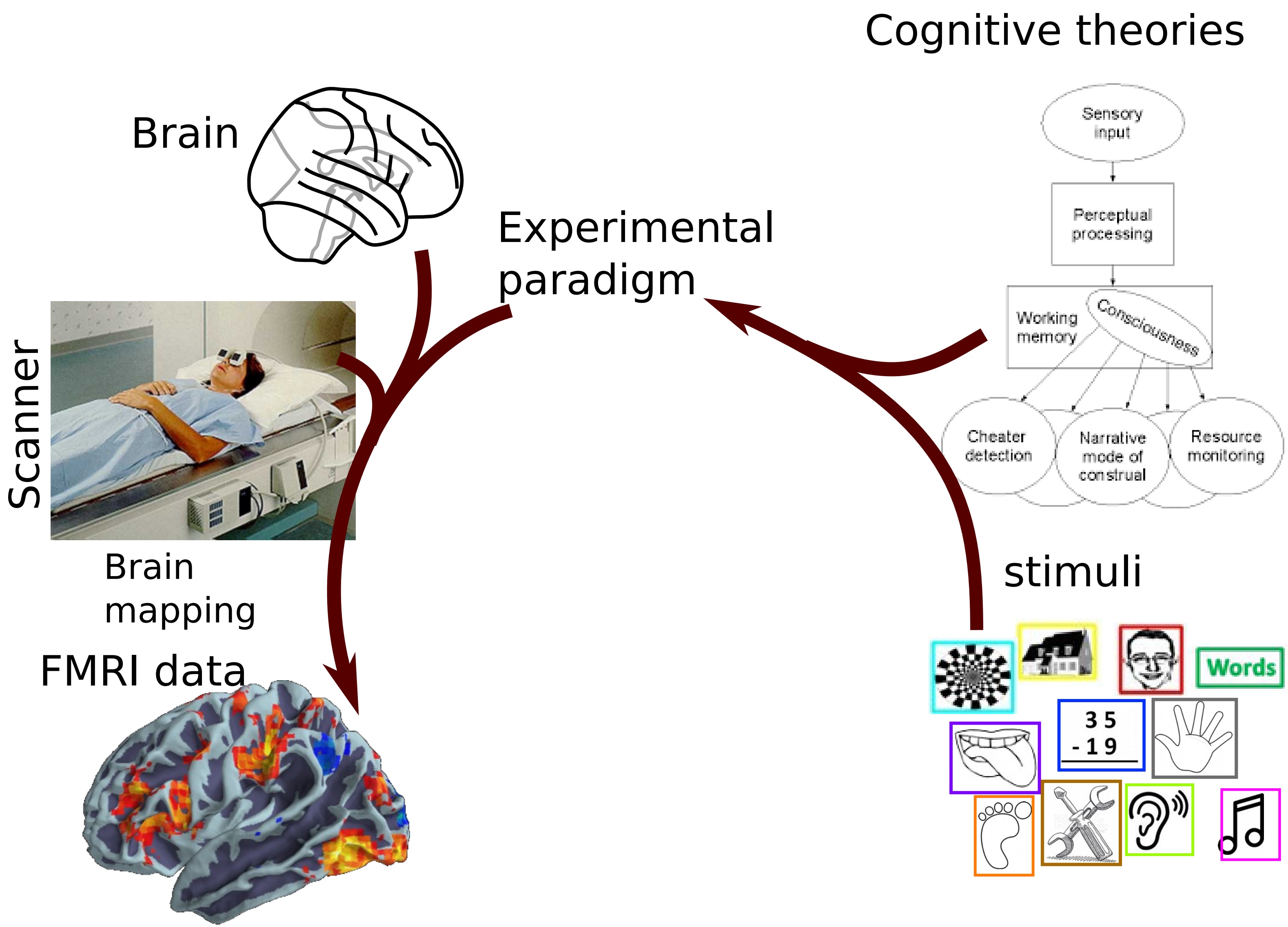
Toward a rigorous statistical framework for brain mapping

