Statistical inference in high-dimension & application to brain imaging

Imaging and machine learning workshop

Bertrand Thirion,

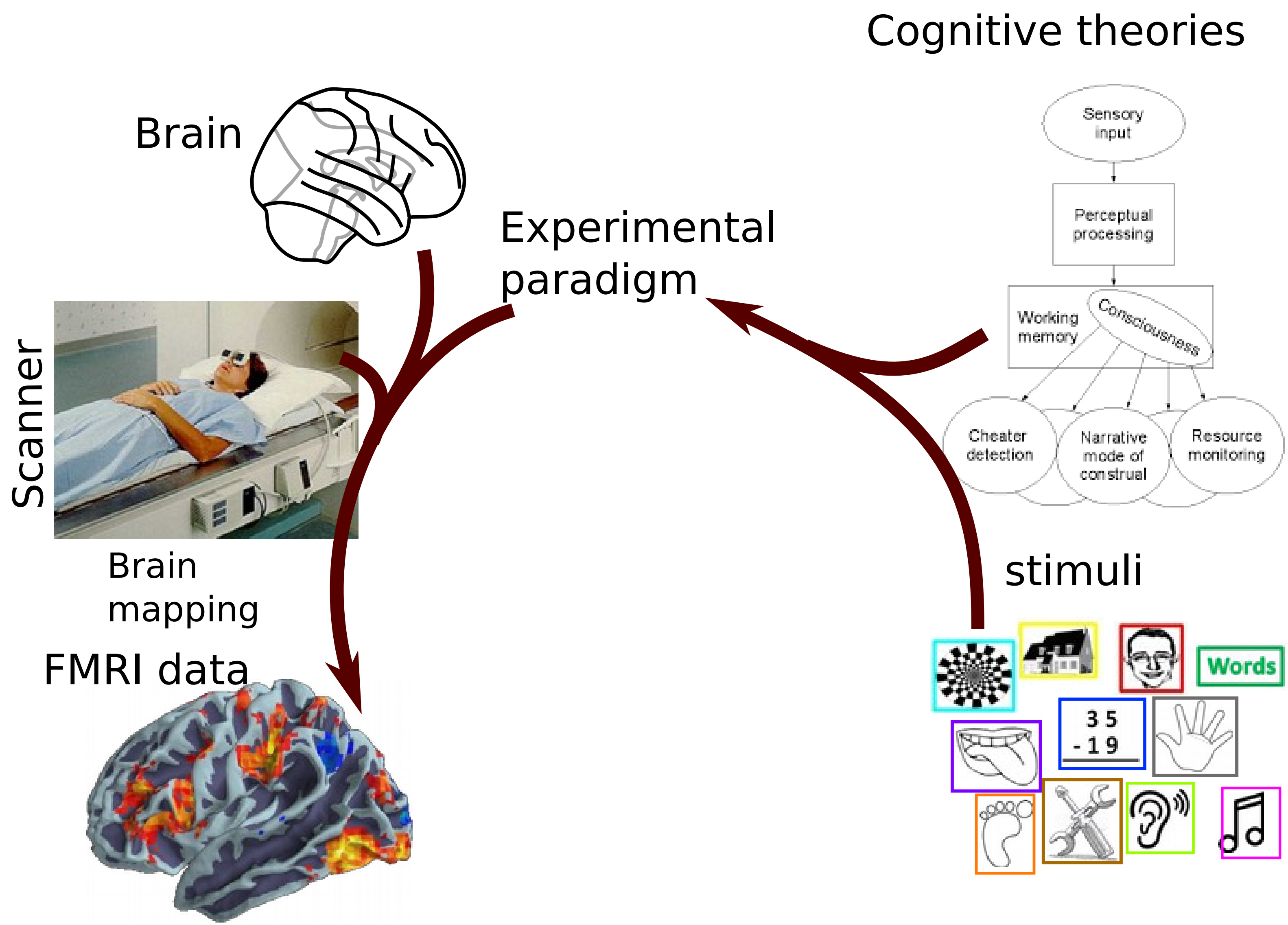
bertrand.thirion@inria.fr

INVENTEURS DU MONDE NUMÉRIQUE

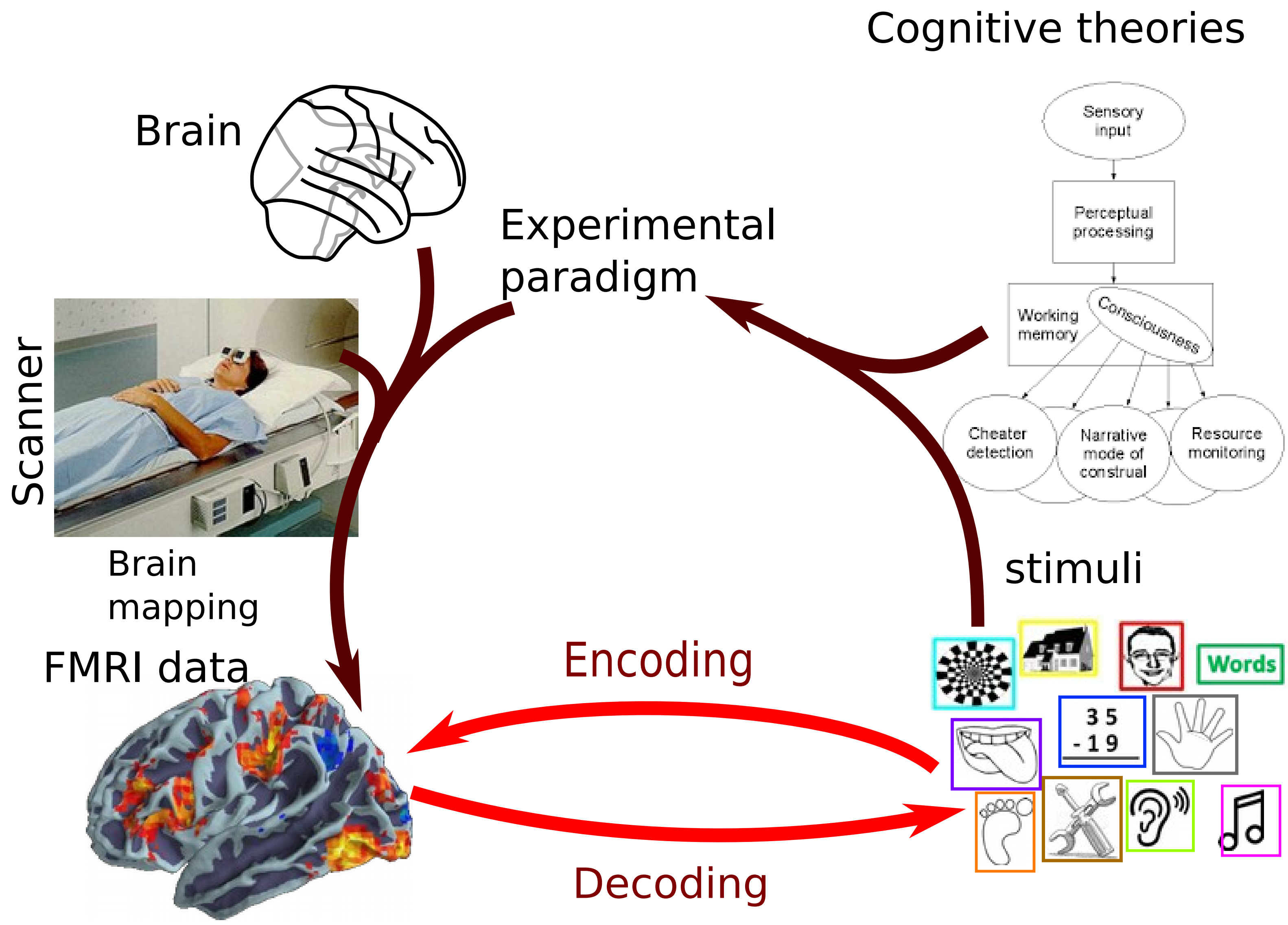
Cognitive neuroscience

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

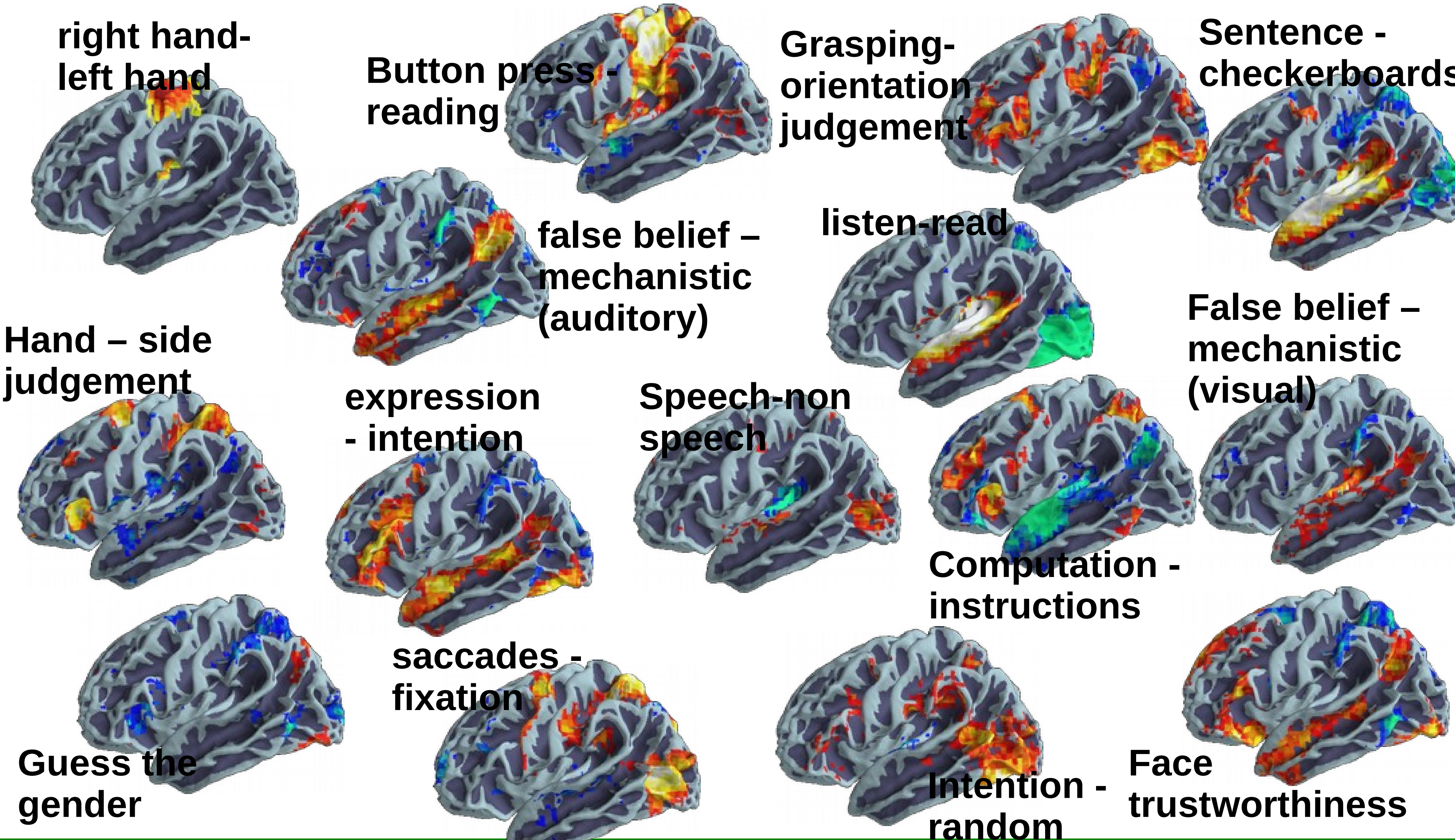
The brain, the mind and the scanner



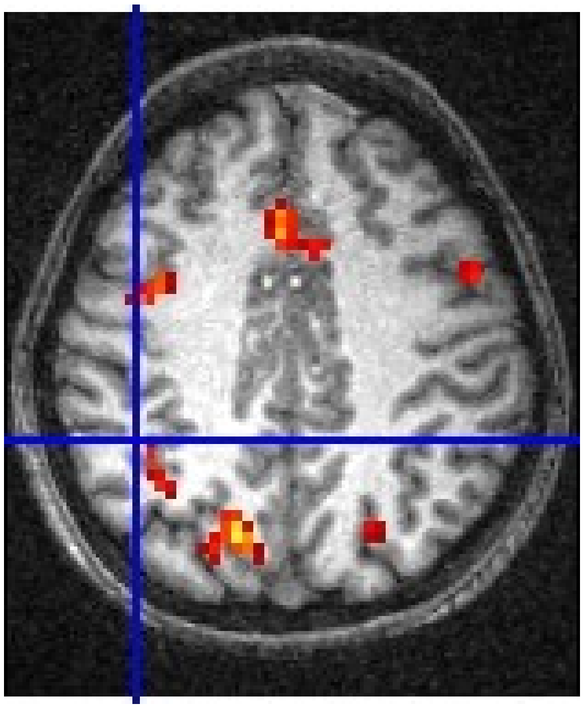
The brain, the mind and the scanner

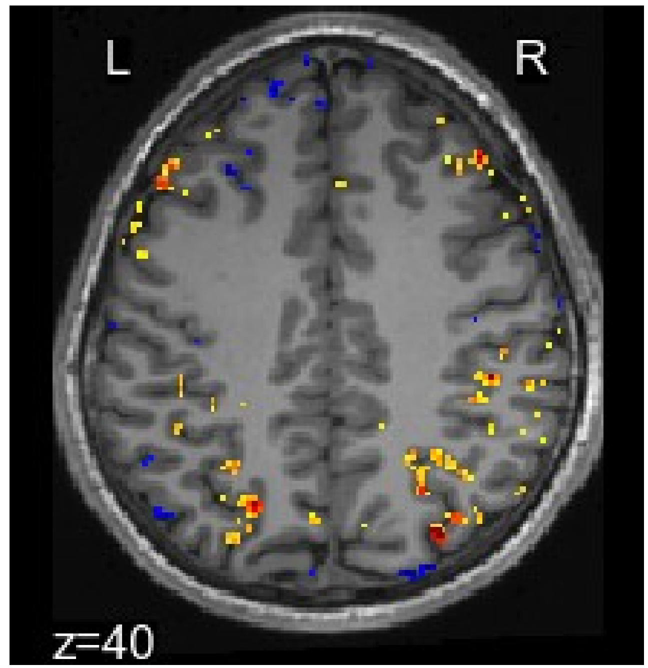


Encoding: mapping cognitive functions to brain activity



Resolution increases





2007:
3 mm

2014:
1.5 mm

2021:
0.5 mm ?

p = 50,000

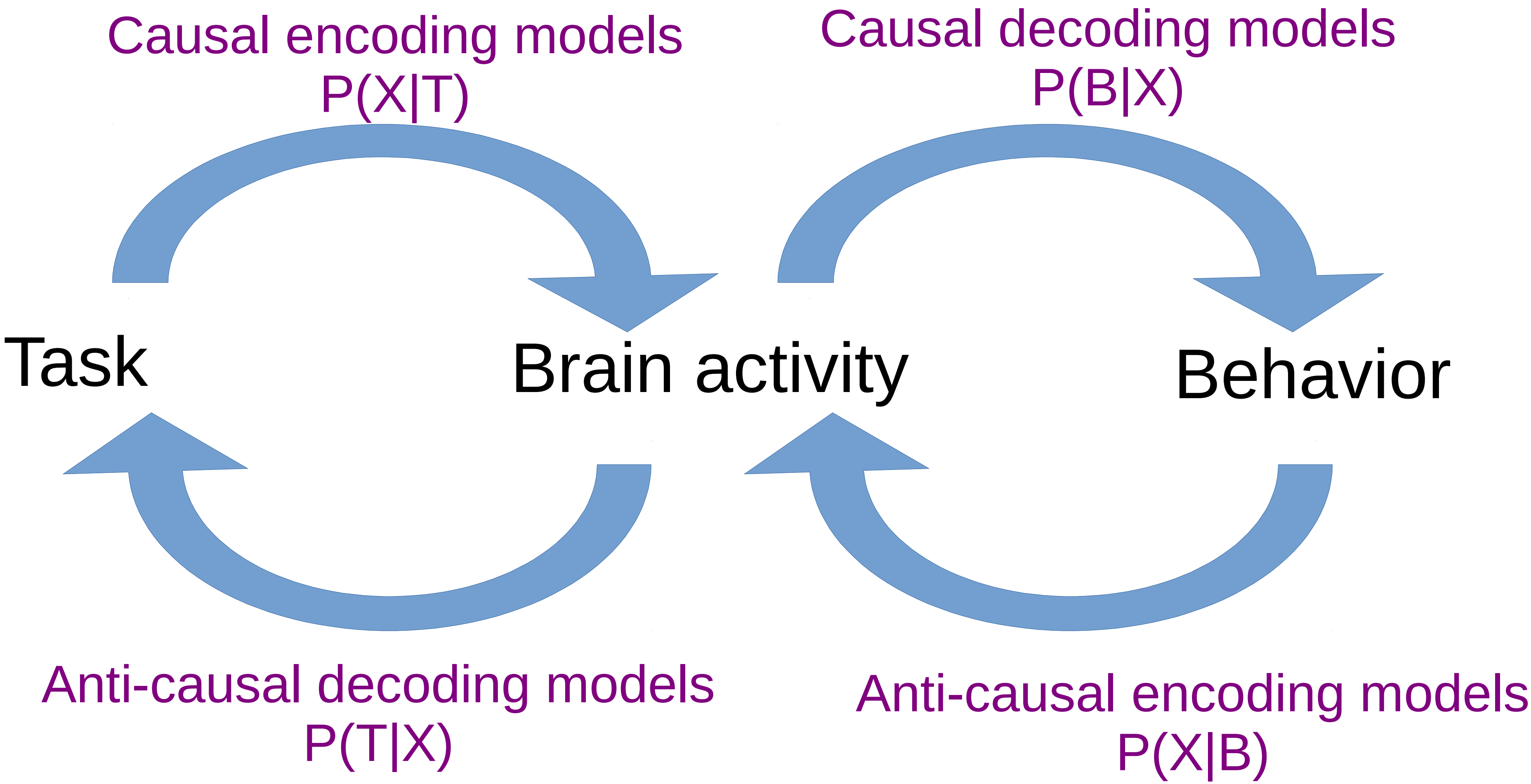
p = 400,000

p = 10^7

better estimators for large-scale brain imaging

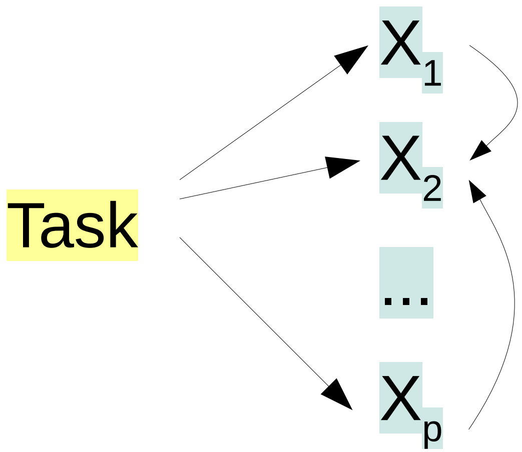
- A causal framework for brain activity decoding
- Dimension reduction for images
- Fast regularized ensembles of models
- Statistical inference for high-dimensional models

Causal reasoning on encoding/decoding



[Weichwald et al Nimg 2015]

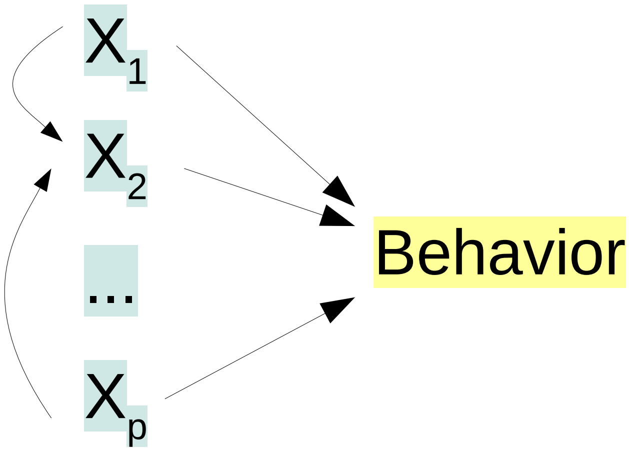
Causal interpretation



$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{T}$$

Encoding: causal
Decoding: anti-causal

