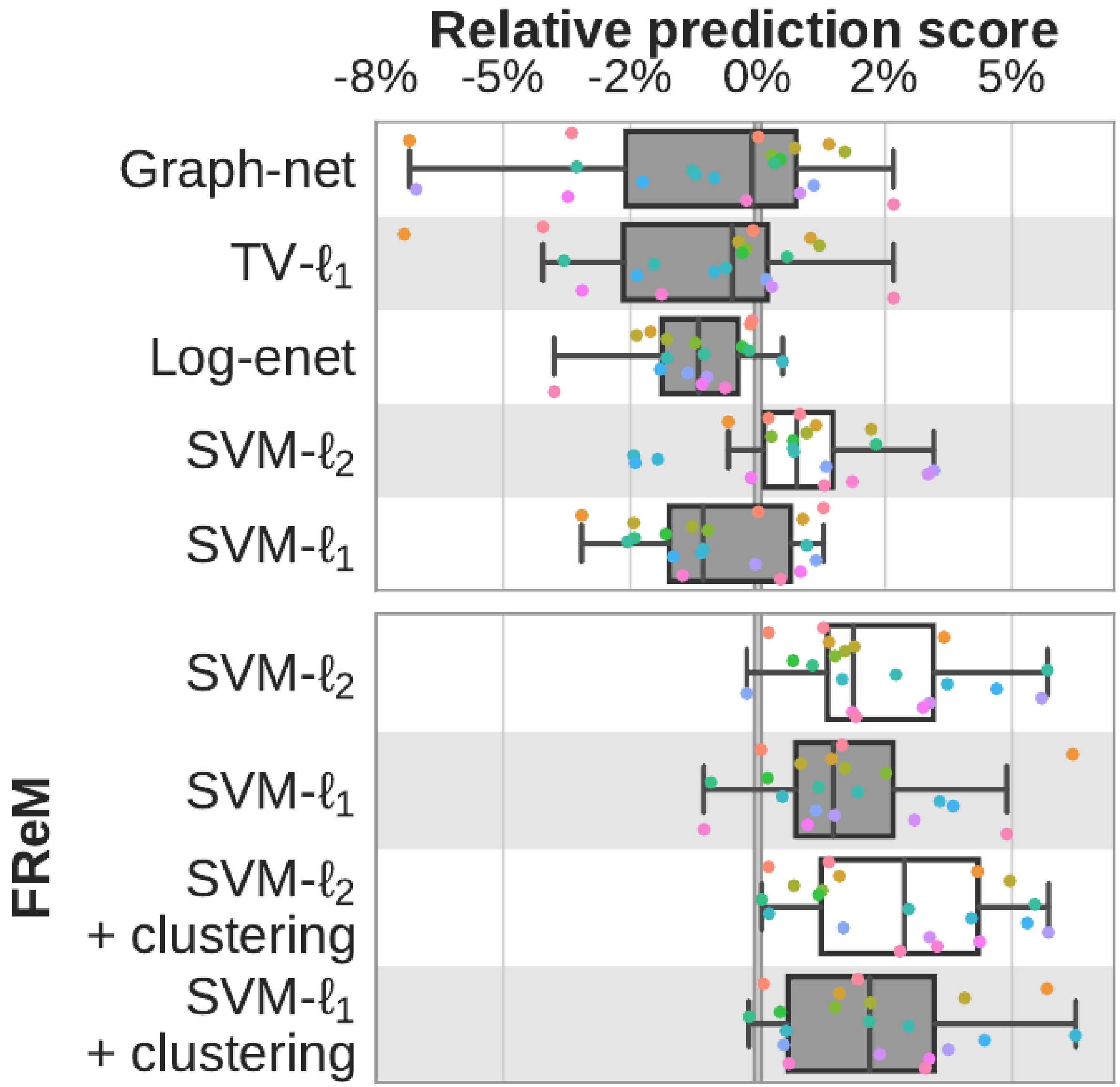


Computationally efficient structure



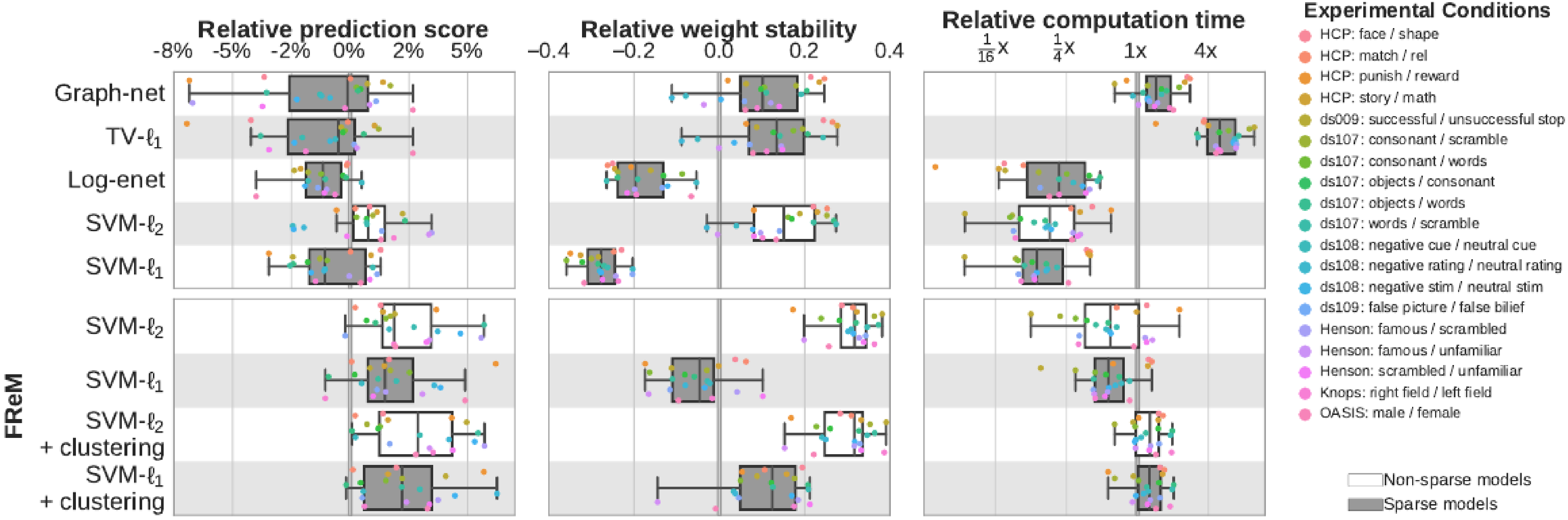
Effect on prediction accuracy

[Hoyos Idrobo et al PRNI 2015, Neuroimage 2017, PAMI 2018]