Bertrand Thirion, bertrand.thirion@inria.fr

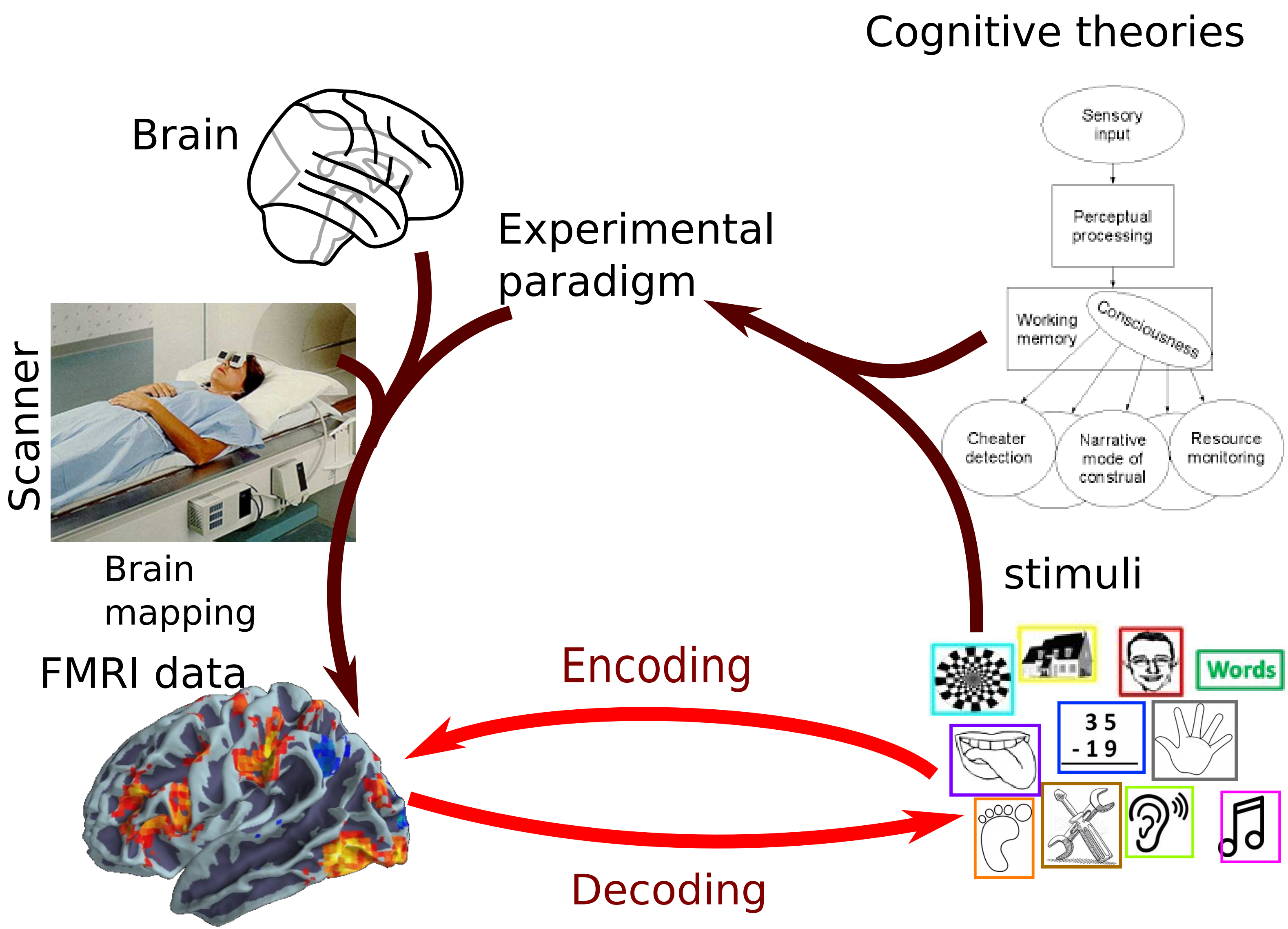
Cognitive neuroscience

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

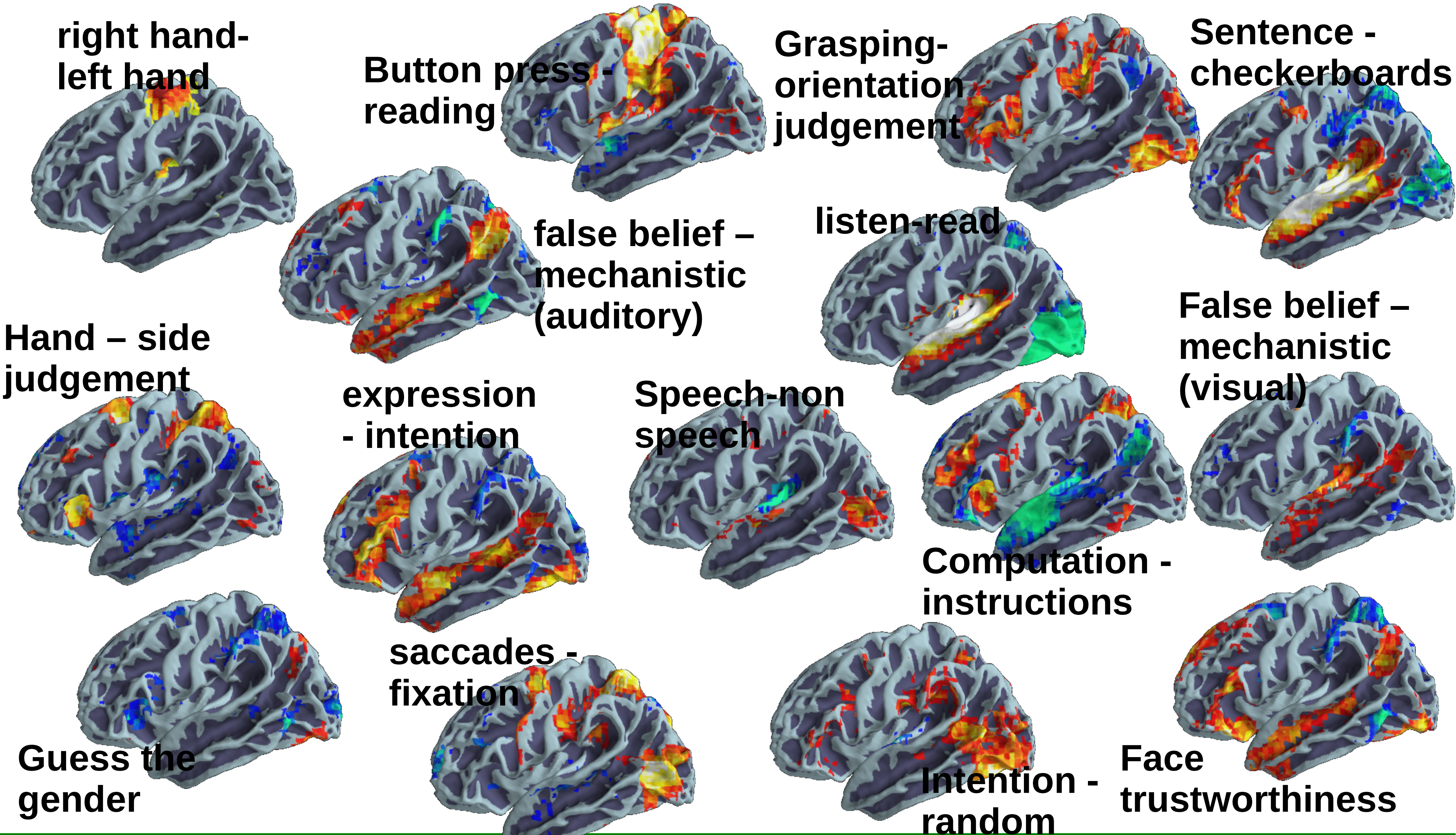
The brain, the mind and the scanner



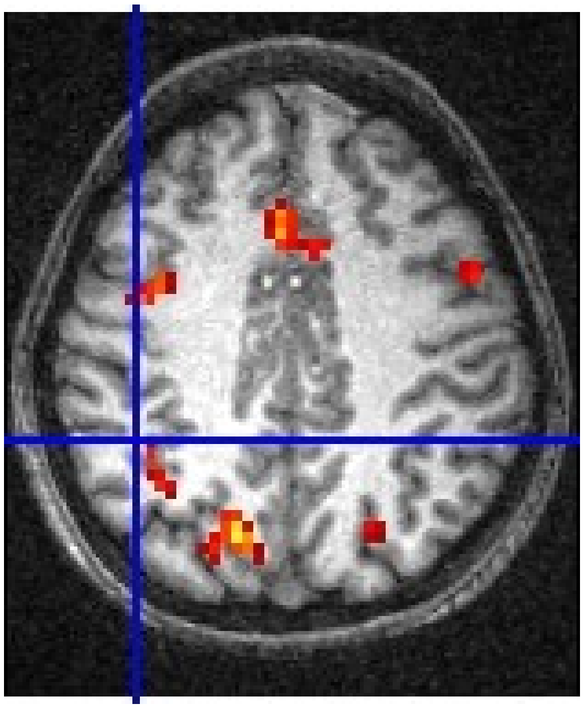
The brain, the mind and the scanner

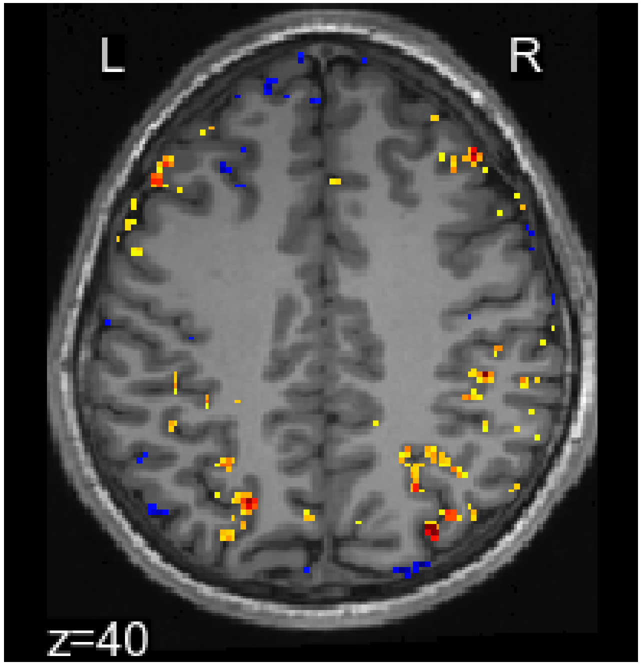


Mapping cognitive functions to brain activity



Resolution increases





2007:
3 mm

2014:
1.5 mm

2020:
0.5 mm ?