$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{T} | (\mathbf{X_j}, \mathbf{j} \neq \mathbf{i})$$



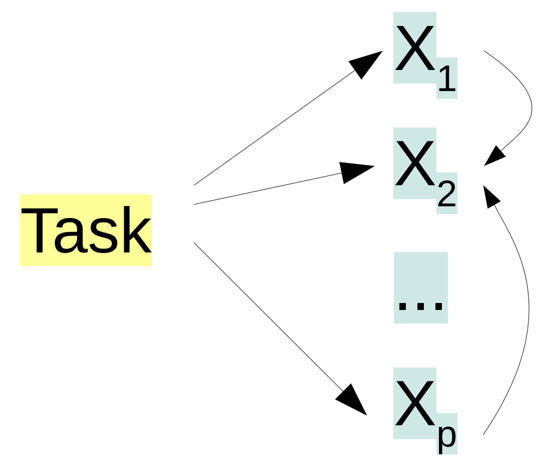
$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{B}$$

Encoding: anti-causal
Decoding: causal

$$\mathbf{X_i} \perp\!\!\!\perp \mathbf{B} | (\mathbf{X_j}, \mathbf{j} \neq \mathbf{i})$$

Causal reasoning on encoding/decoding

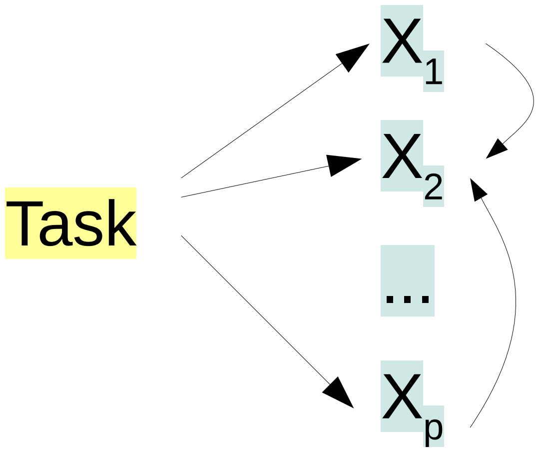
		Feature X_i relevant?		
		Encoding	Decoding	Causal interpretation
Experimental setting	Task	$\times$		$T \perp\!\!\!\perp X_i \Rightarrow X_i$ is no effect of T
		$\checkmark$		$T \not\perp\!\!\!\perp X_i \Rightarrow X_i$ is an effect of T
	Behaviour			



[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

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



[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

Experimental setting		Feature X_i relevant? Encoding	Feature X_i relevant? 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		$\times$	$T \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
	Task		$\surd$	$T \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ 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		$\times$	$B \perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive
	Behaviour		$\surd$	$B \not\perp\!\!\!\perp X_i \mid \mathbf{X} \setminus X_i \Rightarrow$ inconclusive