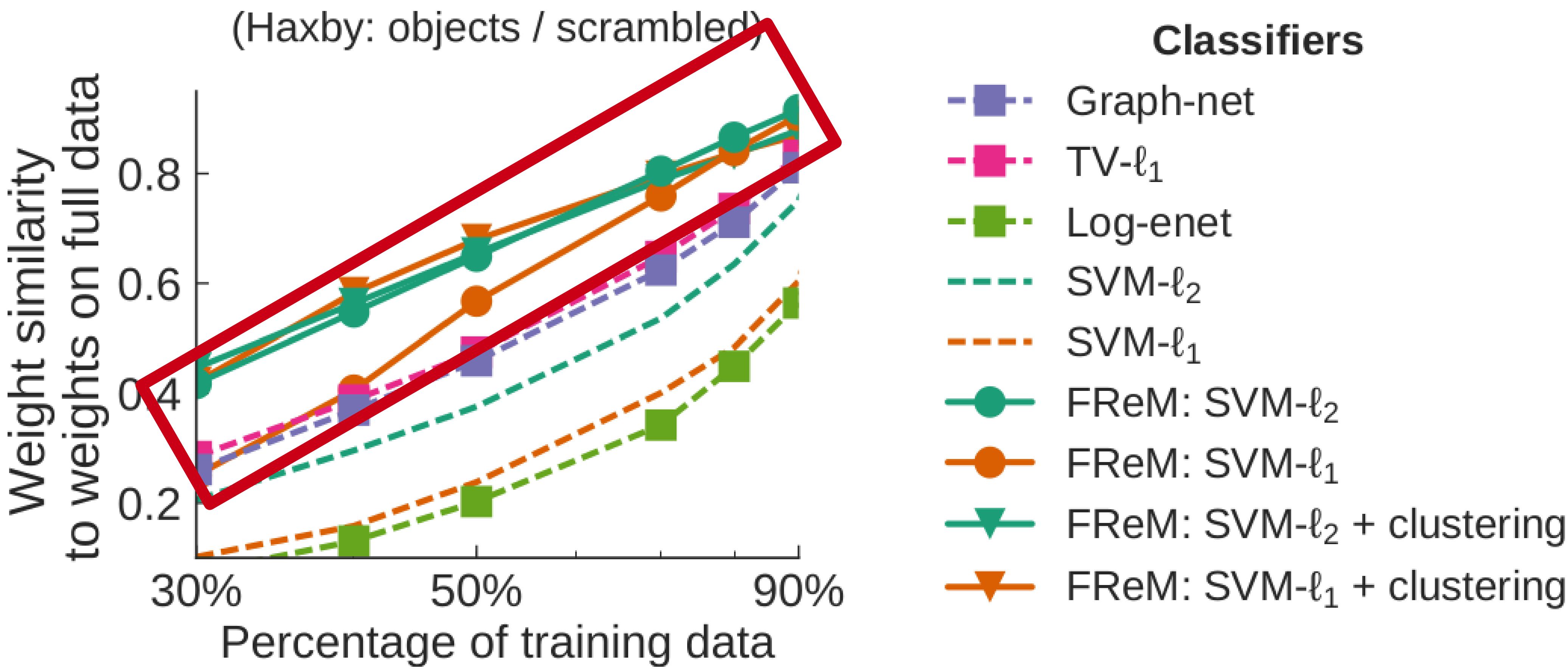
“fast regularized ensembles of models”

More results



[Hoyos Idrobo et al PRNI 2015, Neuroimage 2017, PAMI in Press]

Learning curve



[Hoyos Idrobo et al PRNI 2015, Neuroimage 2017]

Outline

- A causal framework for brain activity decoding
- Dimension reduction for images
- Fast regularized ensembles of Models
- Statistical inference for high-dimensional models

Statistical inference on w

- Inference: find {j: $\mathbf{w}_j > 0$} with some statistical guarantees
- Standard solutions for high-dimensional linear models ($p > n$)
  - Corrected ridge [Bühlmann 2013]
  - Desparsified Lasso [Zhang & Zhang 2014, Montanari 2014]
  - Multi-split [Meinshausen 2009], knockoffs [Candès 2015+]
- Fail for $p \gg n$

Desparsified Lasso

- **Objective:** construct confidence bounds on the coefficients of $\boldsymbol{w}^*$
- **Principle:** [Zhang & Zhang 2014 Series B Stat Meth]
  - construct an unbiased estimator of $\boldsymbol{w}^*$ (generalization of $\hat{\boldsymbol{w}}^{\mathrm{OLS}}$)
  - compute its covariance matrix
- **Heuristic argument:** in low dimension we can prove that:

$$\hat{w}_j^{\mathrm{OLS}} = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{x}_j} \quad ,$$

where $\mathbf{z}_j$ is the residual of the OLS regression of $\mathbf{x}_j$ versus $\mathbf{X}^{(-j)}$:

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{P}_{\mathbf{X}^{(-j)}} \mathbf{x}_j \quad ,$$

where $\mathbf{P}_{\mathbf{X}^{(-j)}}$ is the projection onto $\mathrm{Span}(\mathbf{X}^{(-j)}) \subset \mathbb{R}^{p-1}$

Desparsified Lasso

- **Desparsified Lasso estimator:** when $n < p$, $\mathbf{z}_j$ is the residual of a Lasso-CV regression of $\mathbf{x}_j$ vs $\mathbf{X}^{(-j)}$ and the debiased estimator is:

$$\hat{w}_j = \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{x}_j} - \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k \hat{w}_k^{(init)}}{\mathbf{z}_j^\top \mathbf{x}_j} \quad ,$$

  where $\hat{\boldsymbol{w}}^{(init)}$ is an initial non linear estimator of $\boldsymbol{w}^*$ (*e.g.*, Lasso)

- **Covariance:** the covariance matrix of this estimator is:

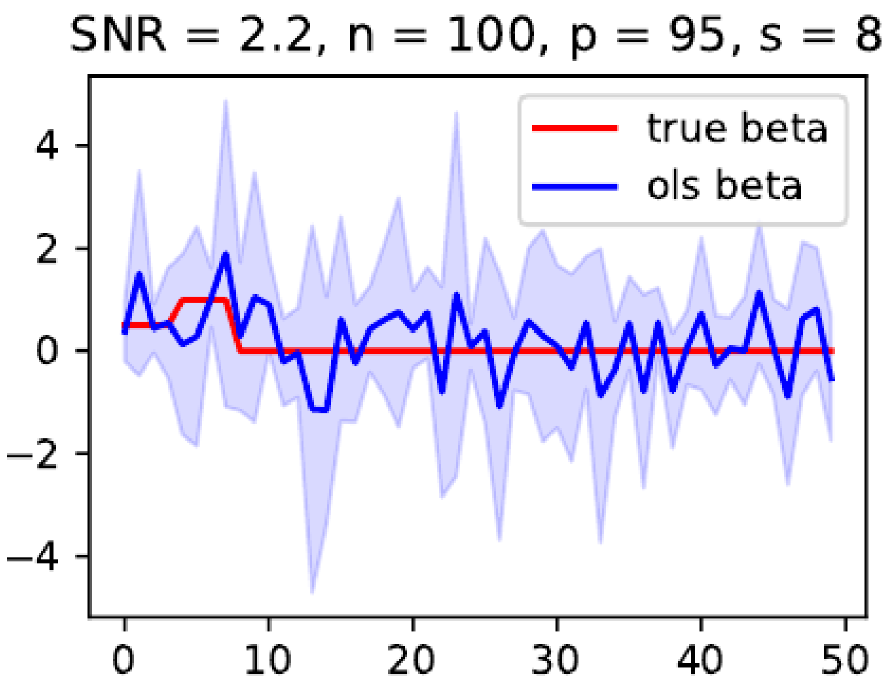
$$\Omega_{jk} = \frac{n \mathbf{z}_j^\top \mathbf{z}_k}{(\mathbf{z}_j^\top \mathbf{x}_j)(\mathbf{z}_k^\top \mathbf{x}_k)}$$

- **Confidence bounds:** under few assumptions (Dezeure et al. [2015]):

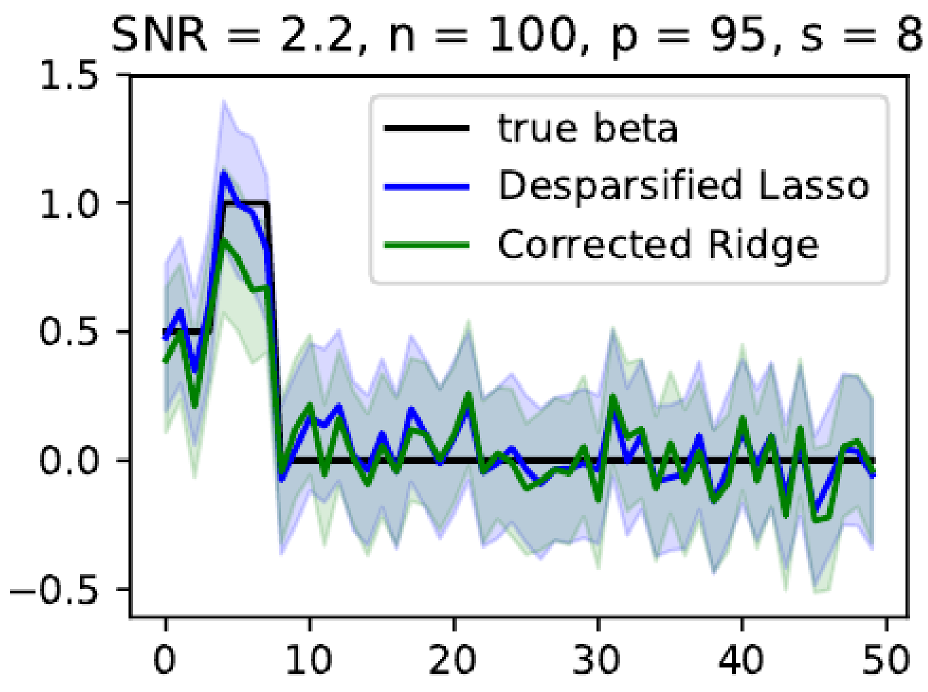
$$\sigma_*^{-1} (\Omega_{jj})^{-1/2} (\hat{w}_j - w_j^*) \sim \mathcal{N}(0, 1)$$

Preliminary assessment

- **Low dimension:** $n = 100$ and $p = 95$
- **OLS versus corrected Ridge and desparsified Lasso:**