p = 50,000

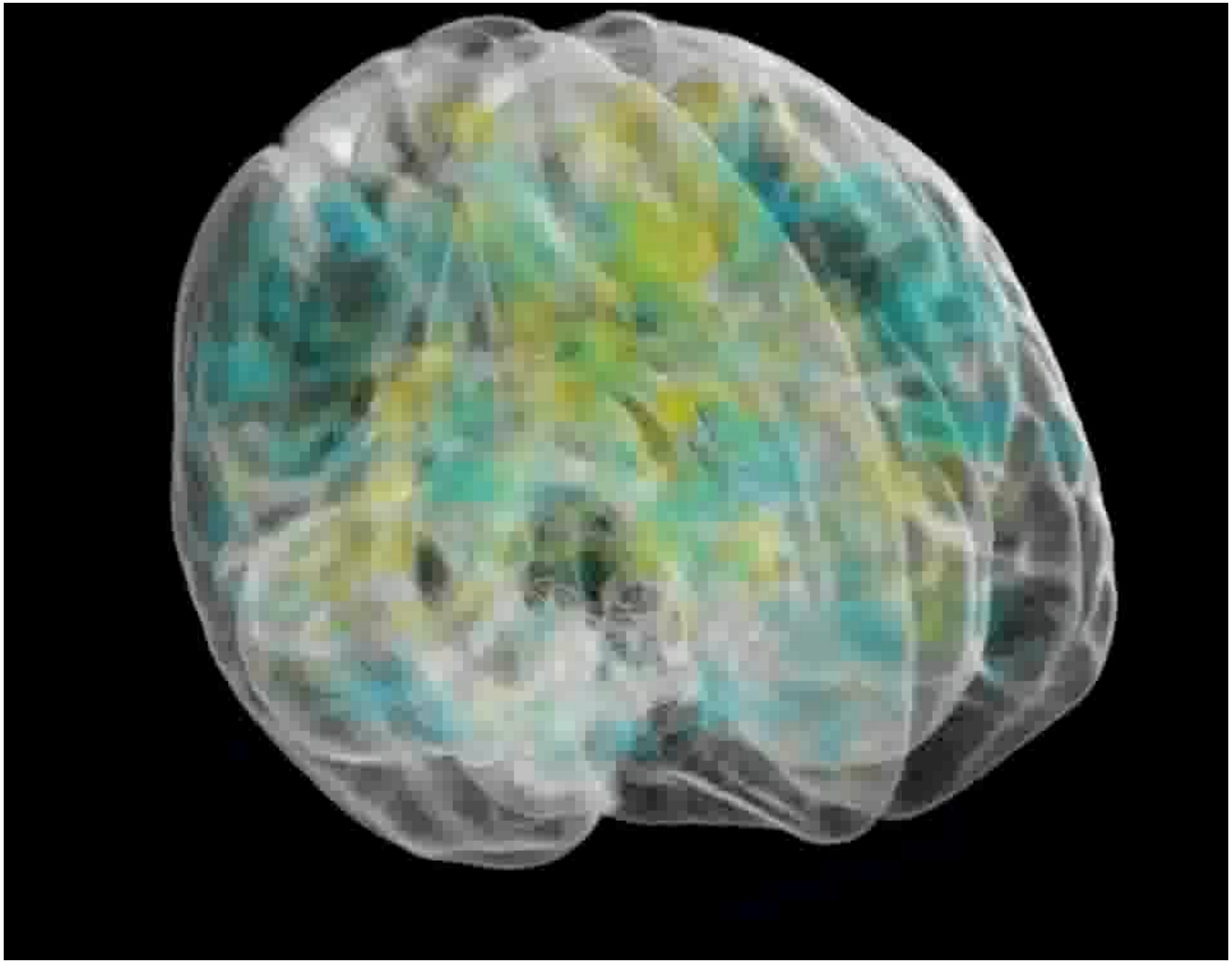
p = 400,000

p = 10^7

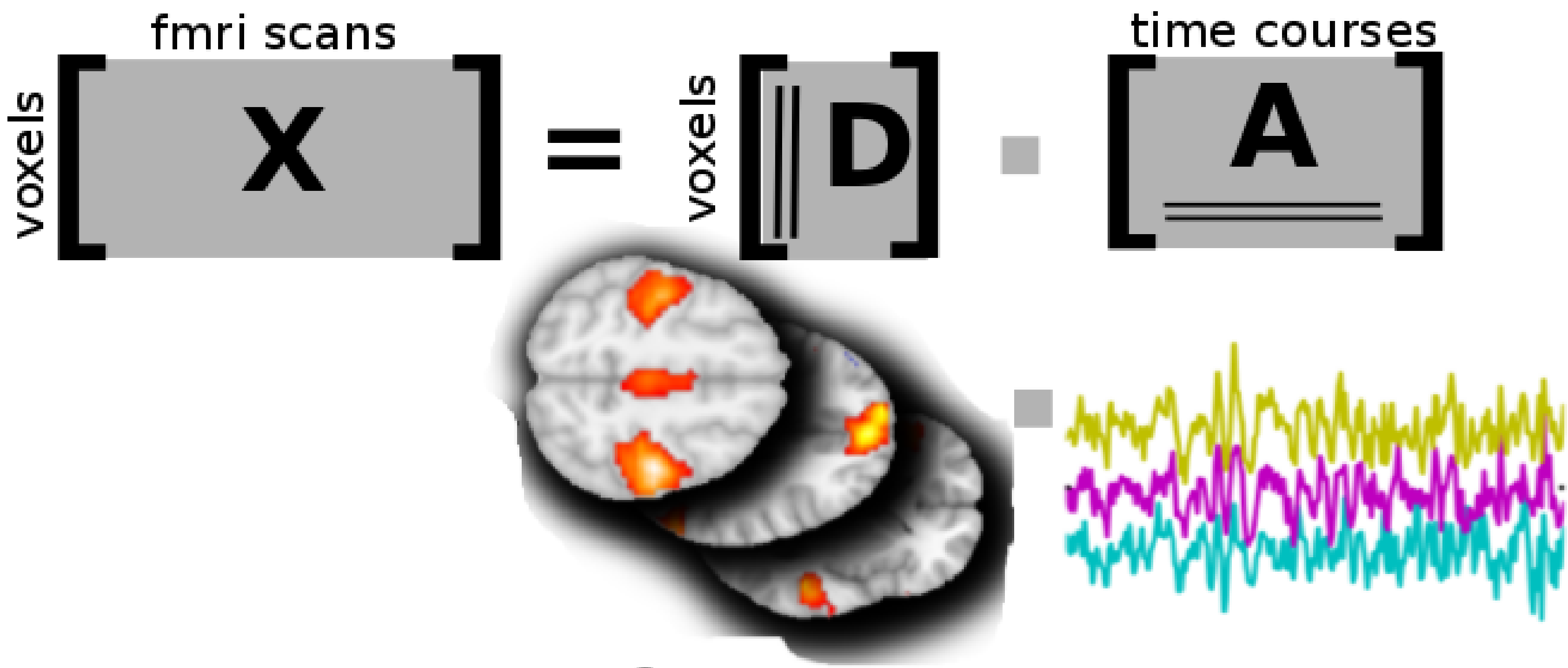
better estimators for large-scale brain imaging

- Massive online dictionary learning
- Dimension reduction for images
- Fast regularized ensembles of models
- Statistical inference for high-dimensional models

fMRI datasets are feature-rich



Discovering structure in fMRI



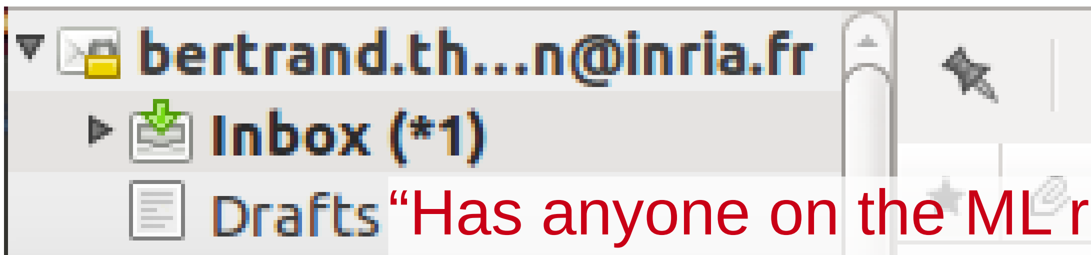
$$\underset{A,\,D}{\mathrm{argmin}} \; \|X\text{-}DA\|^2 + \lambda \|D\|_1$$

Can be captured by dictionary learning / sparse coding [Olshausen Nature 1996]

→ Use of sparse PCA

High-dimensional fMRI

- n = number of samples, 10^2 to 10^6
- p = number of voxels, 10^5-10^6



"Has anyone on the ML run group-wise analysis on the HCP resting state data, and if so what tools did you use?

I **am having memory issues when running more than 10 subjects** and I was wondering if anyone has a way of getting around the large memory requirements when concatenating in time."

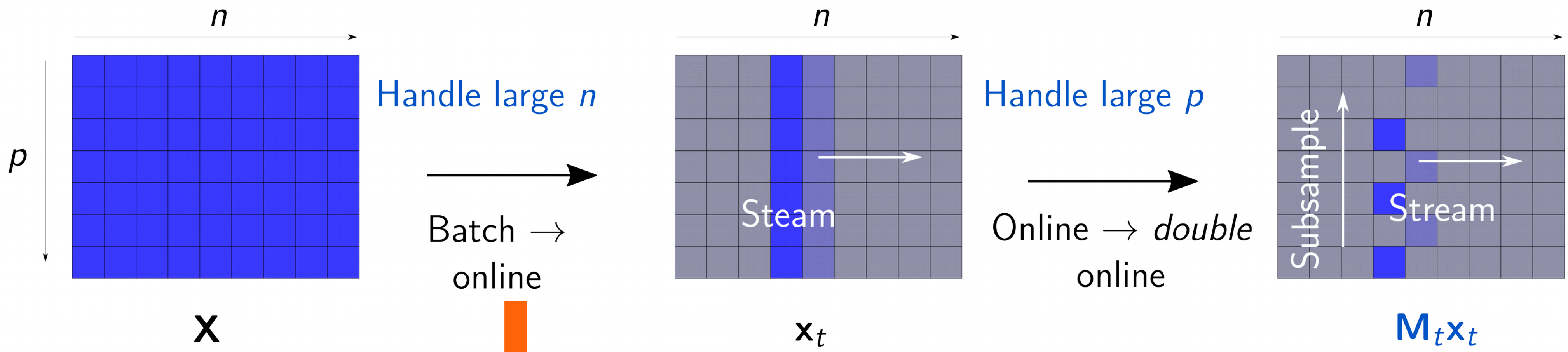
Huge ?