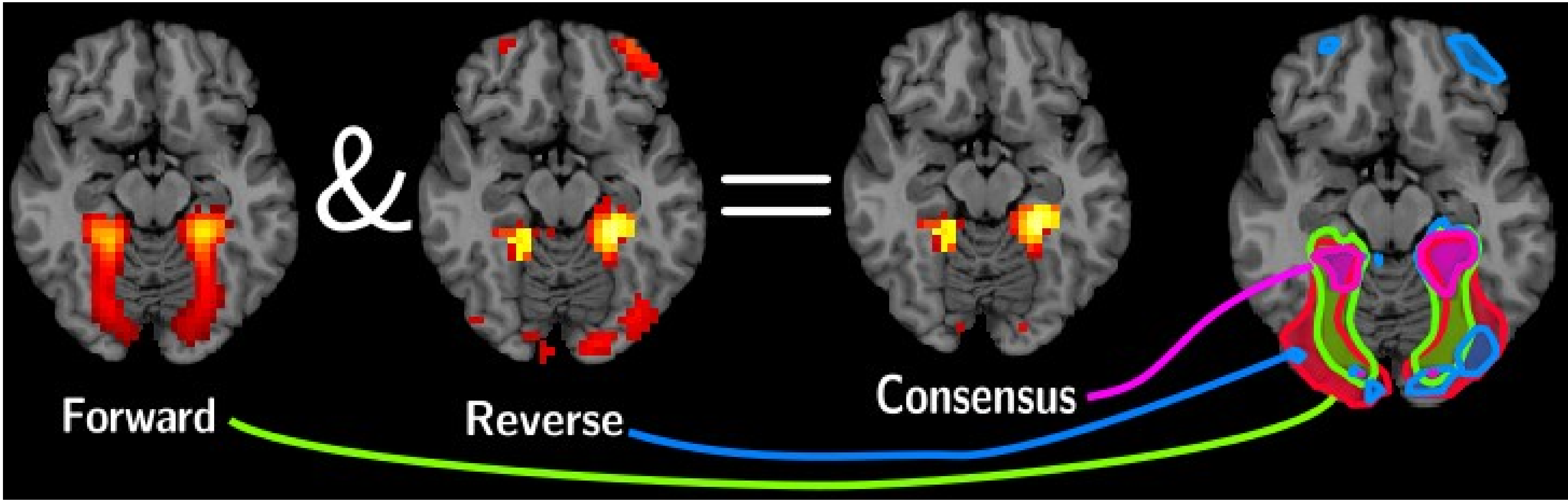
[Weichwald et al. NIMG 2015]

Causal reasoning on encoding/decoding

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

[Weichwald et al. NIMG 2015]

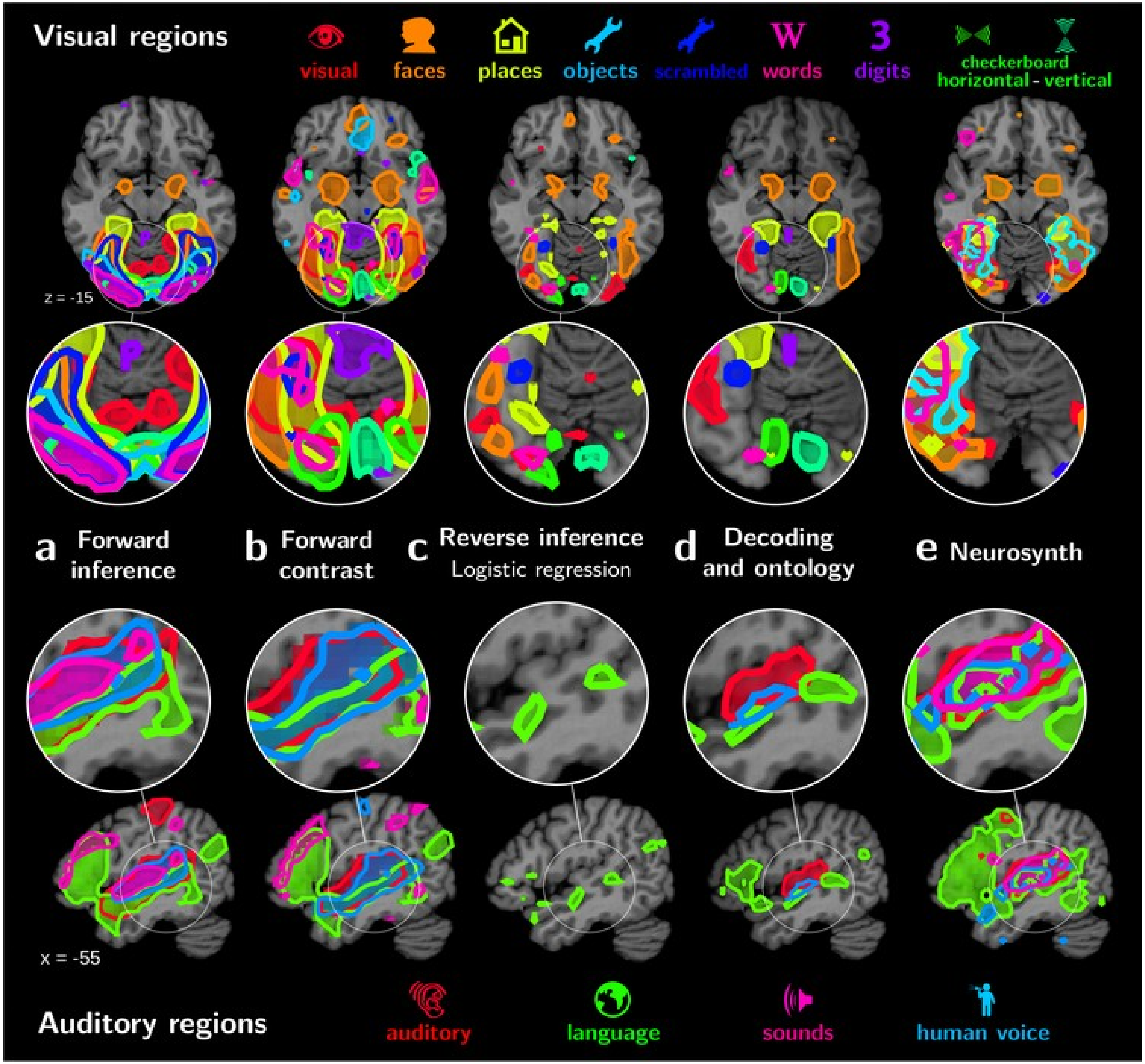
Joint encoding and decoding



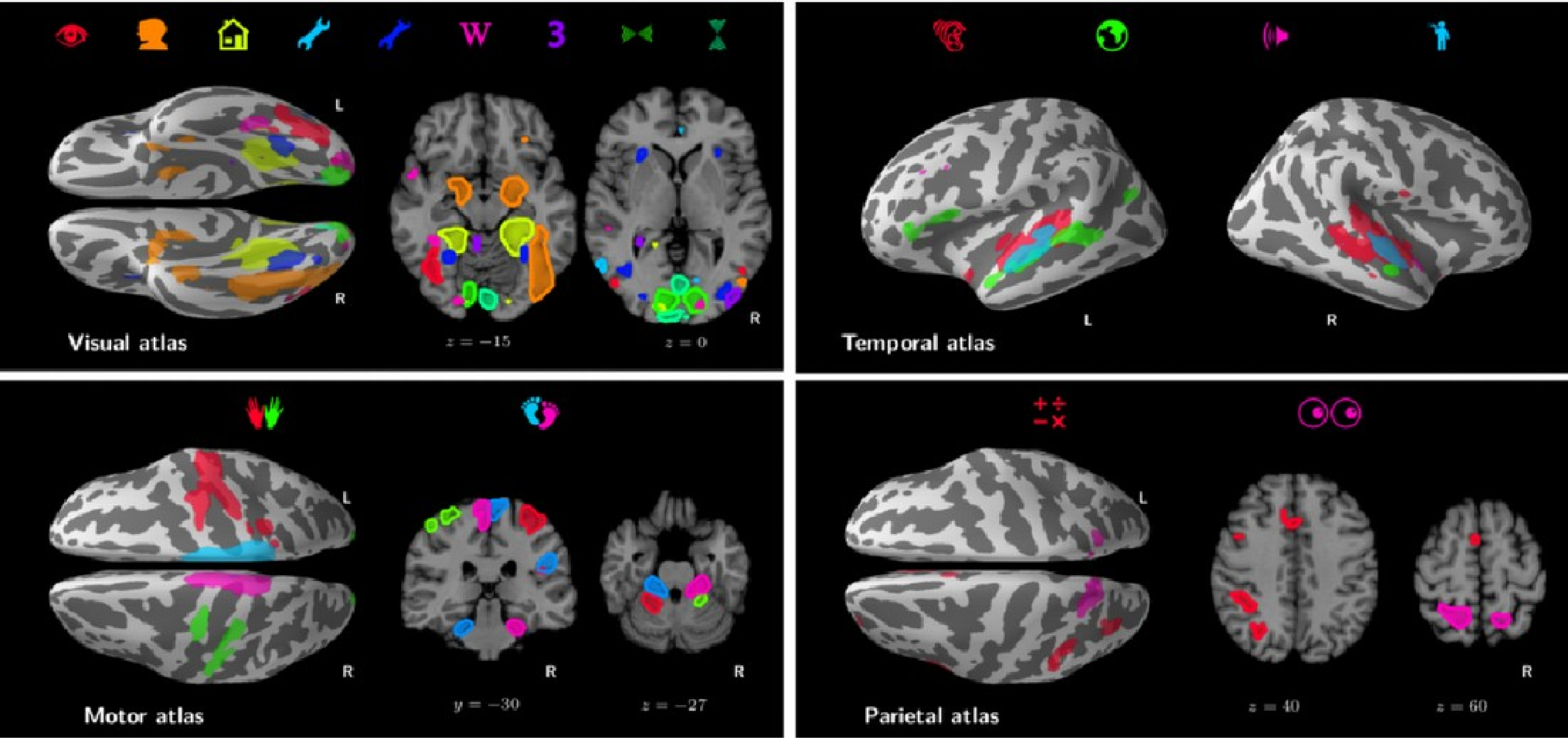
“Encoding” “Decoding”

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

Decoding maps



Joint encoding and decoding

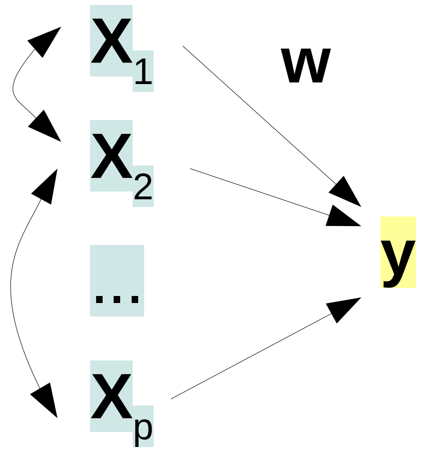


[Schwartz et al. NIPS 2013, Varoquaux et al. PCB 2018]

Statistical associations and causal reasoning

- Problems:
  - Establish non-independence based on finite datasets → statistical tests
  - **Large number of conditioning variables**
  - Encoding models: **Multiple comparison issues**
  - Decoding problem: **statistical tests in multiple regression**