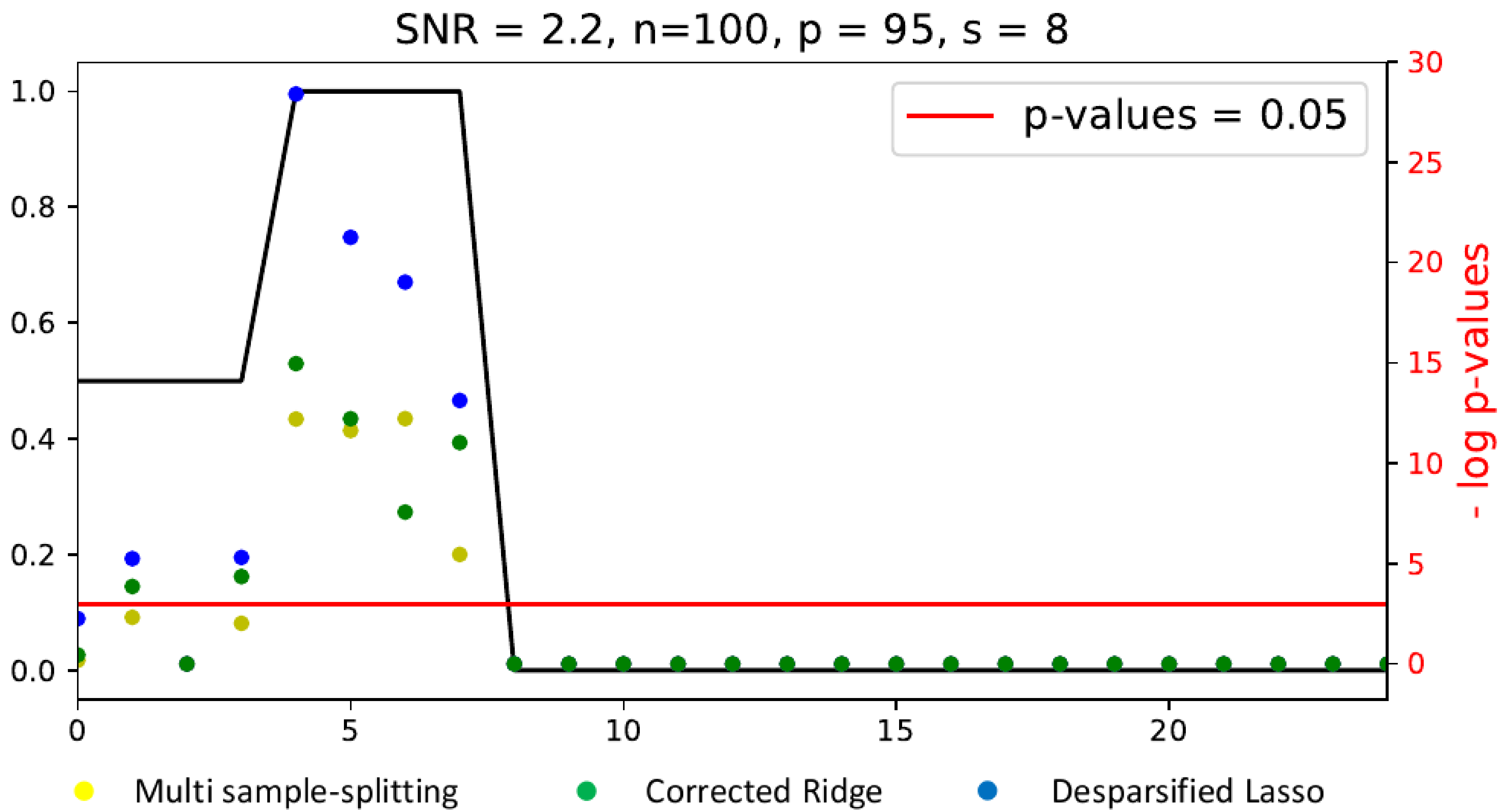


OLS regression when $p \approx n$



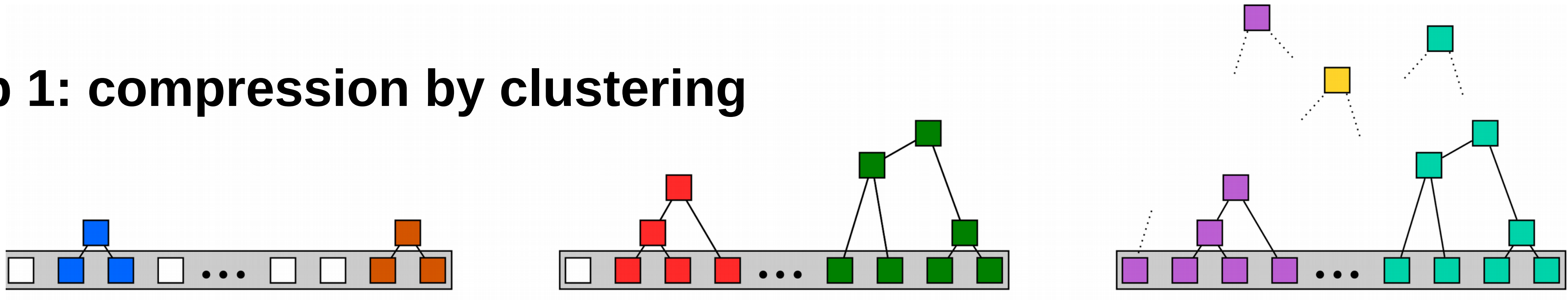
Corrected Ridge and **Desparsified Lasso** when $p \approx n$

Preliminary assessment



Adaptation to brain imaging

Step 1: compression by clustering



Step 2: inference on compressed representations

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

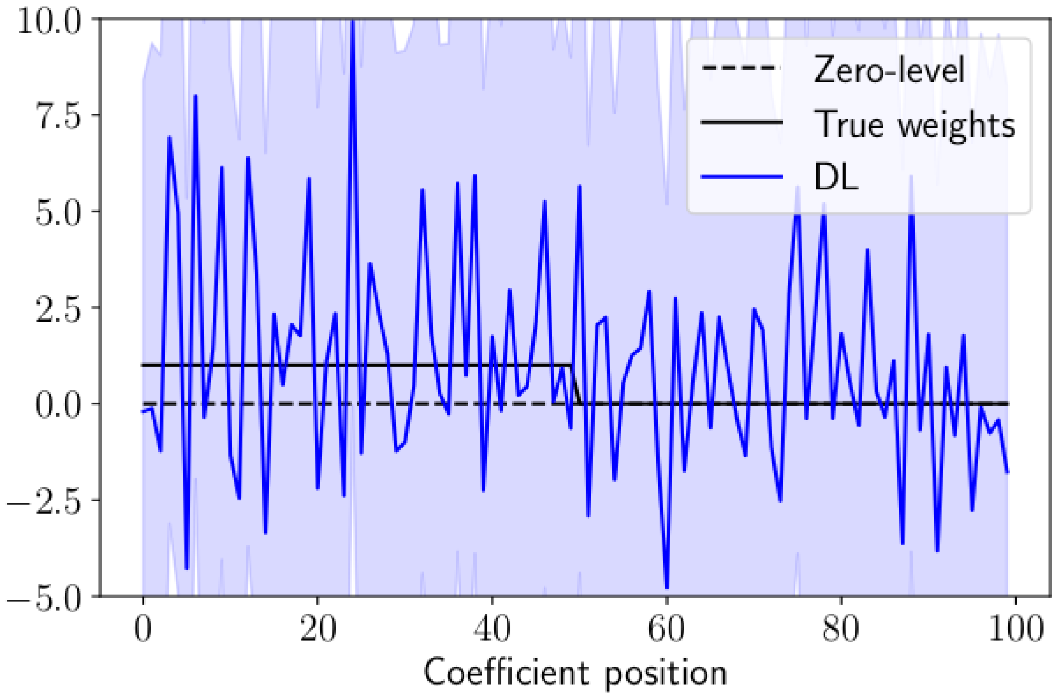
***C**lustered **D**esparsified **L**asso*