- Human Connectome project $n=2.10^6$, $p=2.10^5$, **2TB** of data
- Online dictionary learning [Mairal et al. ICML 2009]
- Constrained rather than penalized formulation
- How to go faster ?
  - Work on batches of images **and** voxels
    - Online method in both samples and feature dimensions

[Mensch et al. ICML 2016, IEEE TSP 2018]

Stochastic gradient approaches

http://amensch.fr/research/2016/06/10/modl.html



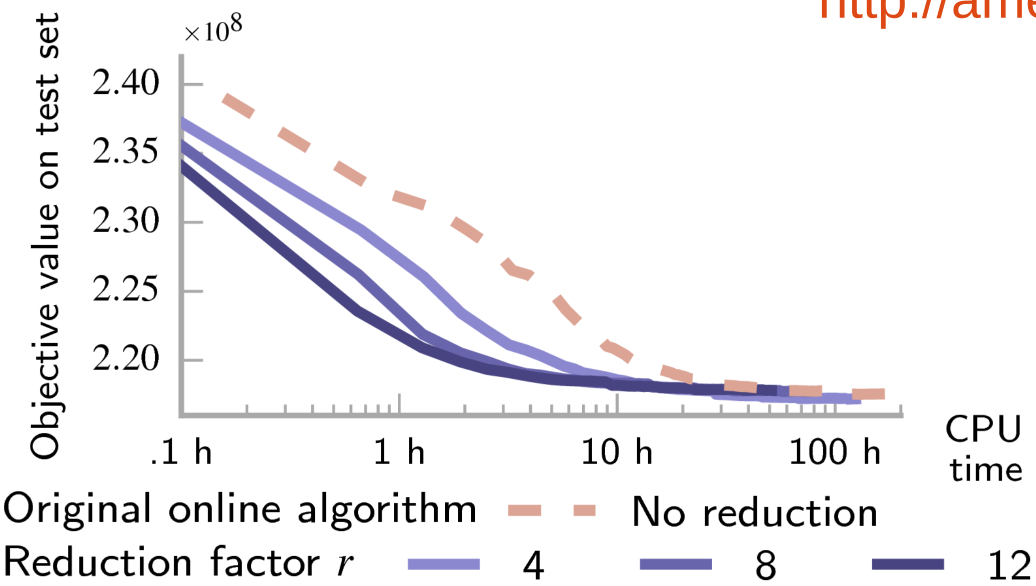
$$\boldsymbol{\alpha}_t(\mathbf{D}) = \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{k\times n}} \|\mathbf{x}_t - \mathbf{D}_{t-1}\boldsymbol{\alpha}_t\|_F^2 + \lambda\Omega(\boldsymbol{\alpha}_t)$$

$$\mathbf{D}_t = \operatorname{argmin}_{\mathbf{D}\in\mathcal{C}} \sum_{i=1}^{t} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_F^2$$

$$\boldsymbol{\alpha}_t(\mathbf{D}) = \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{k\times n}} \|\mathbf{M}_t(\mathbf{x}_t - \mathbf{D}_{t-1}\boldsymbol{\alpha}_t)\|_F^2 + \lambda\frac{s}{p}\Omega(\boldsymbol{\alpha})$$

Stochastic gradient approaches

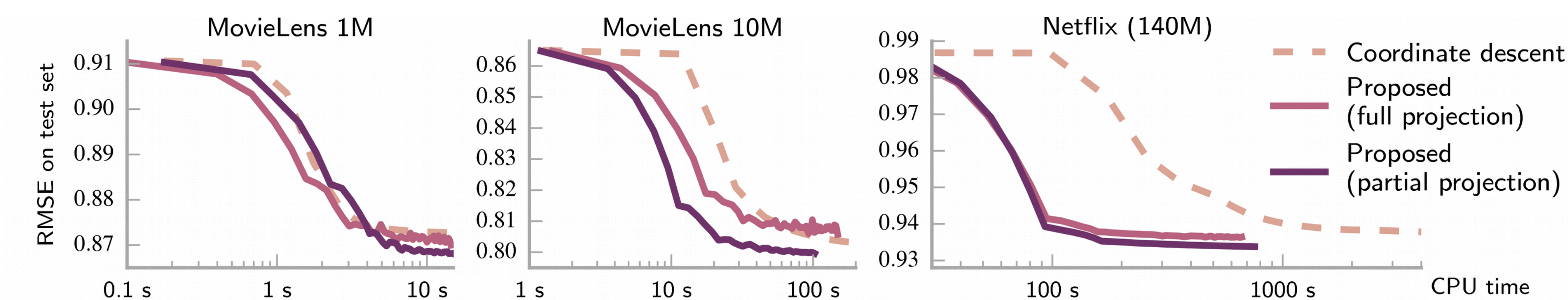
http://amensch.fr/research/2016/06/10/modl.html



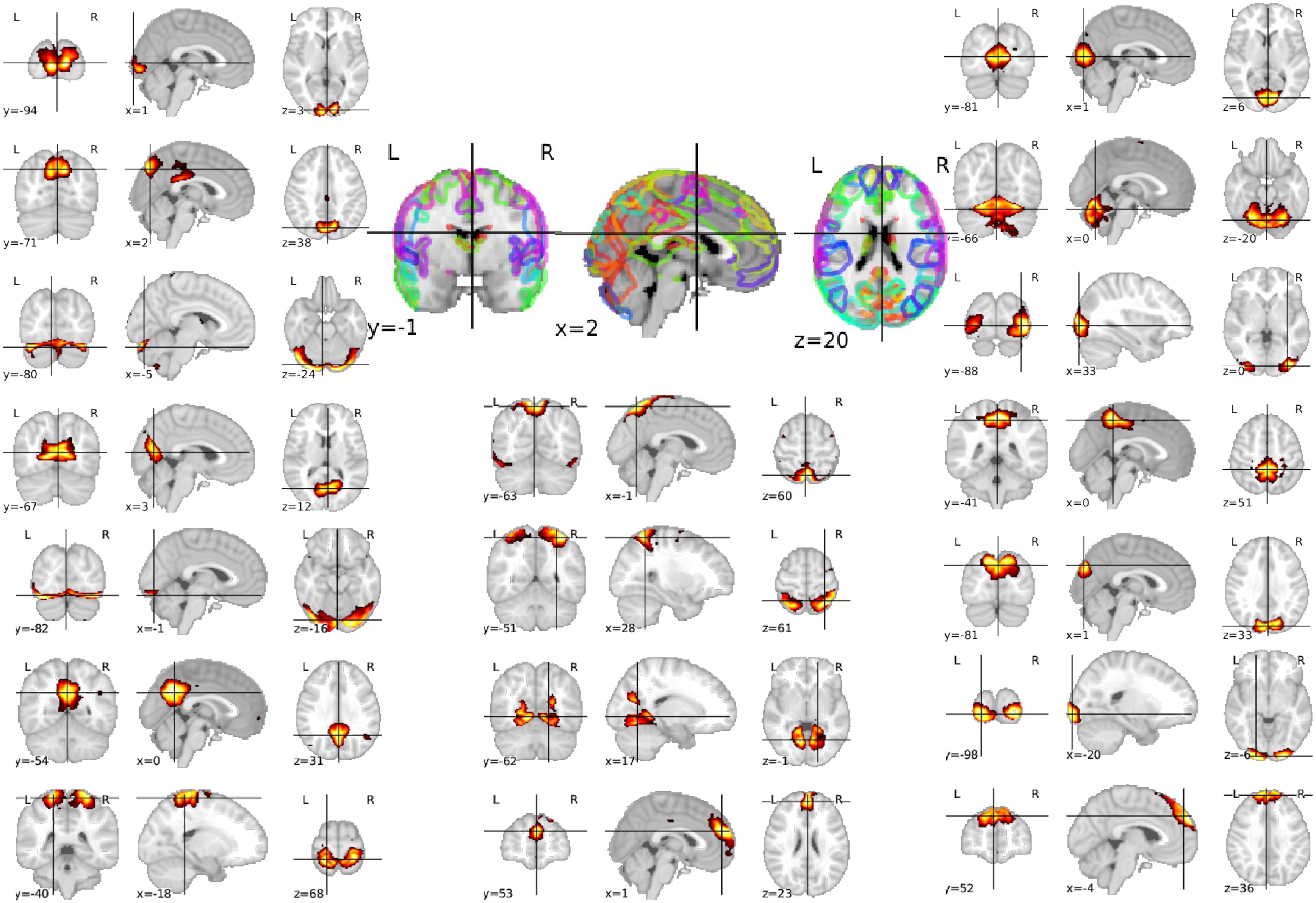
10-fold gain in CPU time without loss in accuracy

[Mensch et al. ICML 2016, IEEE TSP 2018]