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 $j \in [p]$

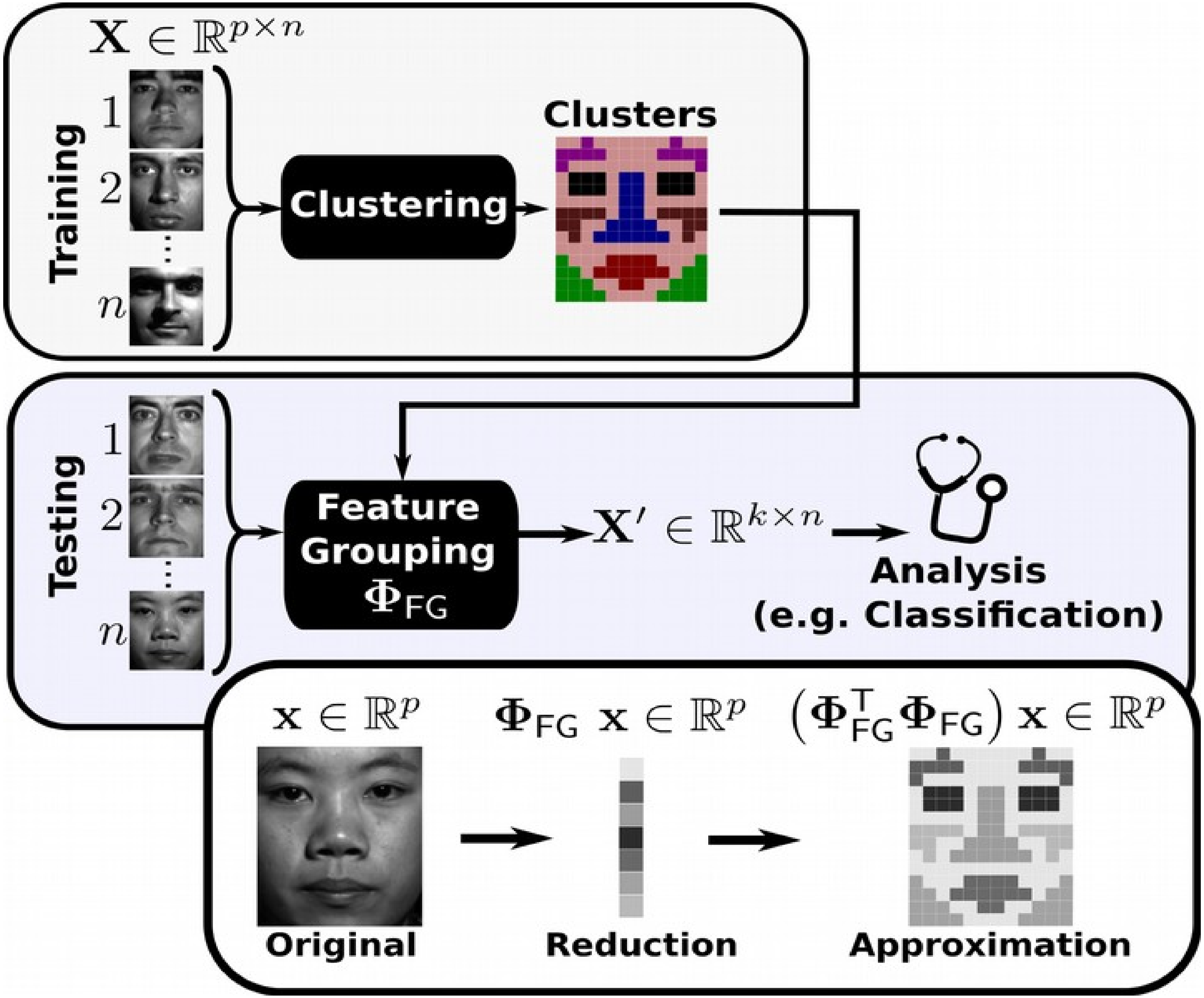
Outline

- 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 $\rightarrow$ feature grouping strategies / clustering: “super-pixels”
  - Fast clustering procedures needed (large-k regime)

Superpixels as an image operator



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 $\mathbf{x} = \mathbf{s} + \mathbf{n}$

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

- Denoising

$$\mathrm{MSE}_{\mathrm{approx}} \leq \mathrm{MSE}_{\mathrm{orig}}$$

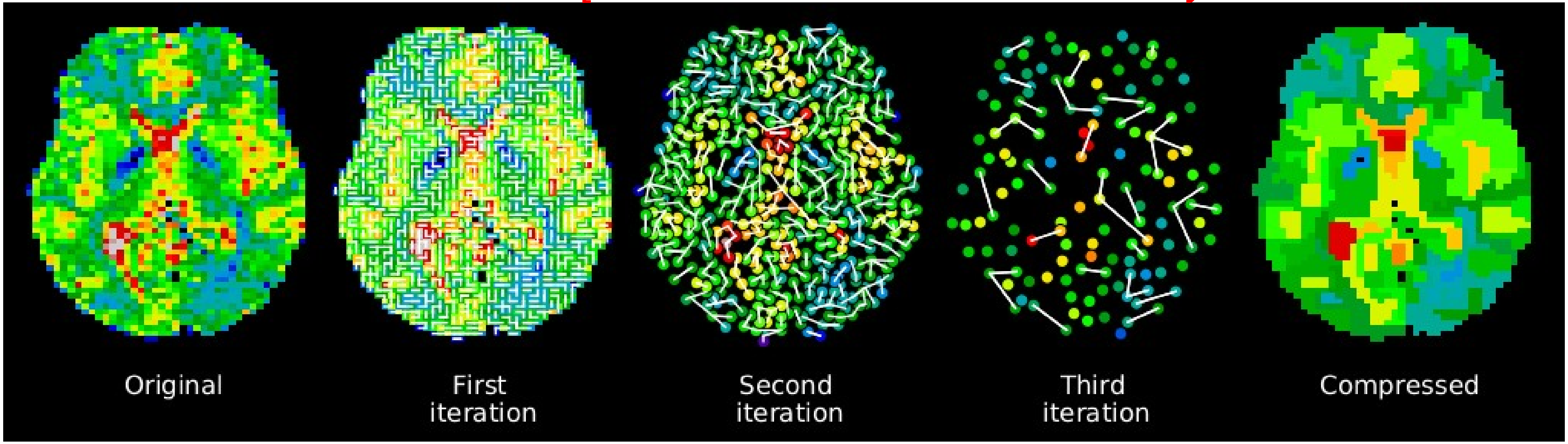
$$L^2 \leq \frac{(p-k)}{\sum_{q=1}^{k} |\mathcal{C}_q| \, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2} \sigma^2$$

- Equal-size clusters

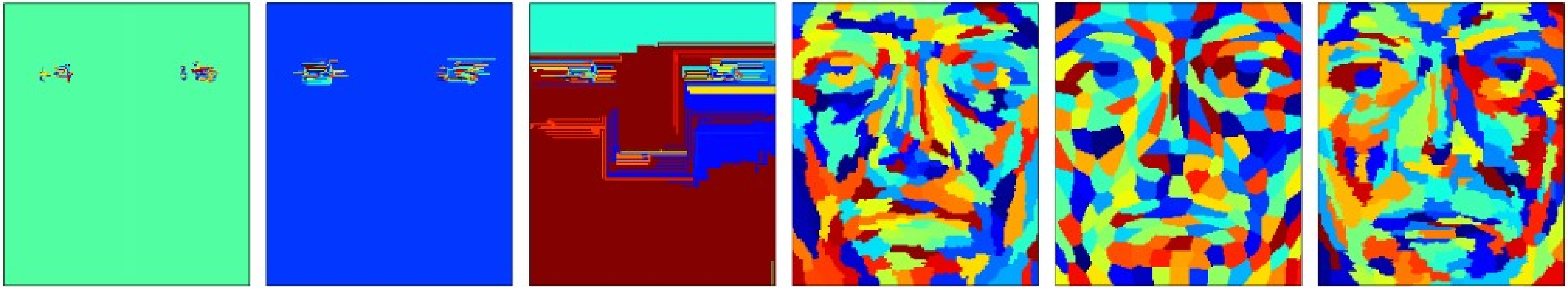
$$\mathrm{MSE}_{\mathrm{approx}} \leq p \left(\frac{L}{k}\right)^2 + \frac{k}{p} \mathrm{MSE}_{\mathrm{orig}} = O\left(\max\left\{\frac{p}{k^2}, \frac{k}{p}\right\}\right)$$

Recursive neighbor Agglomeration

[Thirion et al. Stamlins 2015, Hoyos Idrobo PAMI 2018]

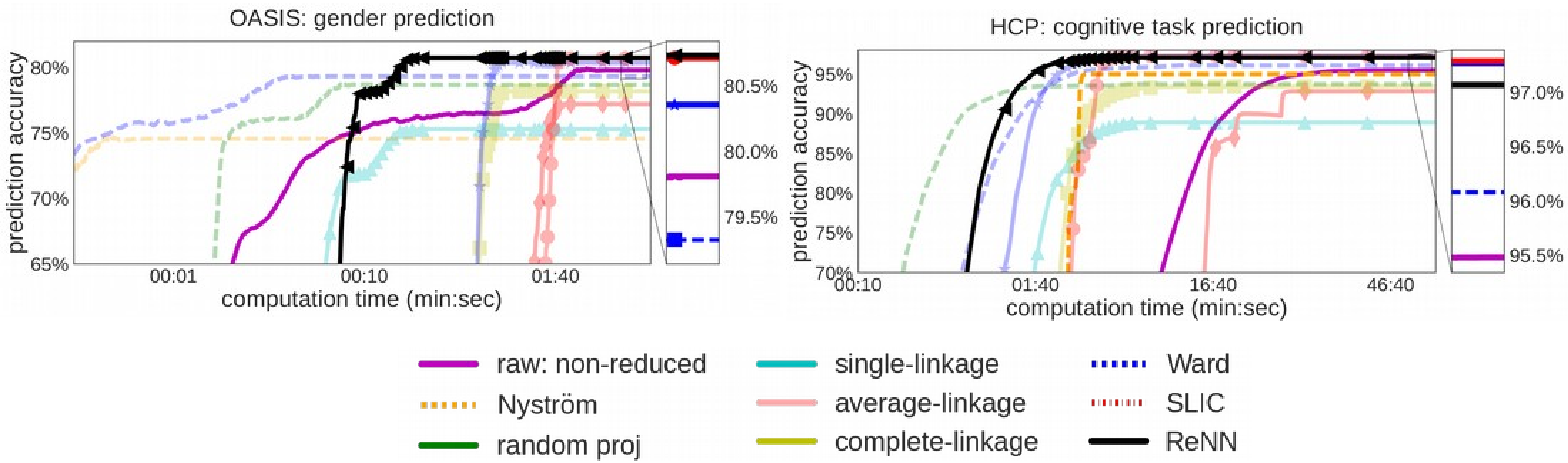


Based on local decisions = fast (linear time) – avoid percolation



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

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 2018]

Outline

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

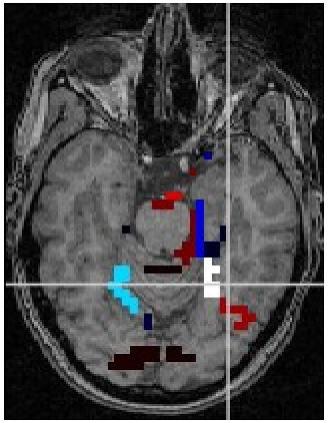
Bagging of clustered models

X

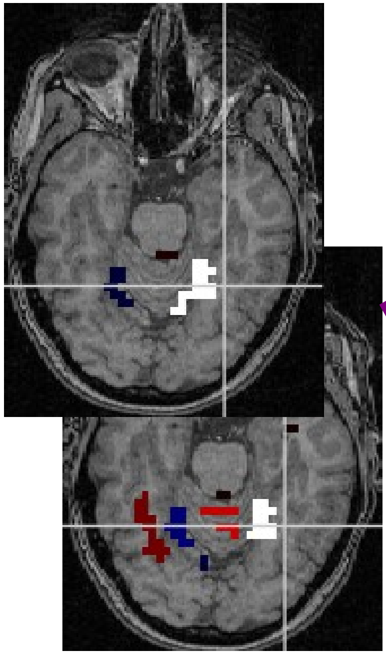
Clustering (create contiguous regions)

y

Solve **regression** on cluster-based representation

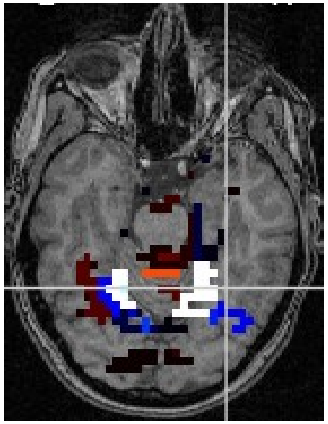


...