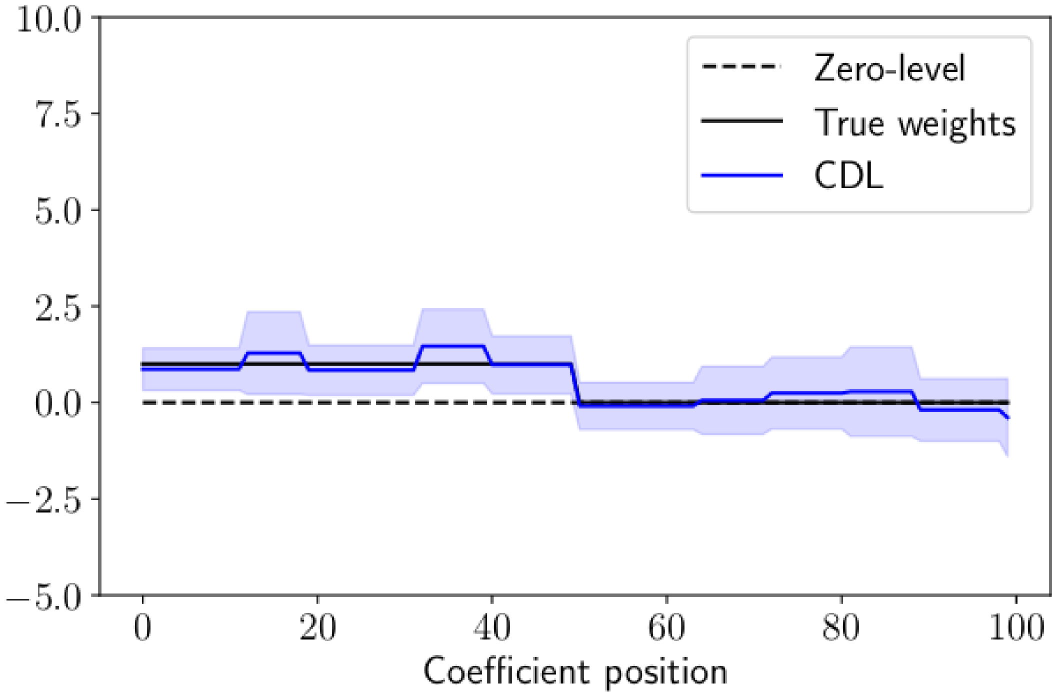
Step 3: ensembling iterate with different parcellations
→ **aggregate p-values** (FReM-like approach)