Can be used for recommender systems



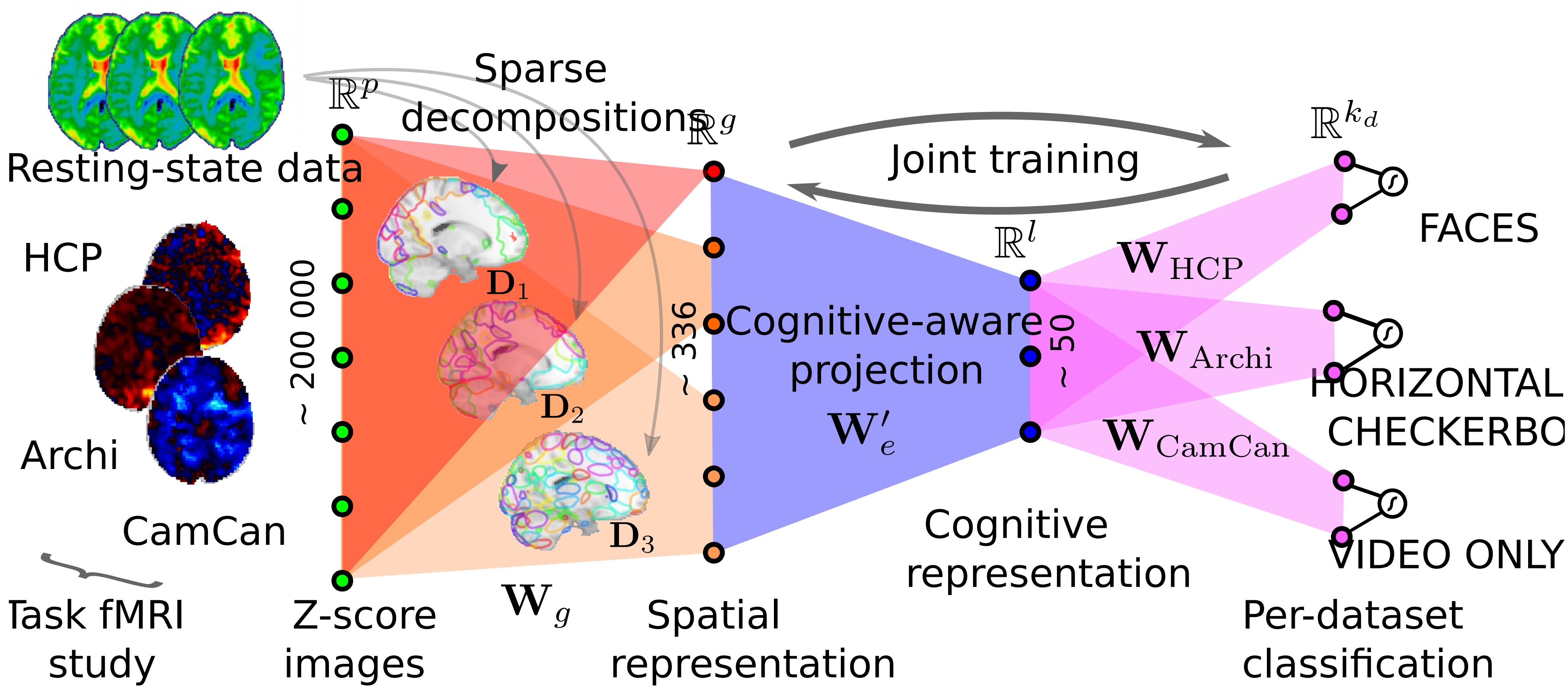
Brain atlases



[Mensch et al. ICML 2016 IEEE TSP 2018]

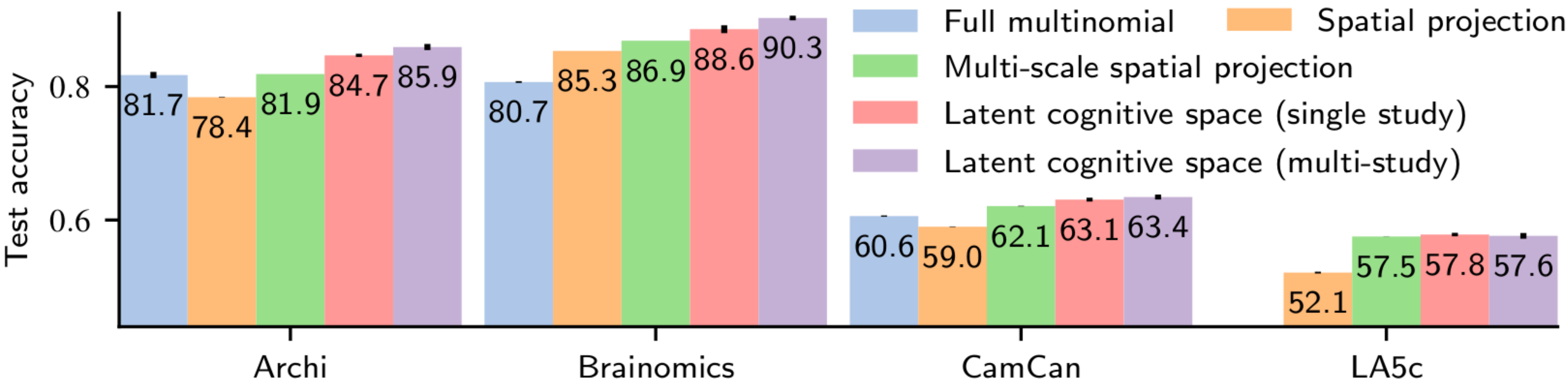
Leveraging rest data for brain decoding

Different datasets share some common patterns



Different datasets share some common representations
[Bzdok et al. Plos Comp Biol 2016, Mensch et al NIPS 2017]

Advantage of large-scale analysis



Information transferred from large datasets (HCP) to smaller ones increases classification accuracy
[Mensch et al NIPS 2017]