various clusterings

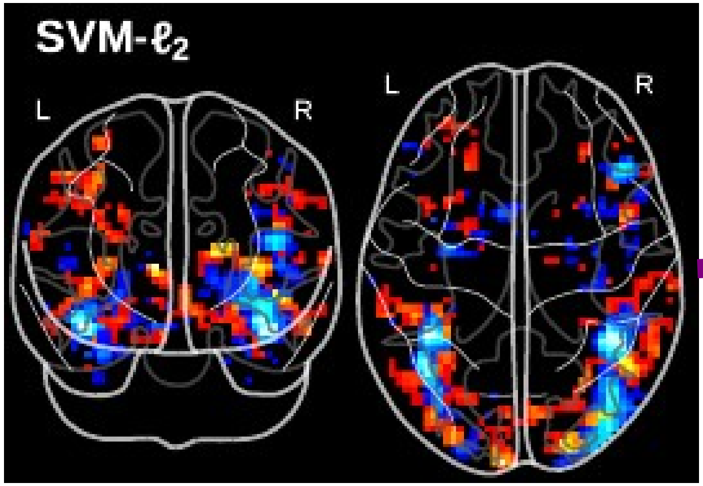
average

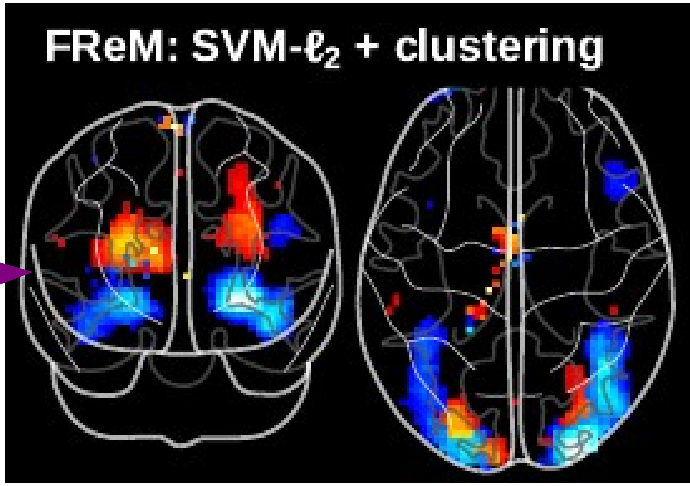


Computationally efficient structure

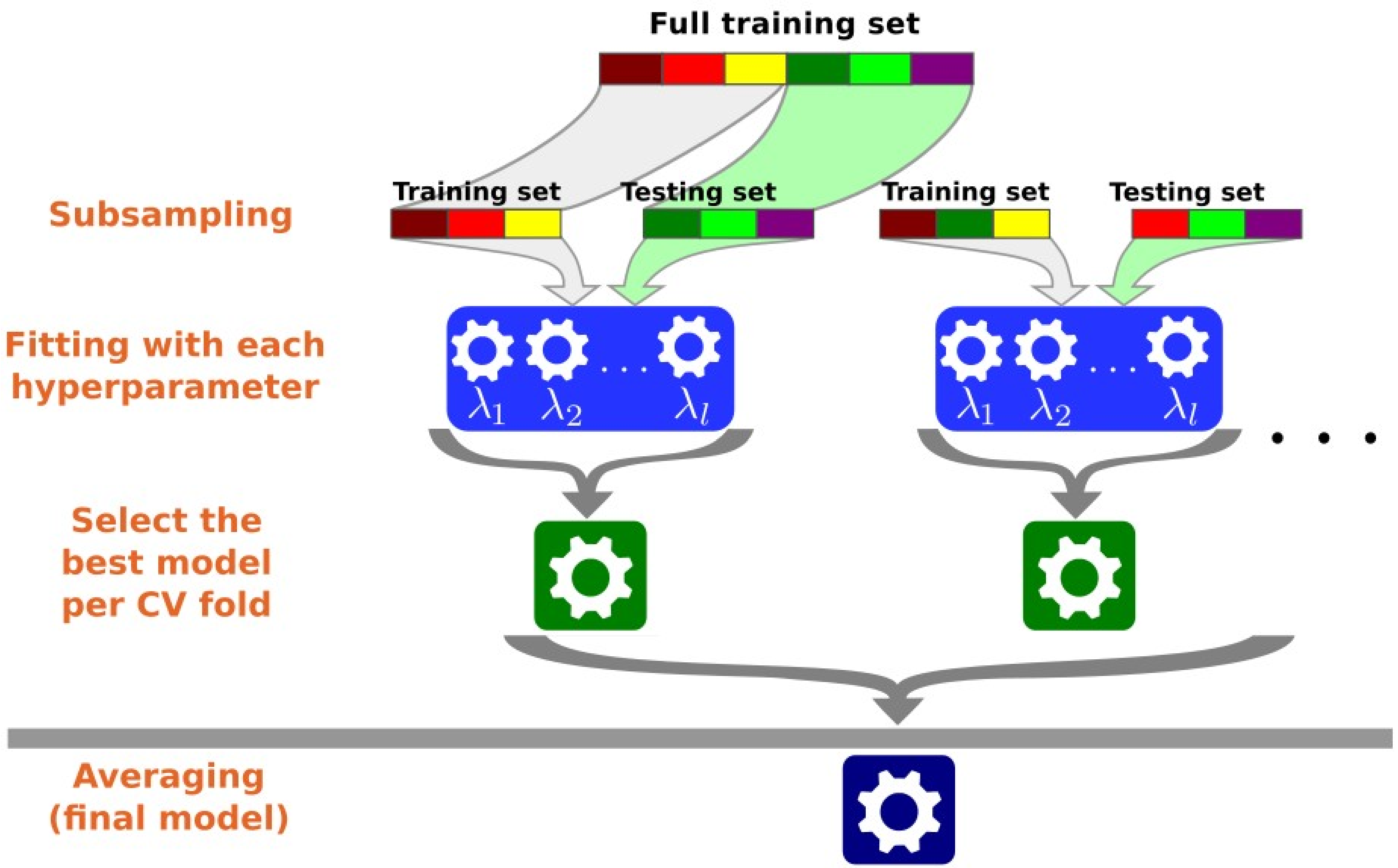
“fast regularized ensembles of models”

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



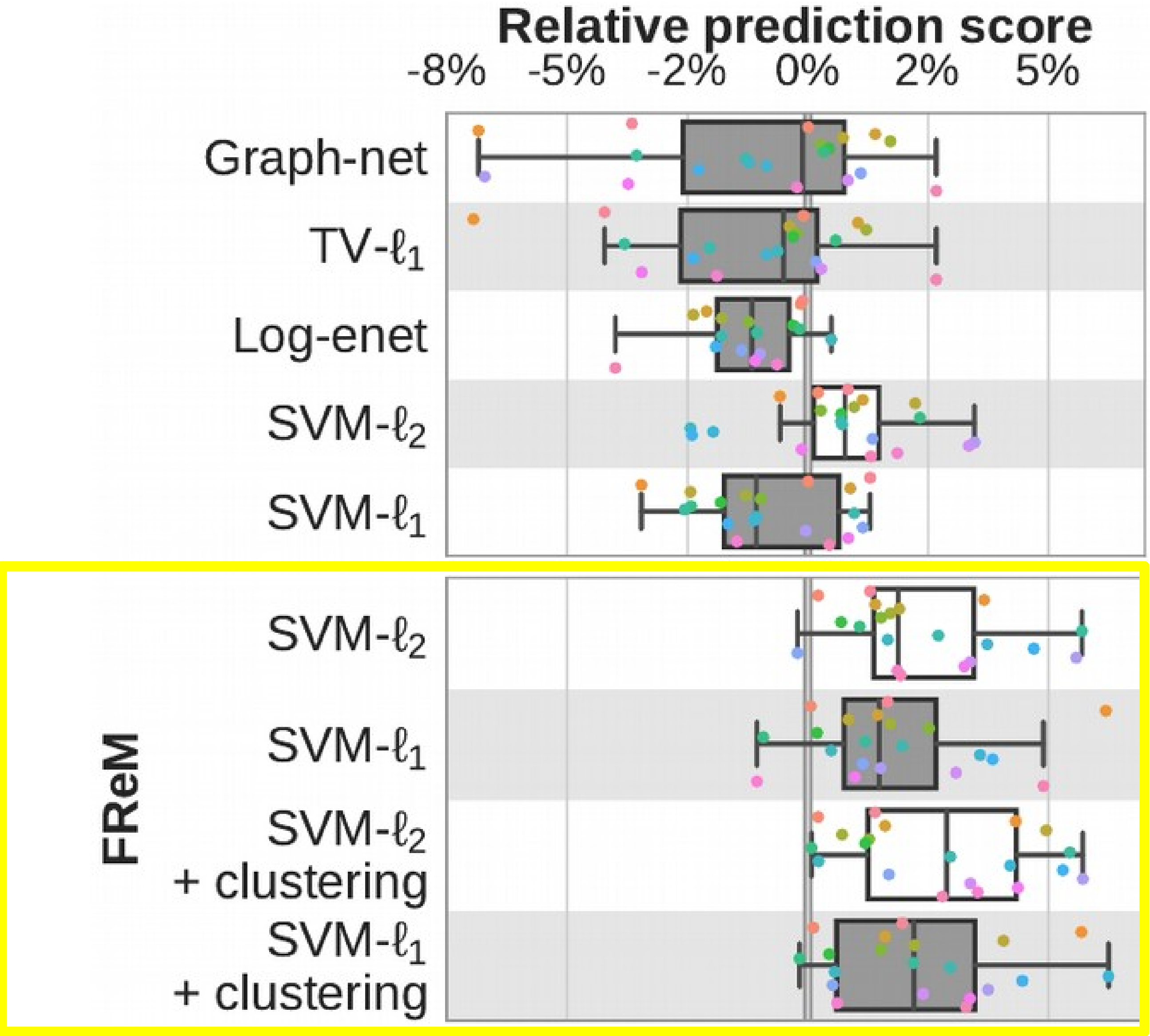


Computationally efficient structure



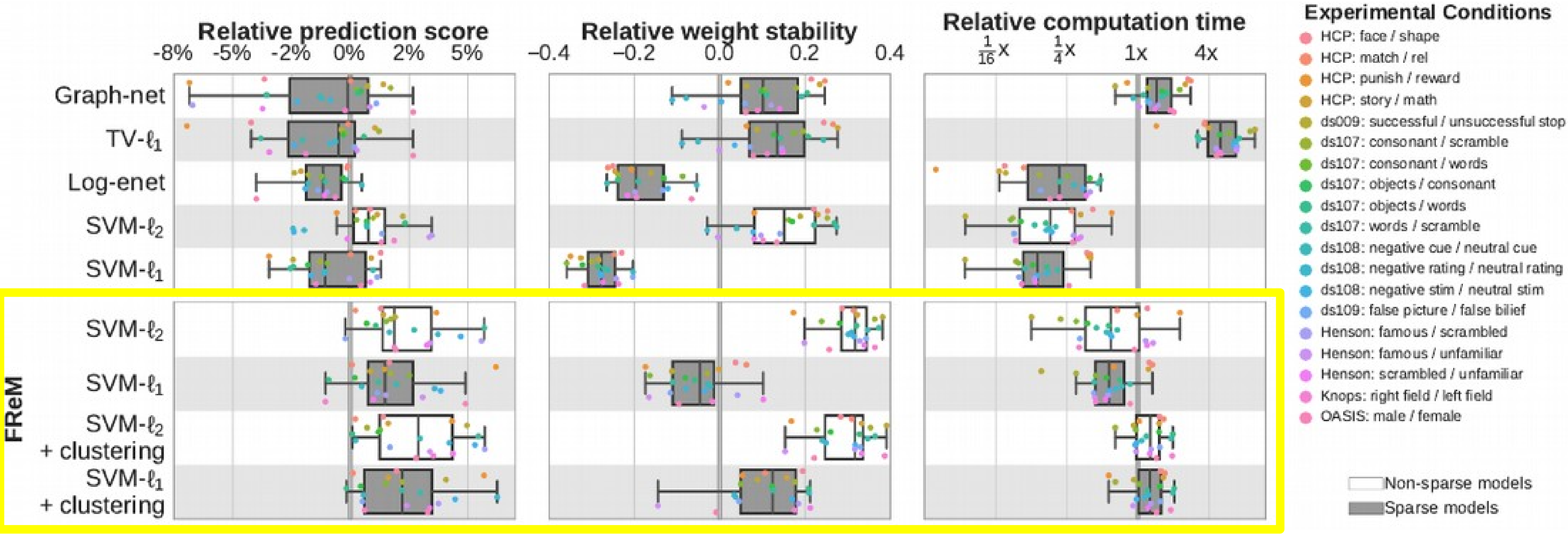
Effect on prediction accuracy

[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 2018]

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 ($p \cong n$)
  - 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 $\hat{\boldsymbol{w}}^{\text{OLS}}$)
  - compute its covariance matrix
- **Heuristic argument:** in low dimension we can prove that:

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

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 $\text{Span}(\mathbf{X}^{(-j)}) \subset \mathbb{R}^{p-1}$

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:

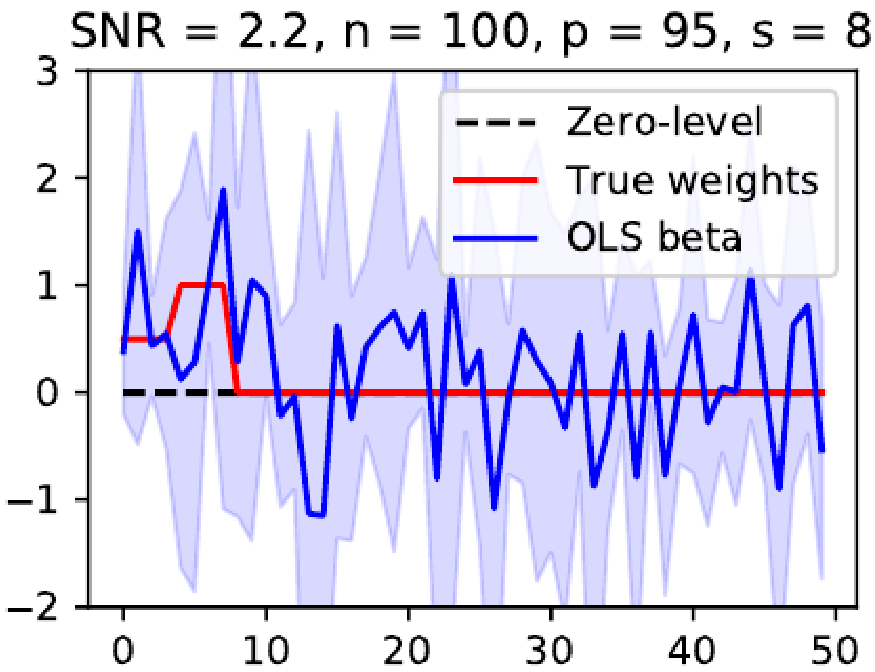
$$\Omega_{jk} = \frac{n \mathbf{z}_j^\top \mathbf{z}_k}{(\mathbf{z}_j^\top \mathbf{x}_j)(\mathbf{z}_k^\top \mathbf{x}_k)}$$

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

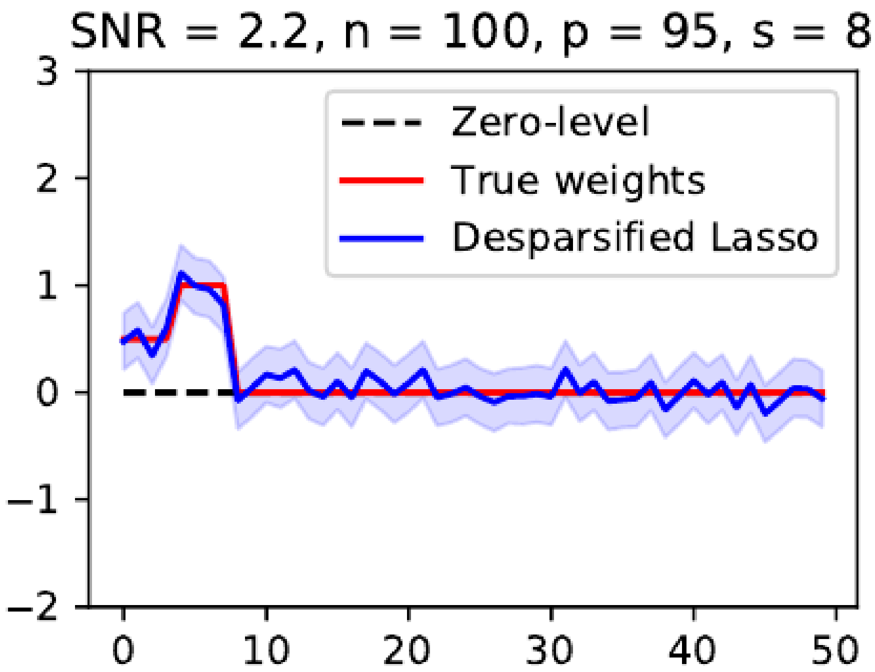
$$\boxed{\sigma_*^{-1} (\Omega_{jj})^{-1/2} (\hat{w}_j - w_j^*) \sim \mathcal{N}(0, 1)}$$

Preliminary assessment

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



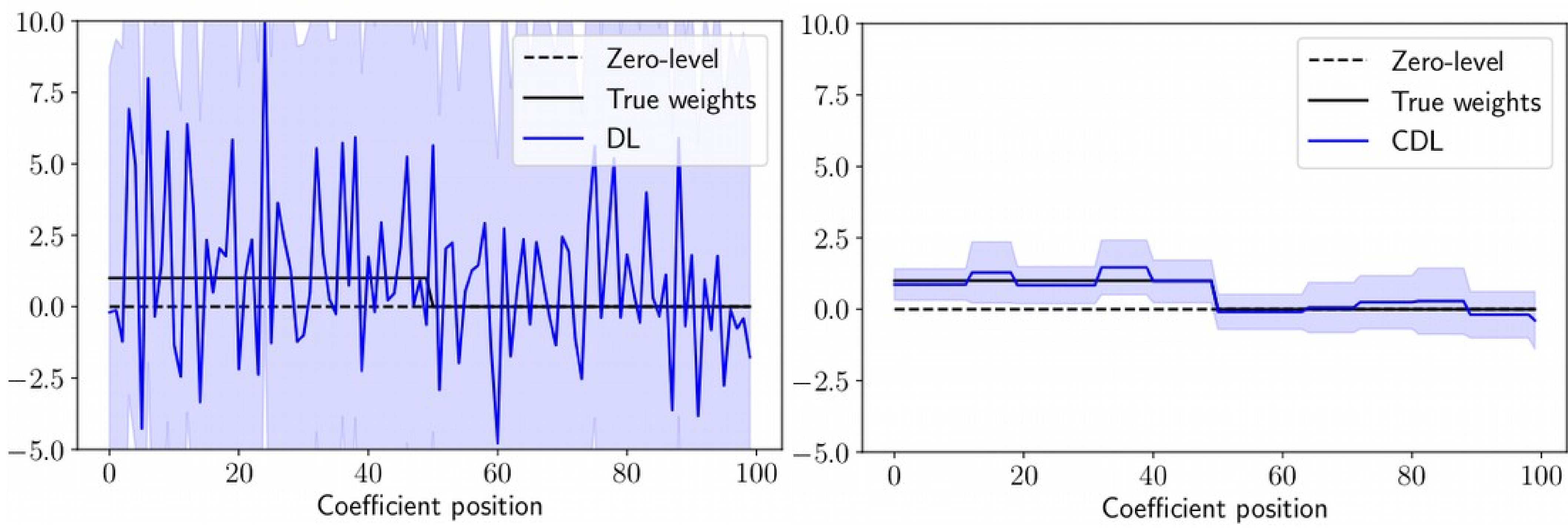
OLS regression when $p \approx n$



Desparsified Lasso when $p \approx n$

Large p → need dimension reduction

p=2000, n=100



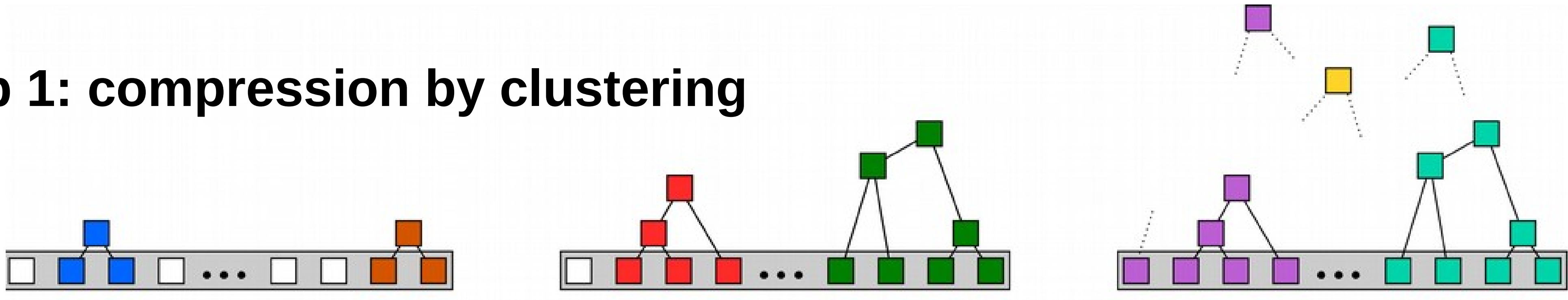
Large p kills statistical power

CDL tames variance

[Chevalier et al. subm. To MICCAI]

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*

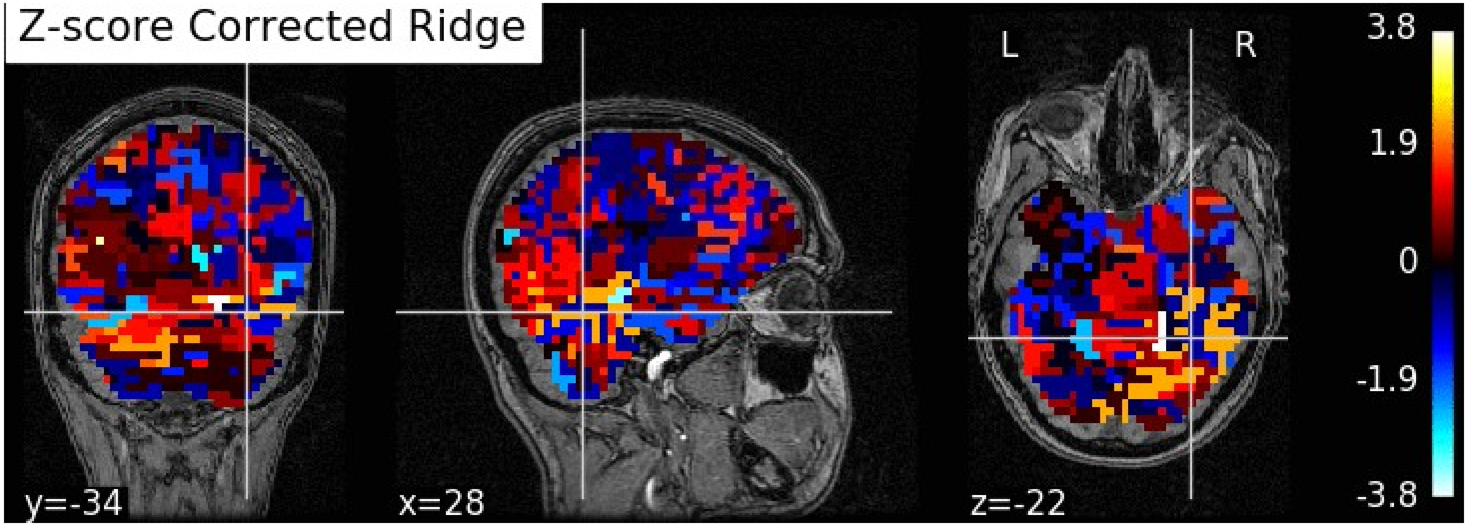
Step 3: ensembling iterate with different parcellations
→ **aggregate p-values** (*see also* FReM)