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

Large p → need dimension reduction

p=2000, n=100



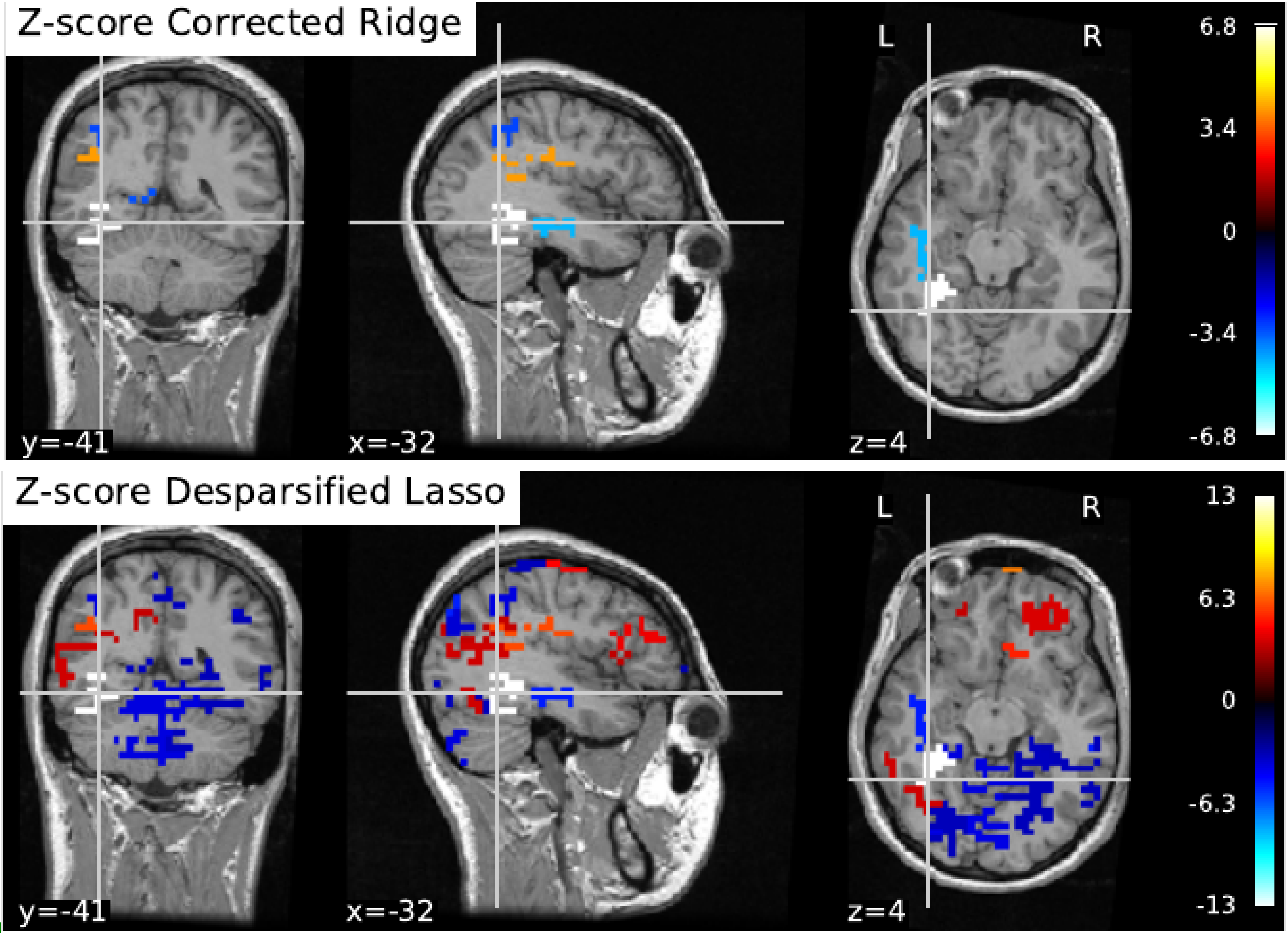


Large p kills statistical power

CDL tames variance

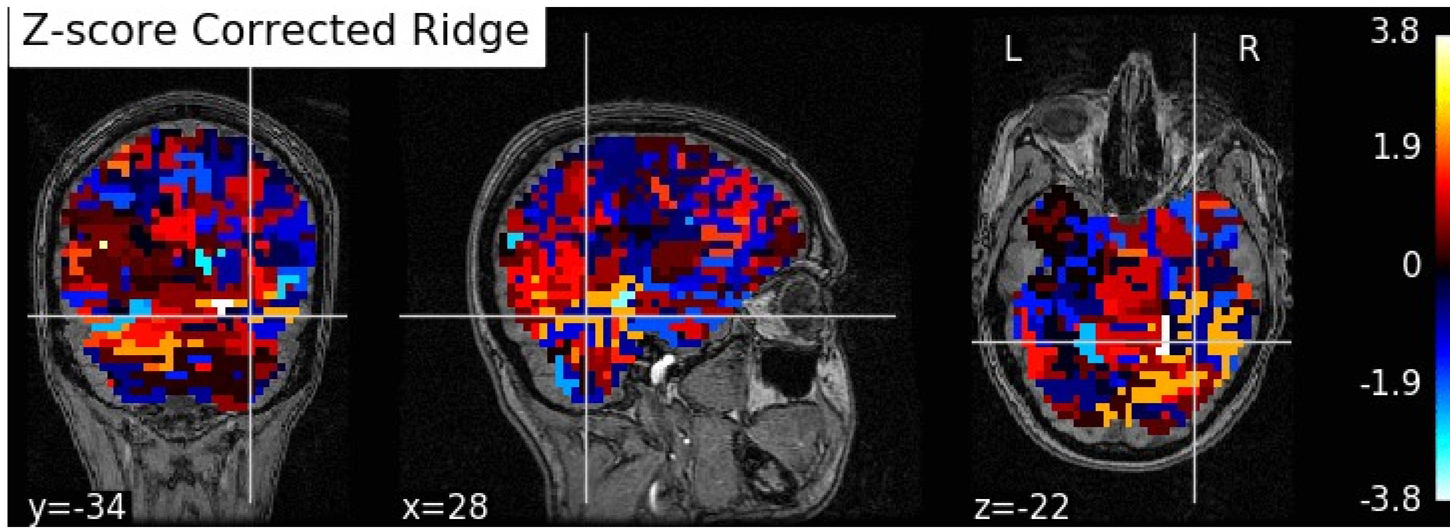
[Chevalier et al. subm. To MICCAI]

Preliminary assessment: CDL

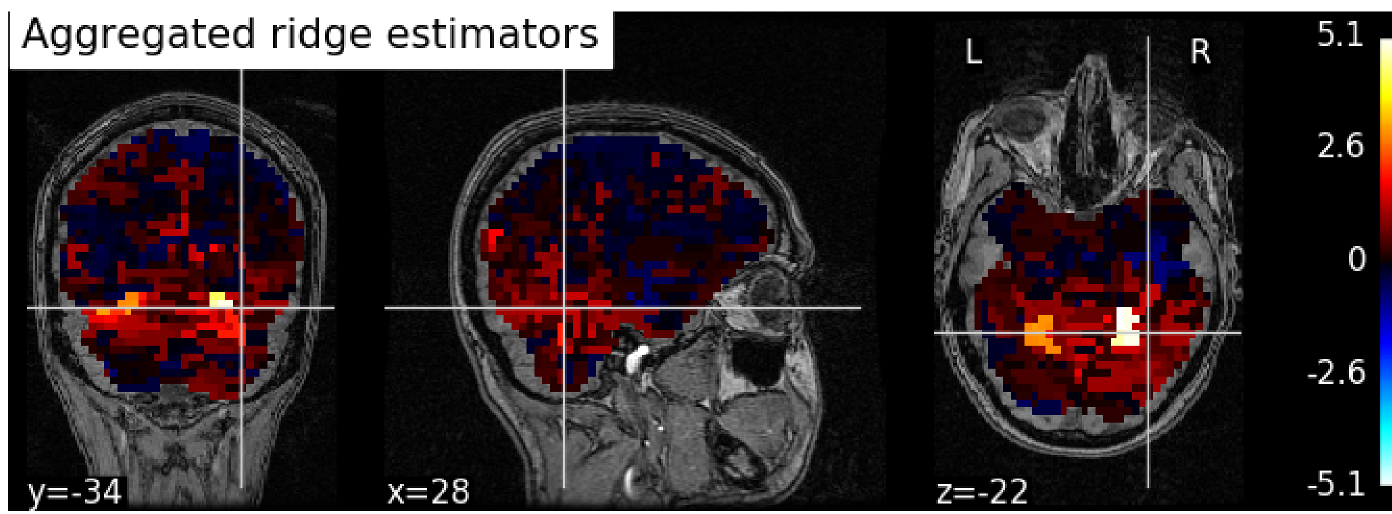


From CDL to ECDL

DL p-values from different clusterings

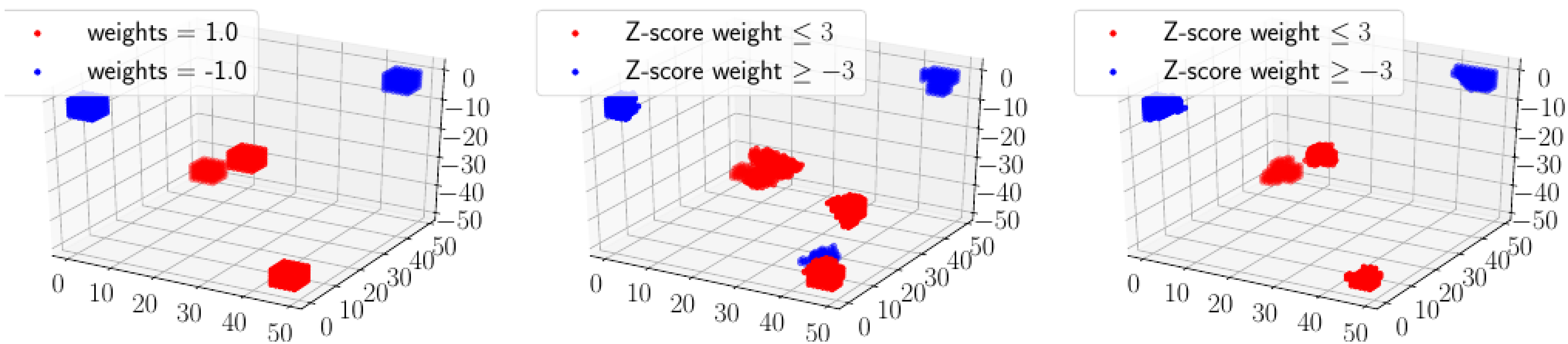


aggregation



Simulations: ECDL > CDL

- **Parameters:** $n = 400$, $H = 50$, $p = H^3 = 125\,000$, $\sigma_{\mathrm{smth}} = 2$
- **Noise:** $\mathrm{SNR}_y = 3$ by taking $\sigma_* = 8$
- **Hyperparameters:** $C = 500$ and $B = 25$
- **Weights:**