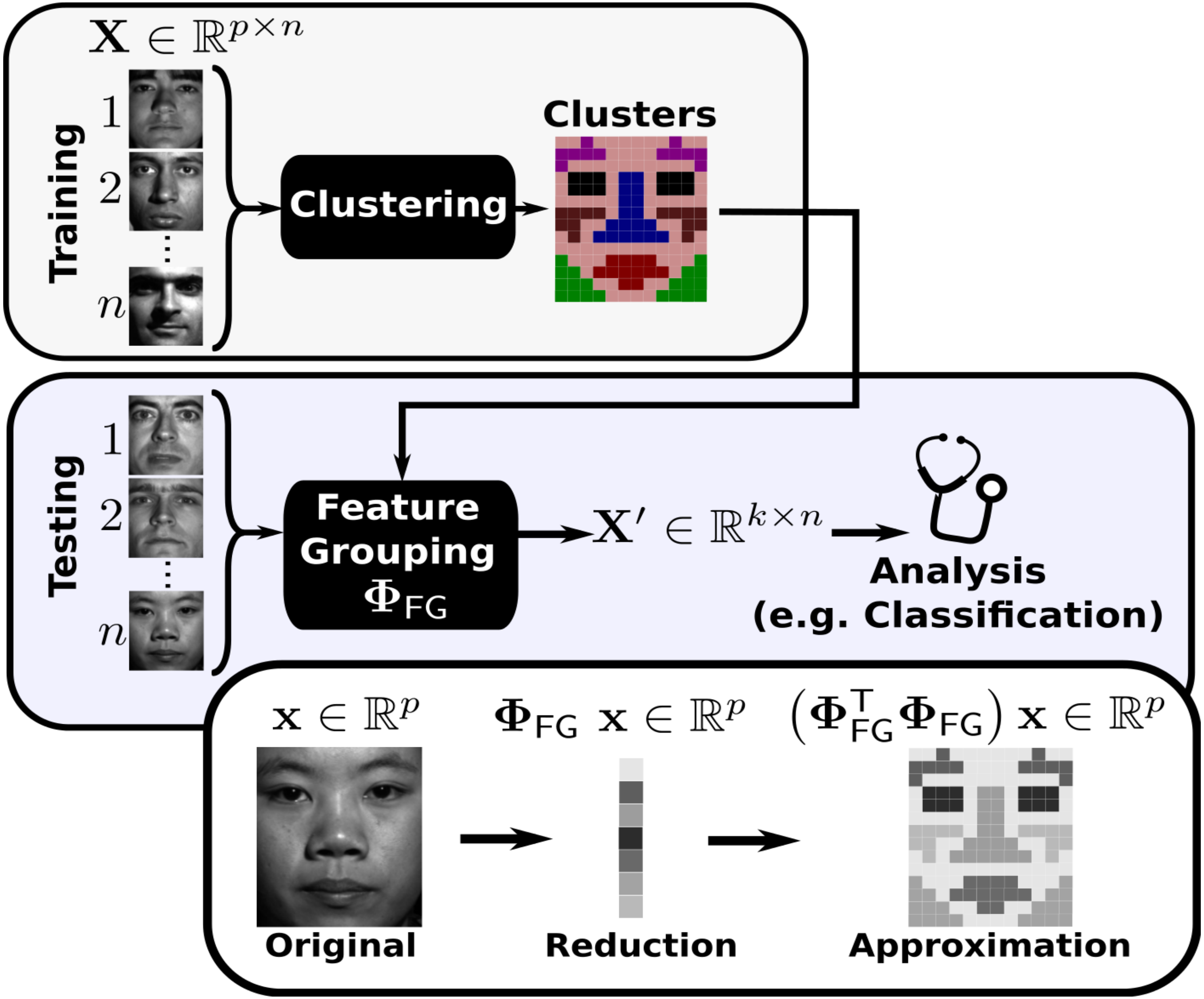
Outline

- Massive online dictionary learning
- 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: $\|\mathbf{x}\|^2 - L^2 \sum_{q=1}^{k} |\mathcal{C}_q|^3 \leq \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|^2 \leq \|\mathbf{x}\|^2$
- And $\|\mathbf{x}\|^2 - p\left(L\frac{p}{k}\right)^2 \leq \mathbb{E}_{|\mathcal{P}|} \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|^2 \leq \|\mathbf{x}\|^2$
- 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

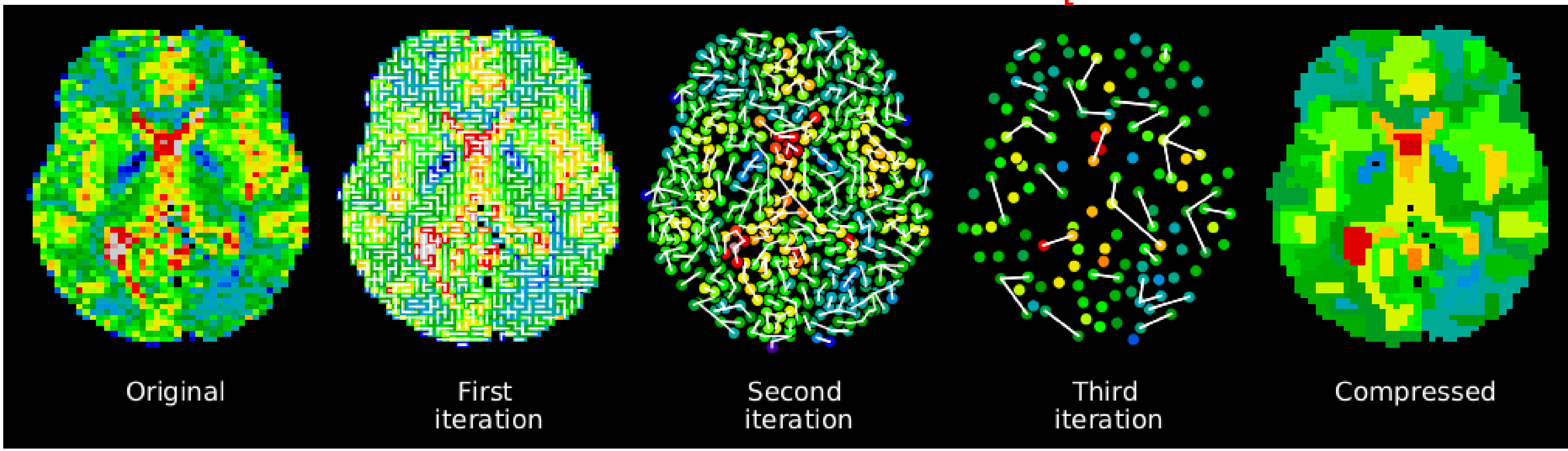
$$\mathrm{MSE}_{\mathrm{approx}} \leq \mathrm{MSE}_{\mathrm{orig}} \qquad L^2 \leq \frac{(p-k)}{\sum_{q=1}^{k} |\mathcal{C}_q| \, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2} \sigma^2$$

- Equal-size clusters

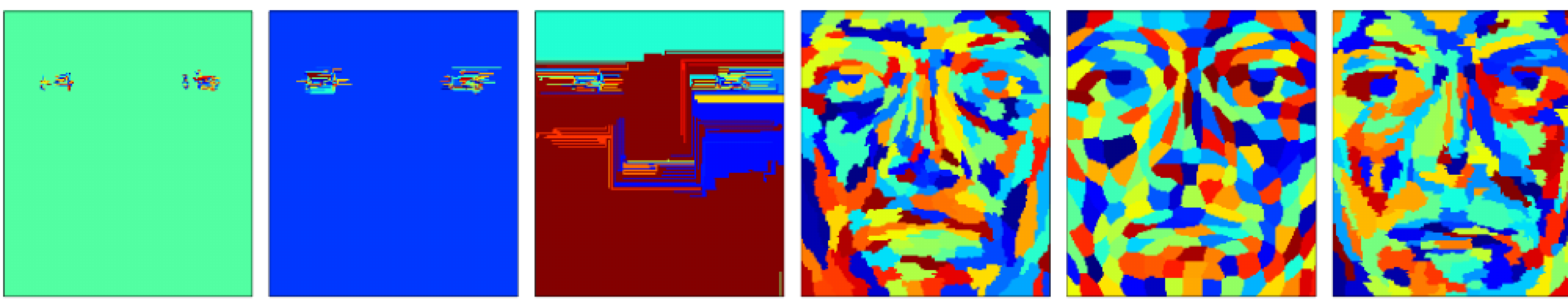
$$\mathrm{MSE}_{\mathrm{approx}} \leq p \left(\frac{L}{k}\right)^2 + \frac{k}{p} \mathrm{MSE}_{\mathrm{orig}} = O\left(\max\left\{\frac{p}{k^2}, \frac{k}{p}\right\}\right)$$

Recursive nearest neighbor

[Thirion et al. Stamlins 2015]