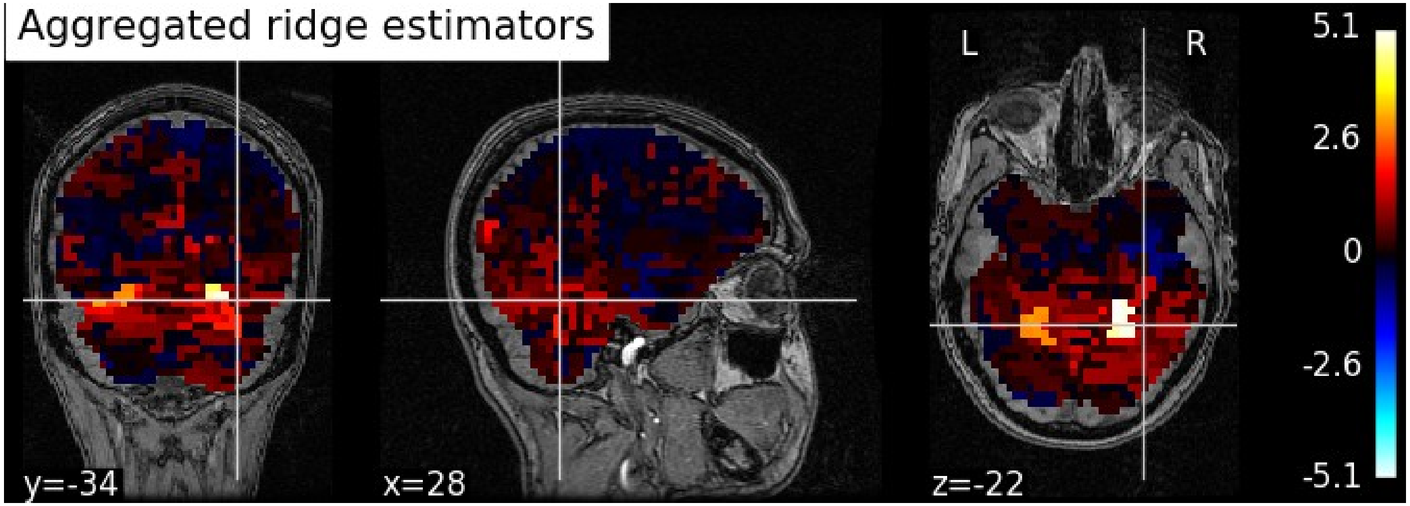
***E**nsemble of*
***C**lustered*
***D**esparsified*
***L**asso*

From CDL to ECDL

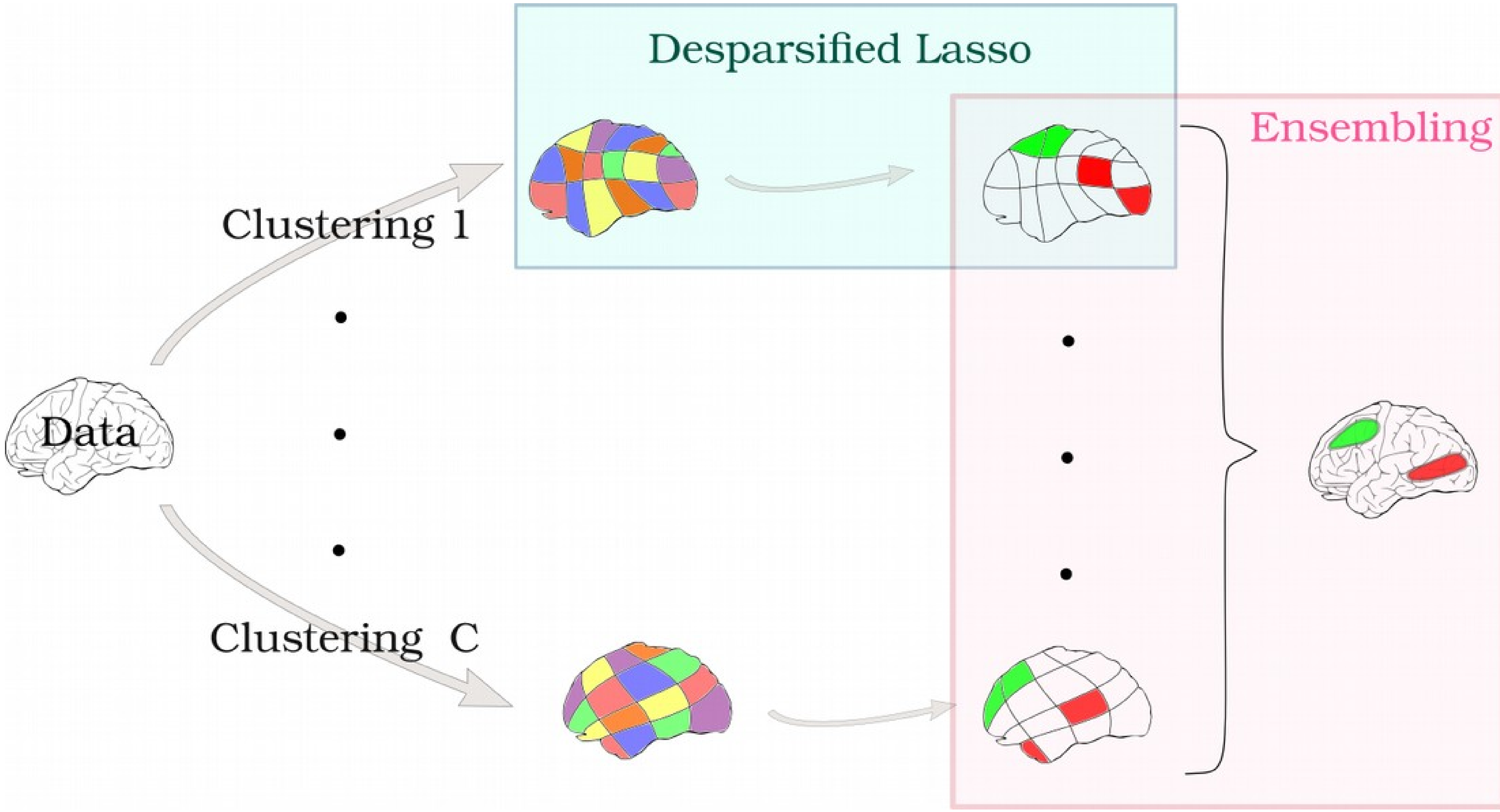
DL p-values
from different
clusterings

aggregation





ECDL for brain imaging



δ-error control

Definition (δ-null region)

The set of covariates that verify the δ-null hypothesis in the true model is called the δ-null region and is denoted by N^δ:

$$N^\delta = \{j \in [p] : \forall k \in [p], d(j,k) \leq \delta \implies \mathbf{w}^*_k = 0\} \quad .$$

Definition (Rejection region)

Given a family of p-values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ and a threshold $\alpha \in (0,1)$, we call rejection region at level α for the family $\hat{p}$ the indexes of covariates whose corresponding p-value are lower than α and denote it by $R_\alpha(\hat{p})$:

$$R_\alpha(\hat{p}) = \{j \in [p] : \hat{p}_j \leq \alpha\} \quad .$$

δ-error control

Definition (δ-type 1 error region)

Given a family of p-values $\hat{p} = (\hat{p}_j)_{j\in[p]}$ and a threshold $\alpha \in (0,1)$, the δ-type 1 error region (or erroneous rejection region with tolerance δ) at level α is the set of covariates indexes belonging both to the δ-null region and to the rejection region at level α:

$$\mathscr{E}_\alpha^\delta(\hat{p}) = N^\delta \cap R_\alpha(\hat{p}) \ .$$

Definition (δ-family wise error rate)

Given a family of p-values $\hat{p} = (\hat{p}_j)_{j\in[p]}$ and a threshold $\alpha \in (0,1)$, the δ-FWER, denoted $FWER_\alpha^\delta(\hat{p})$, is the probability that the δ-type 1 error region at level α is not empty:

$$FWER_\alpha^\delta(\hat{p}) = \mathbb{P}(|\mathscr{E}_\alpha^\delta(\hat{p})| \geq 1) = \mathbb{P}(\min_{j\in N^\delta} \hat{p}_j \leq \alpha) \ .$$

δ-FWER control

Definition (δ-FWER control)

We say that the family of p-values $\hat{p} = (\hat{p}_j)_{j \in [p]}$ *controls the* δ*-FWER if, for all* $\alpha \in (0, 1)$*:*

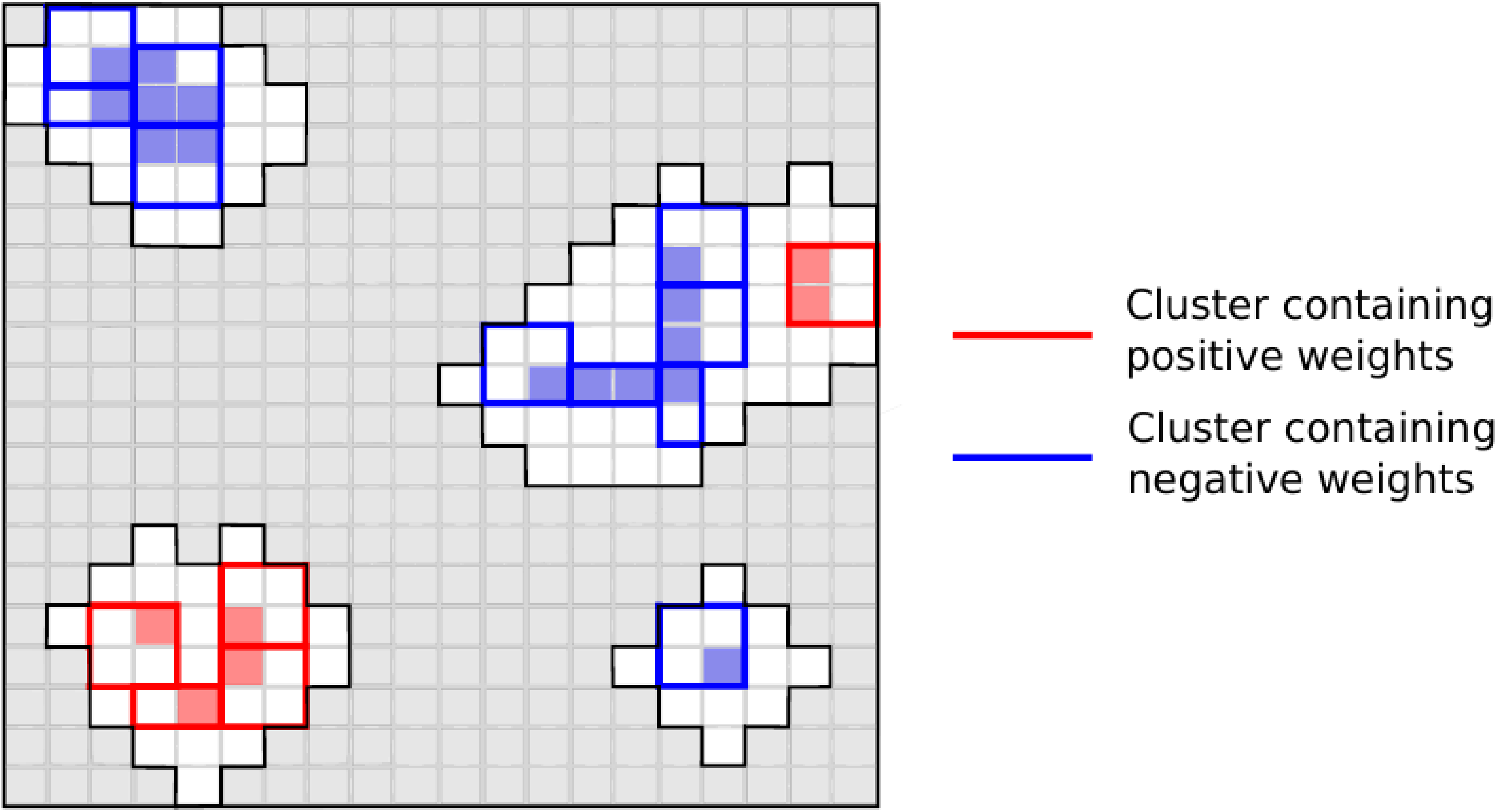
$$FWER_{\alpha}^{\delta}(\hat{p}) \leq \alpha \ .$$

Proposition

Under the assumptions for the vanilla Desparsified Lasso and assumptions on the weight map and on the data structure we have the following result:

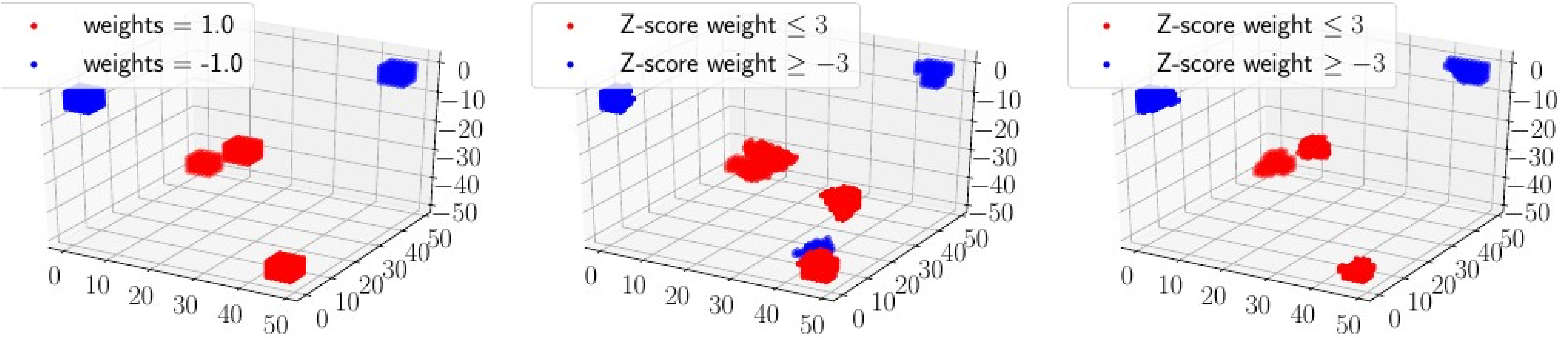
If the diameters of the clusters are all smaller than δ *then the p-value family computed through the ECDL algorithm controls the* δ*-FWER.*

δ-FWER-control



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:**



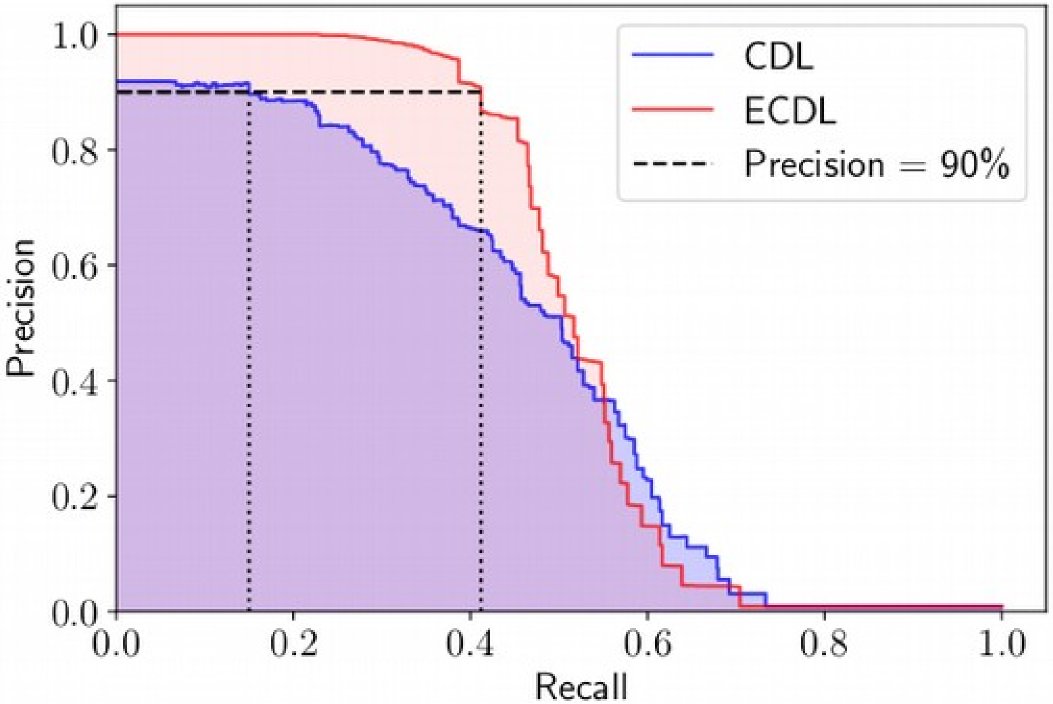
(a) weight vector: $\mathbf{w}^*$ (b) CDL (c) ECDL

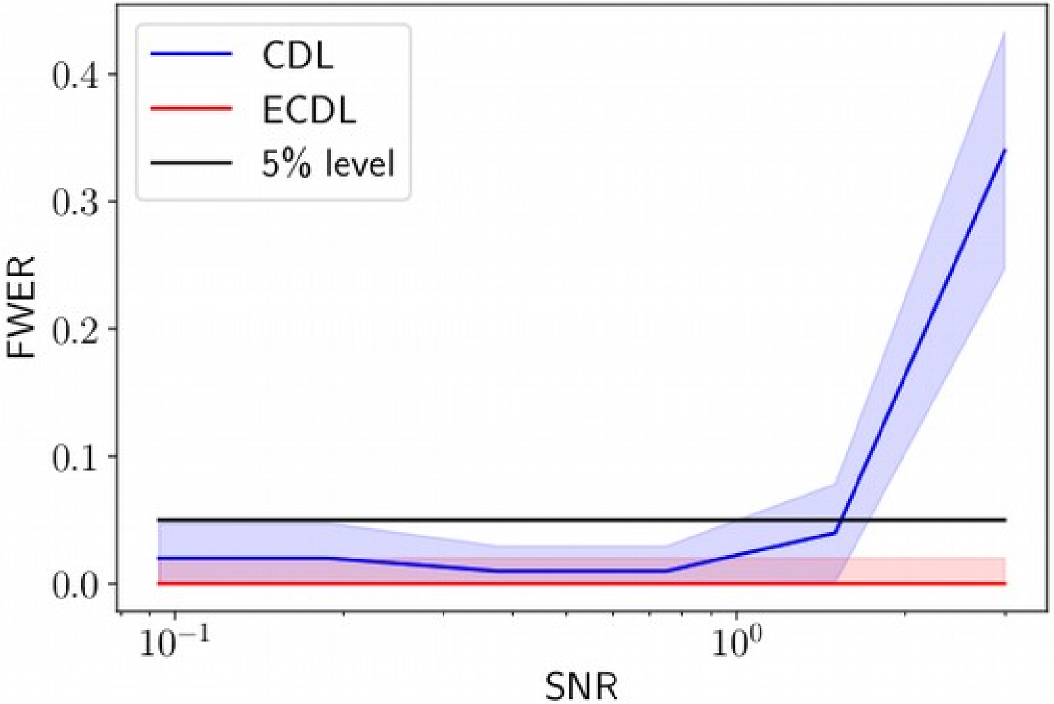
[Chevalier et al. MICCAI 2018]

Experiments: PR and FWER control

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

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





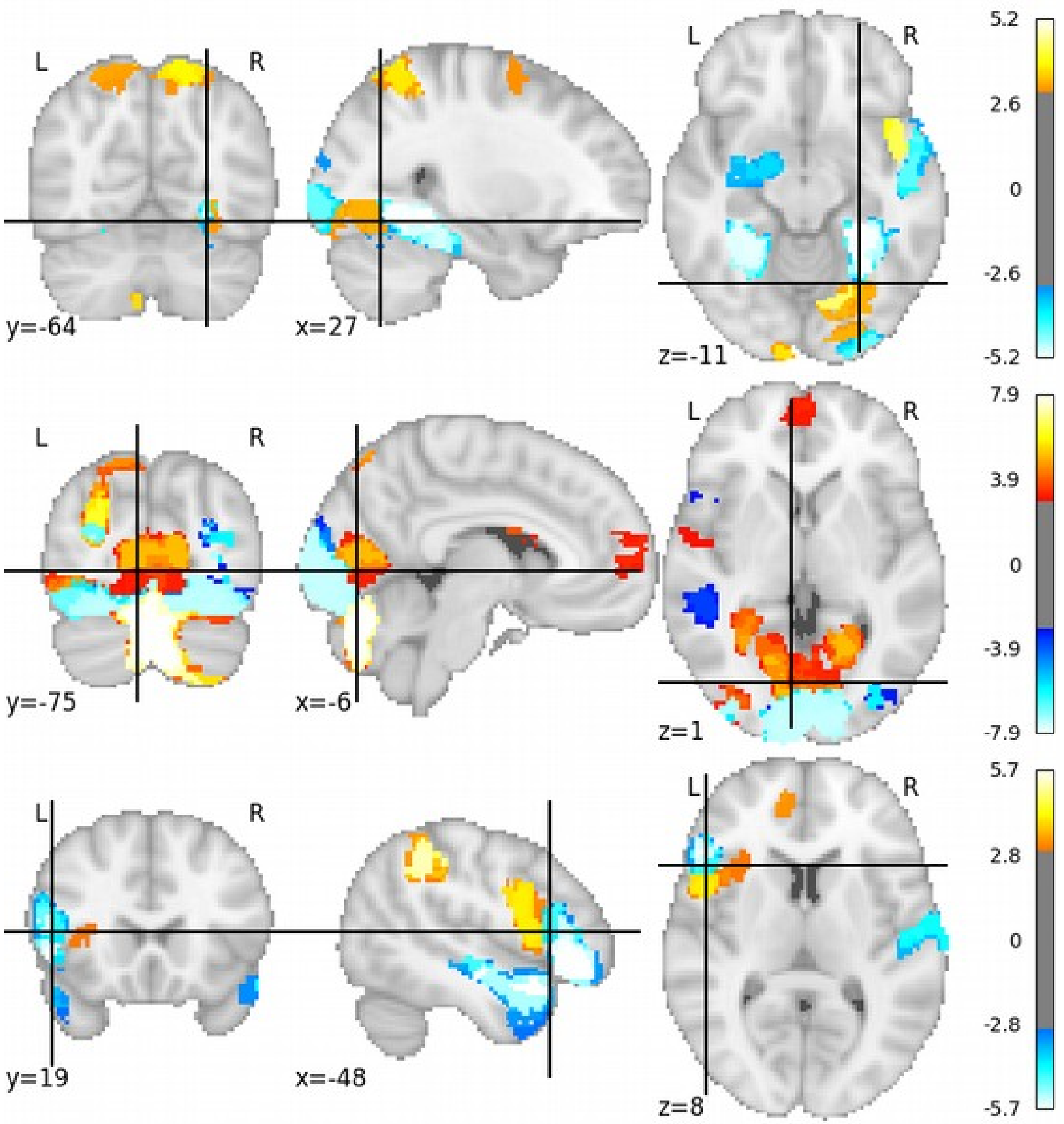
Better PR with ECDL

+ More accurate FWER control

[Chevalier et al. MICCAI 2018]

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]

Conclusion

- Causal reasoning →
  conditional association analysis
- Large-p data bring challenges:
  - Computation cost
  - Difficulty of statistical inference
- Solutions: ensembling, subsampling, compression
- Efficient stochastic regularizers
- Ongoing comparison with knockoff

[Nguyen et al. IPMI 2019]

WIP

- Classification setting
- Use of bootstrap

[Aydore et al. subm]

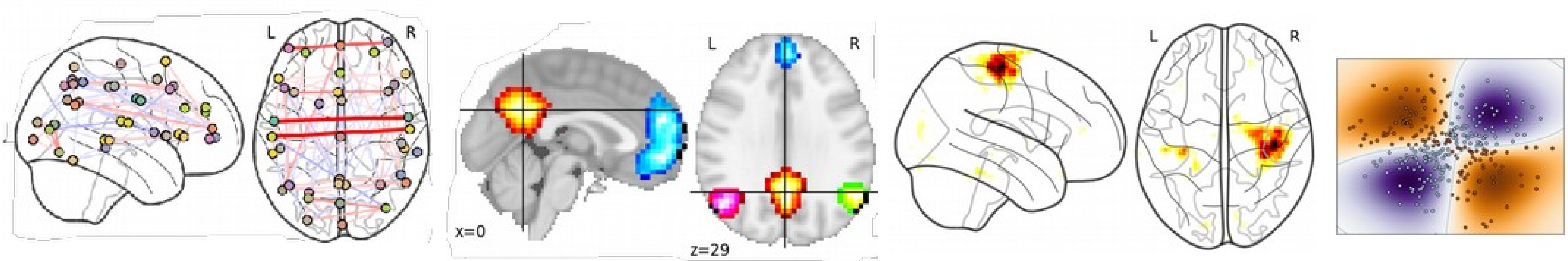
From good ideas to good practices: software

scikits learn machine learning in Python

ni learn

MNE
MEG + EEG ANALYSIS & VISUALIZATION

- Machine learning in Python
- Machine learning for neuroimaging
  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,
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

INSTITUT DATAIA
Science des données, Intelligence & Société

HBP Human Brain Project

université PARIS-SACLAY

AGENCE NATIONALE DE LA RECHERCHE ANR