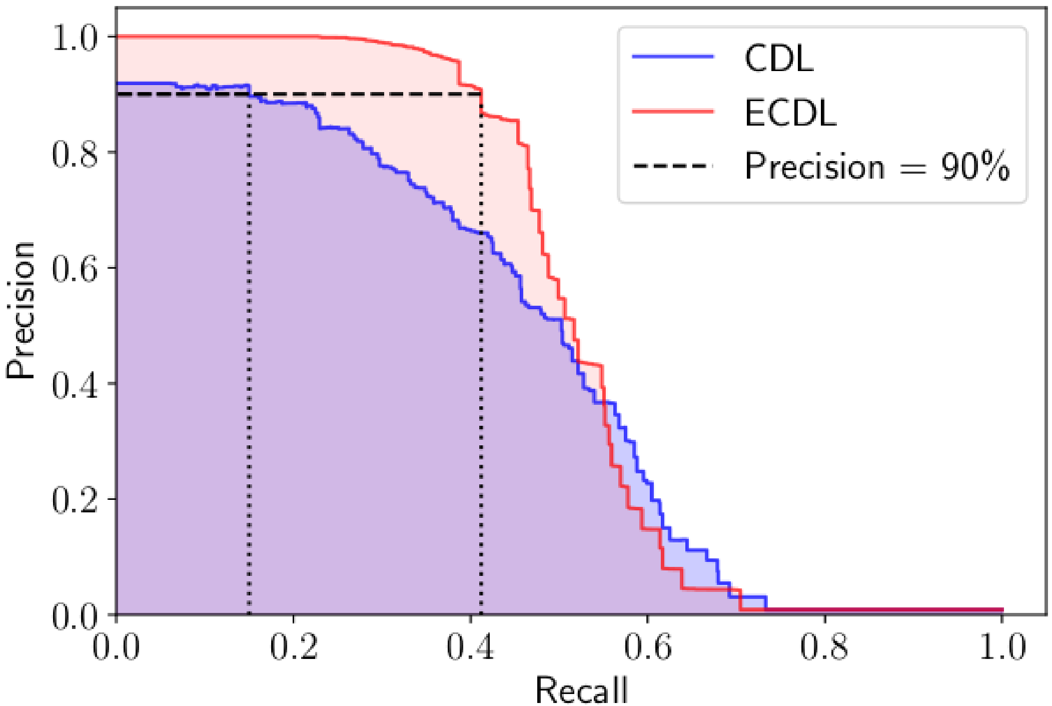
(a) weight vector: $\mathbf{w}^*$ (b) CDL (c) ECDL

[Chevalier et al. subm. To MICCAI]

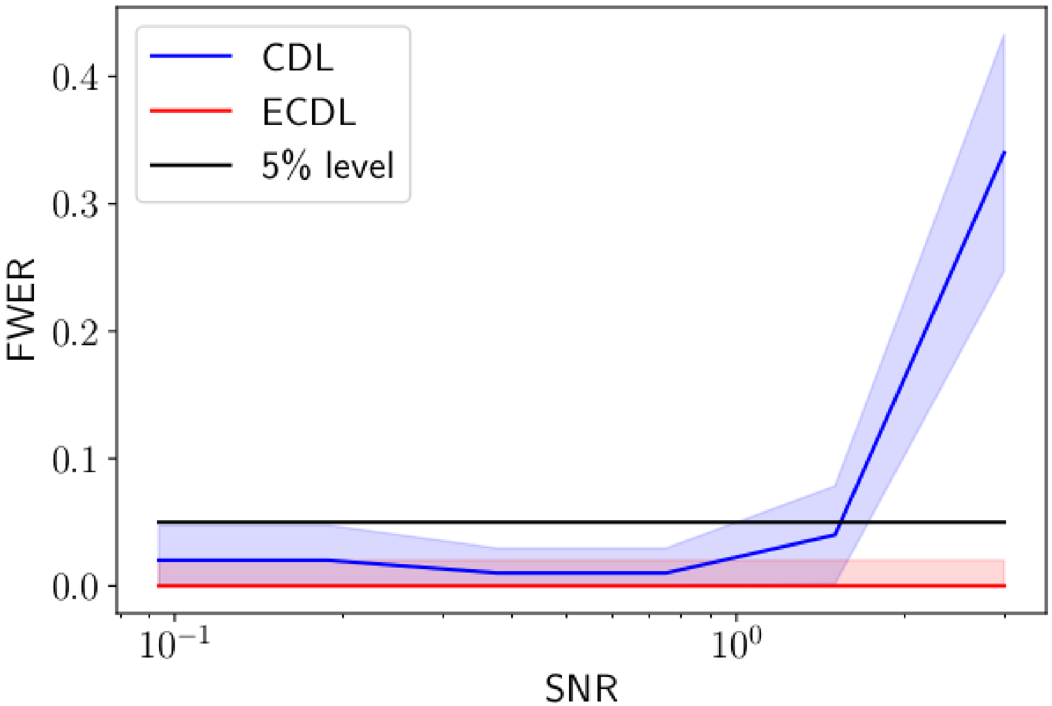
Experiments: PR and FWER control

$$\text{Recall} = \frac{\text{Number of true positive}}{\text{Size of the active set}} \quad \text{Precision} = \frac{\text{Number of true positive}}{\text{Number of discoveries}}$$