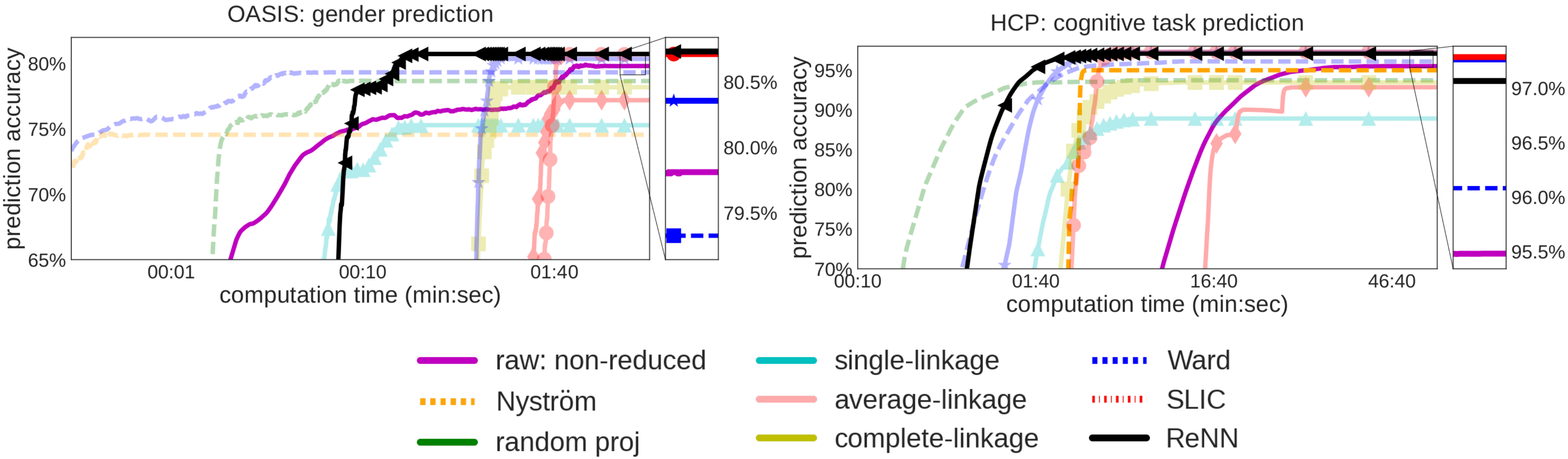


Based on local decisions = fast (linear time) – avoid percolation



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

Effect on data analysis tasks

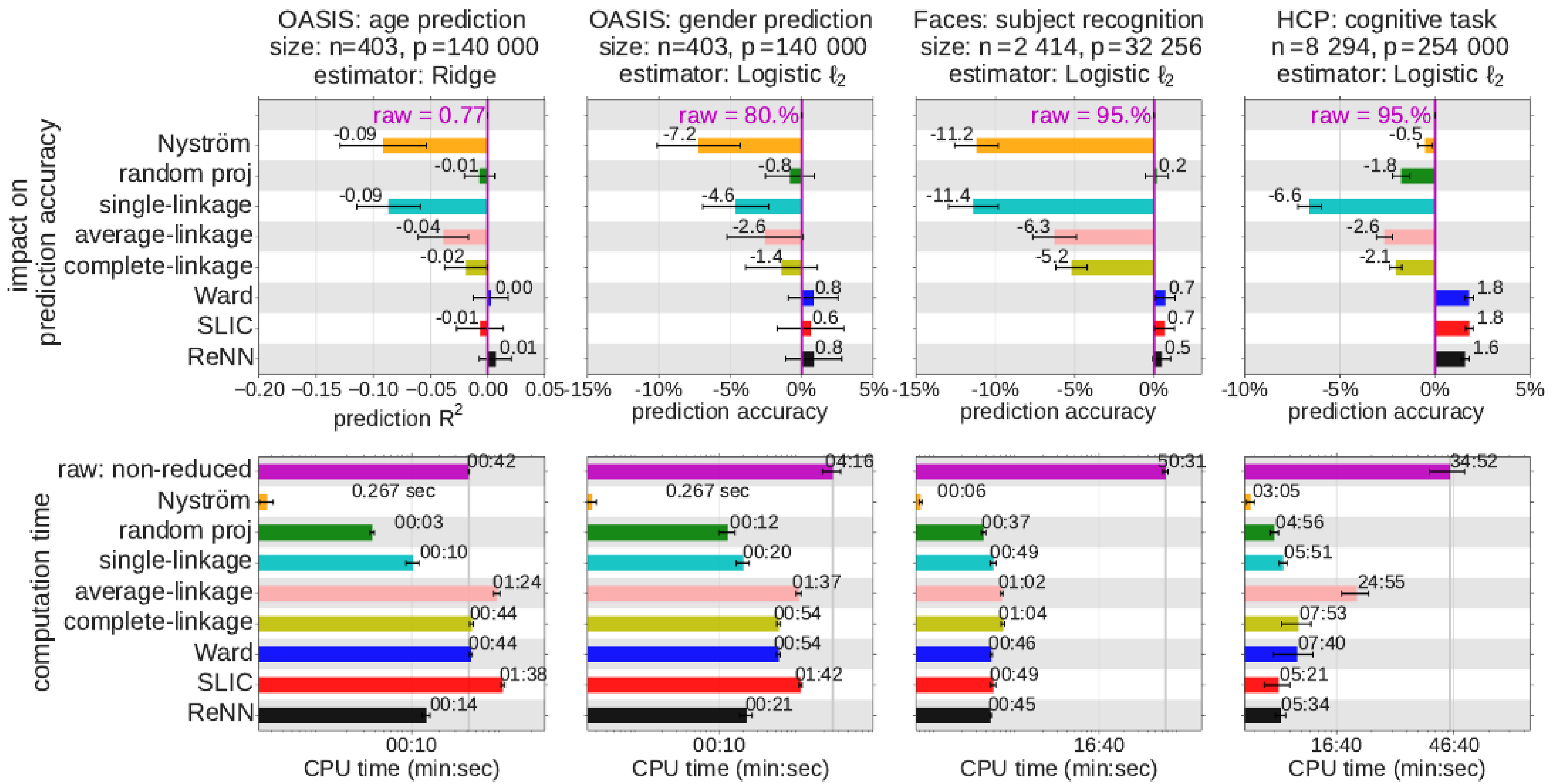


Impressive speed-up and increased accuracy with respect to non-compressed representation

- Clustering has a denoising effect

[Hoyos Idrobo IEEE PAMI under revision]

More results

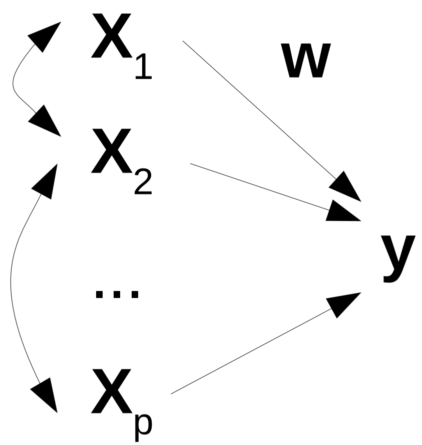


[Hoyos Idrobo IEEE PAMI under revision]

Outline

- Massive online dictionary learning
- 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 $j \in [p]$

Penalized linear regression

Minimize the empirical regularized risk

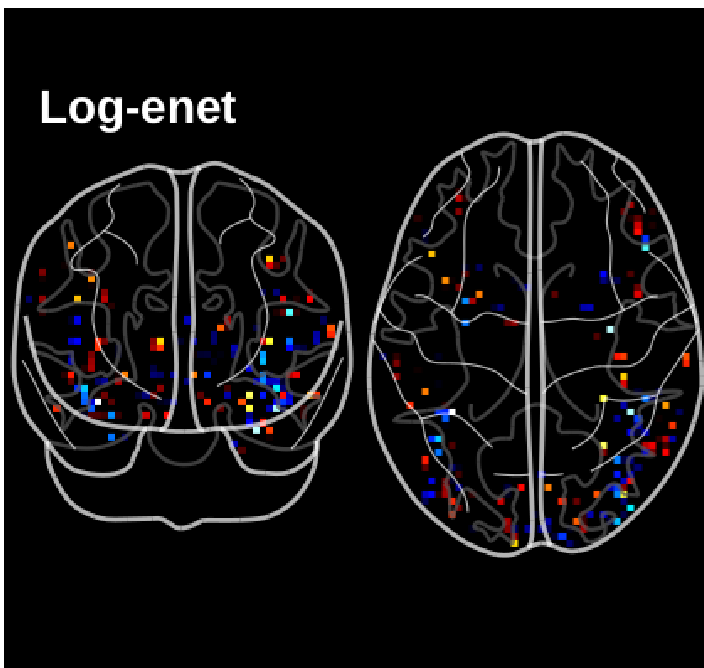
$$\hat{\mathbf{w}} = \underset{w}{\operatorname{argmin}}\{\underbrace{\mathcal{L}(\mathbf{X}, \mathbf{y}; \mathbf{w})}_{\text{Data fidelity}} + \underbrace{\lambda\Omega(\mathbf{w})}_{\text{Regularizer}}\}$$

> convex optimization
> set hyperparameters by cross-validation

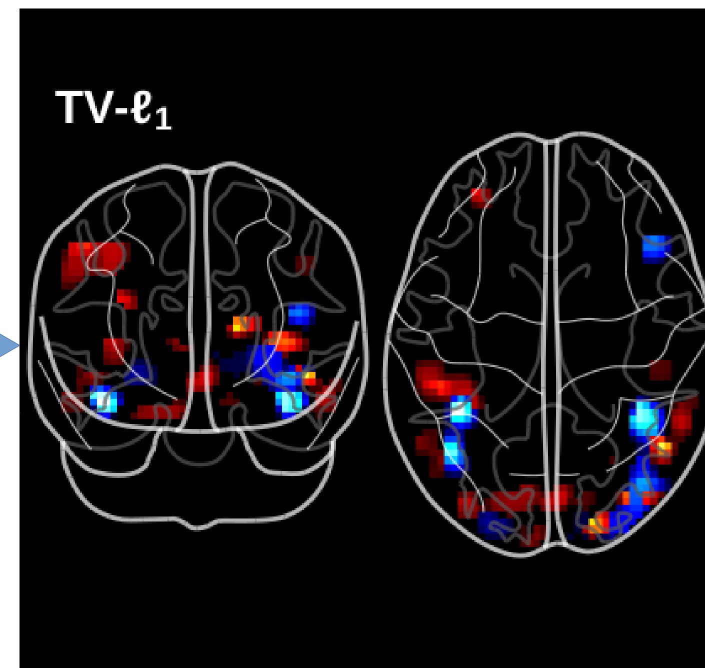
$\lambda\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$	Ridge (shrinkage)
$\lambda\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$	Lasso (very sparse)
$\lambda\Omega(\mathbf{w}) = \lambda\left(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2^2\right)$	Elastic net (sparsity + grouping)
$\lambda\Omega(\mathbf{w}) = \lambda\left(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\nabla\mathbf{w}\|_2^2\right)$	Smooth lasso (sparsity + smoothness)
$\lambda\Omega(\mathbf{w}) = \lambda\left(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\nabla\mathbf{w}\|_{2,1}\right)$	Total variation (piecewise sparsity)