$$\text{FWER} = \text{Prob}(\text{Number of false positive} \geq 1)$$



Better PR with ECDL



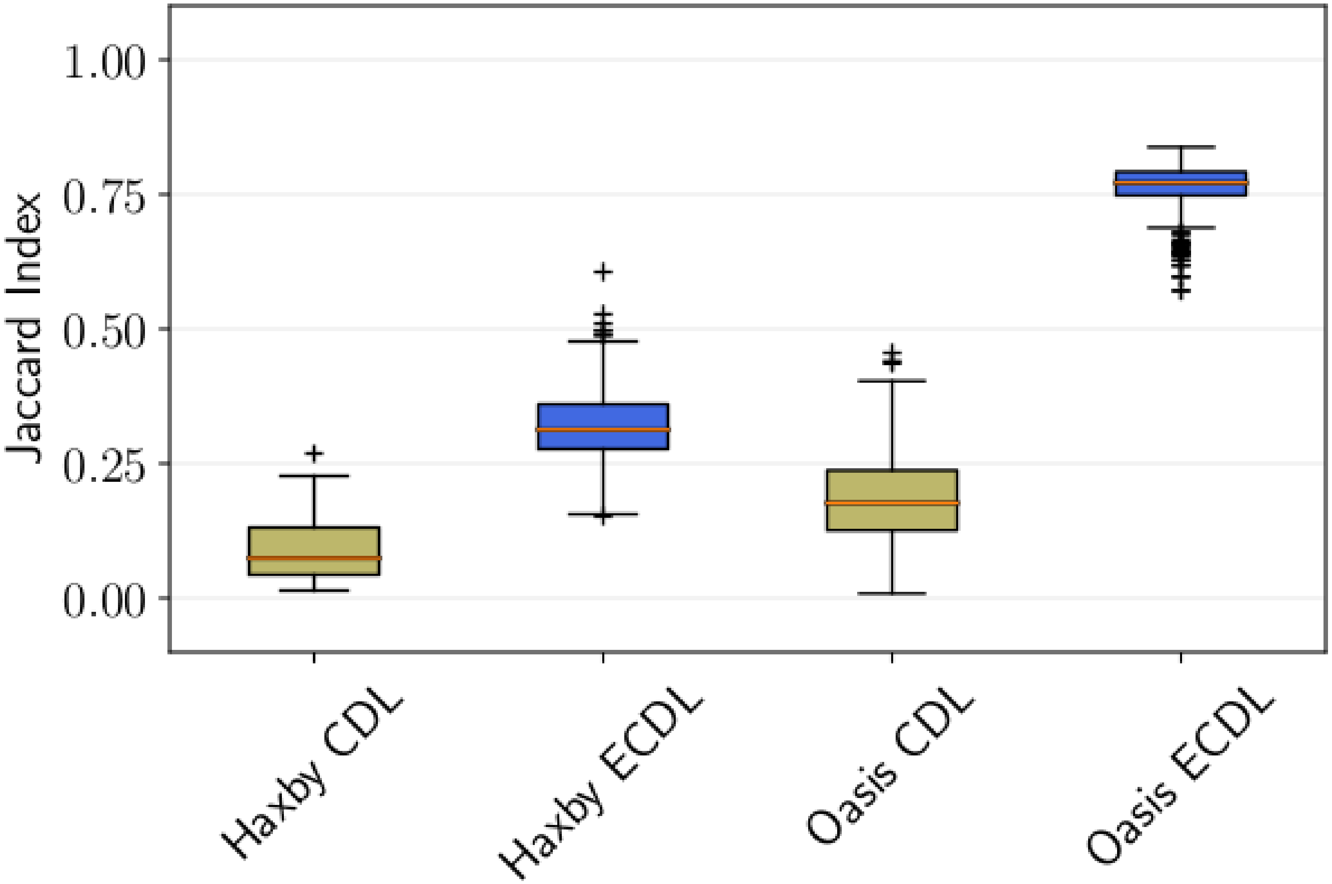
+ More accurate FWER control

[Chevalier et al. subm. To MICCAI]

Stability gains on real data

Similarity across bootstrap replications of the inference

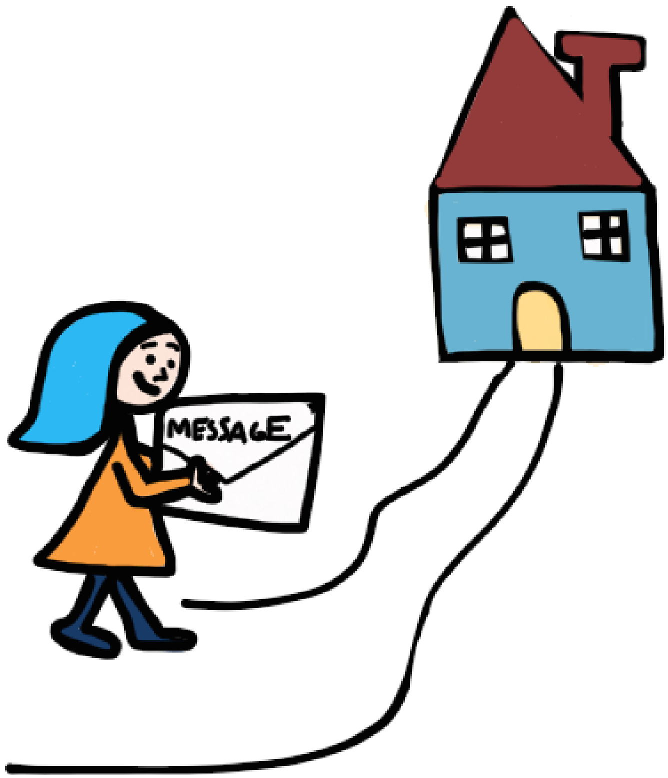
(same result with other metrics)



On two datasets, ECDL improves reproducibility

[Chevalier et al. subm. To MICCAI]

Conclusion



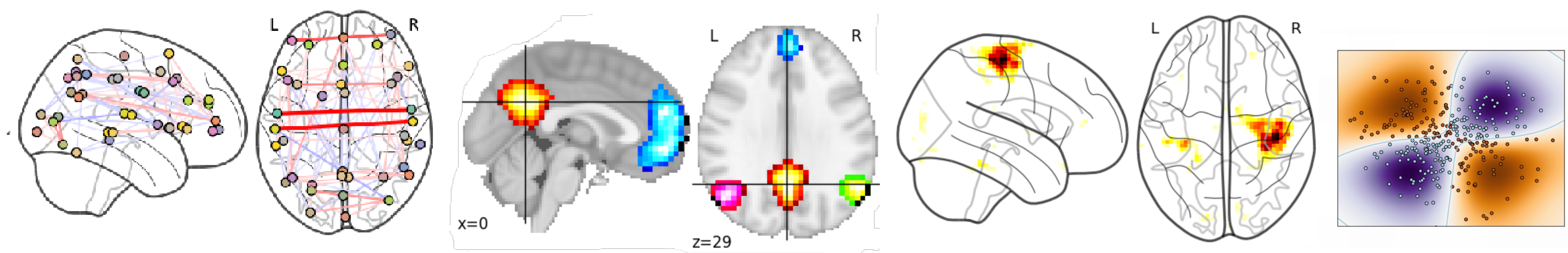
- Large-p data bring challenges:
  - Computation cost
  - Overfit
  - Difficulty of statistical inference
  - … of causal reasoning
- Solutions: online learning, subsampling, compression
- Ensembling improves estimators
- Go & get more data

WIP

- too conservative ?
- Classification ?
- Use of bootstrap
- knockoffs

From good ideas to good practices: software

- Machine learning in Python
- Machine learning for neuroimaging
  http://nilearn.github.io
- BSD, Python, OSS
  - Classification of (neuroimaging) data
  - Network analysis



Acknowledgements

Parietal

G. Varoquaux,
A. Gramfort,
P. Ciuciu,
D. Wassermann,
D. Engemann,
A. Manoel,
D. Chyzhyk
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
C. Petitot

Other collaborators
(thanks for the data)
S. Dehaene
R. Poldrack,
J. Haxby
C. F. Gorgolevski
J. Salmon

Human Brain Project

université PARIS-SACLAY

AGENCE NATIONALE DE LA RECHERCHE ANR