Structure-inducing priors

- Large p → redundancy, latent structure
- Brain imaging: spatial regularity → small total variation



HCP dataset:
“shape” versus
“face” contrast
across subjects



[Michel et al TMI 2011, Gramfort et al. 2013 Eickenberg et al. MICCAI 2015, Dohmatob et al. PRNI 2014, 2015]

Optimizing TV takes time

Data: w_0

$ISTA \leftarrow False, v_1 \leftarrow w_0, k \leftarrow 0, t_1 \leftarrow 1, dgtol \leftarrow 0.1;$

while *not converged* **do**

  $k \leftarrow k + 1;$

  $w_k \leftarrow \mathrm{prox}_{G/L}(v_k - (1/L)\nabla F(v_k), dgtol);$

  if $\mathcal{L}(w_k) > \mathcal{L}(w_{k-1})$ **then**

    $w_k \leftarrow w_{k-1};$

    $v_k \leftarrow w_{k-1};$

    if *ISTA* **then**

      $dgtol \leftarrow dgtol/2;$

      while

      $\mathcal{L}(\mathrm{prox}_{G/L}(v_k - (1/L)\nabla_w F(v_k), dgtol)) > \mathcal{L}(w$

        $dgtol \leftarrow dgtol/2$

    $ISTA \leftarrow True;$

  else

    if *ISTA* **then**

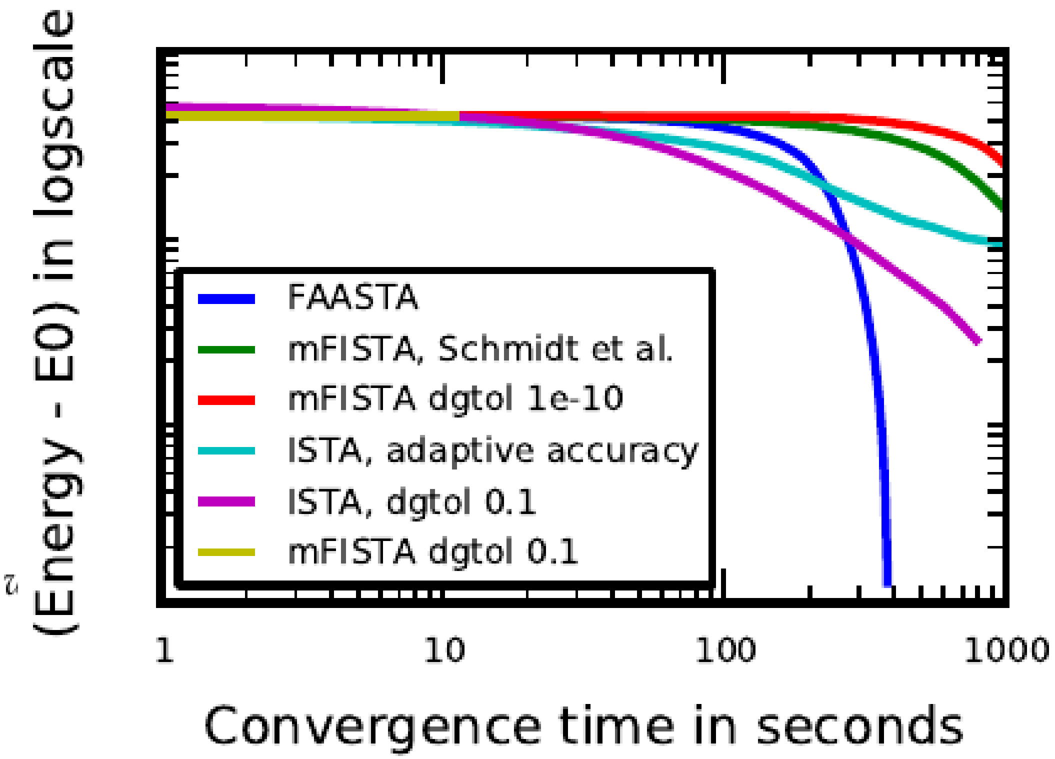
      $v_k \leftarrow w_k$

    else

      $t_k \leftarrow \frac{1+\sqrt{1+4t_{k-1}^2}}{2};$

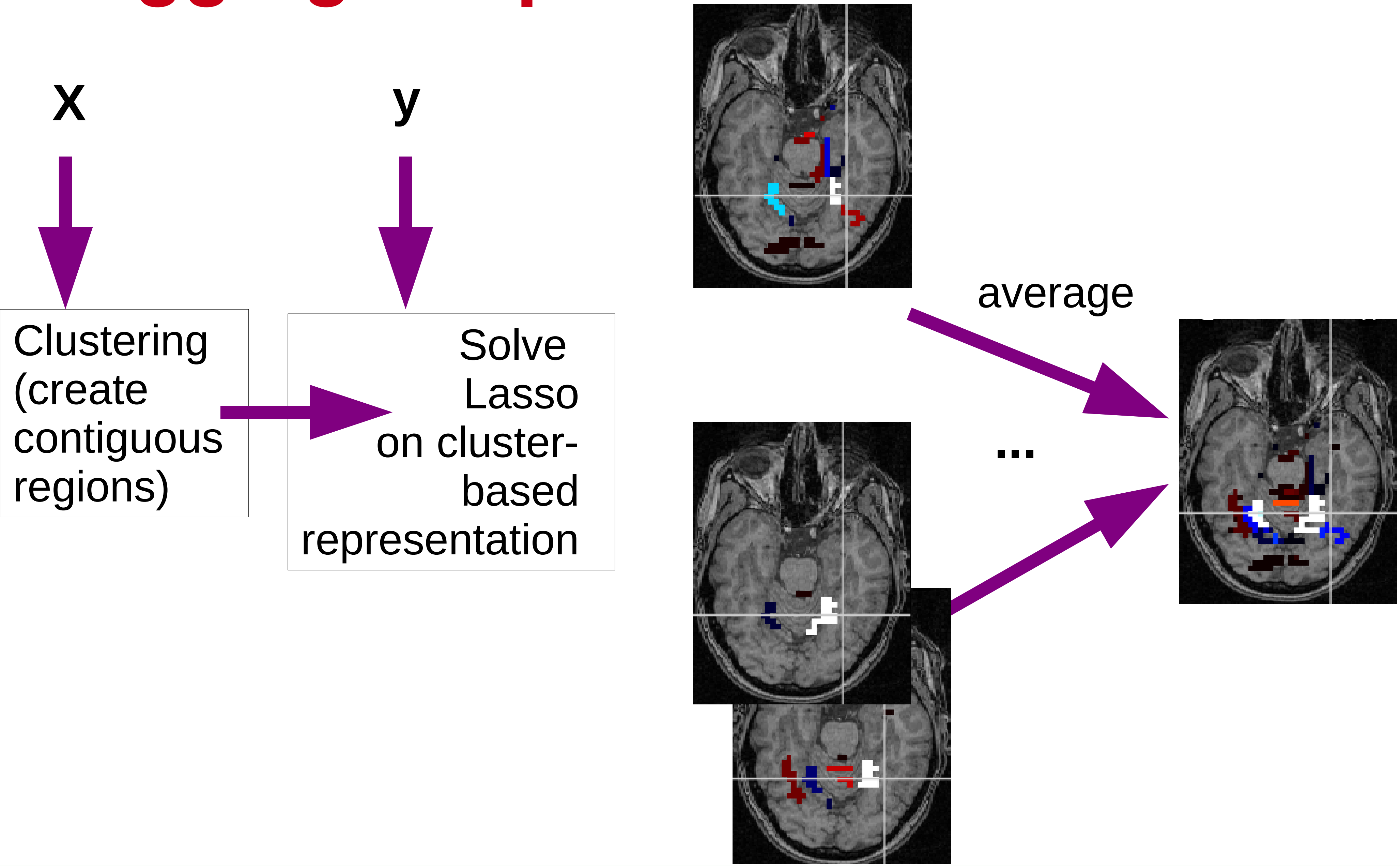
      $v_k \leftarrow w_k + \frac{t_{k-1}-1}{t_k}(w_k - w_{k-1});$

    $ISTA \leftarrow False$



[Varoquaux et al. GRETSI 2015]

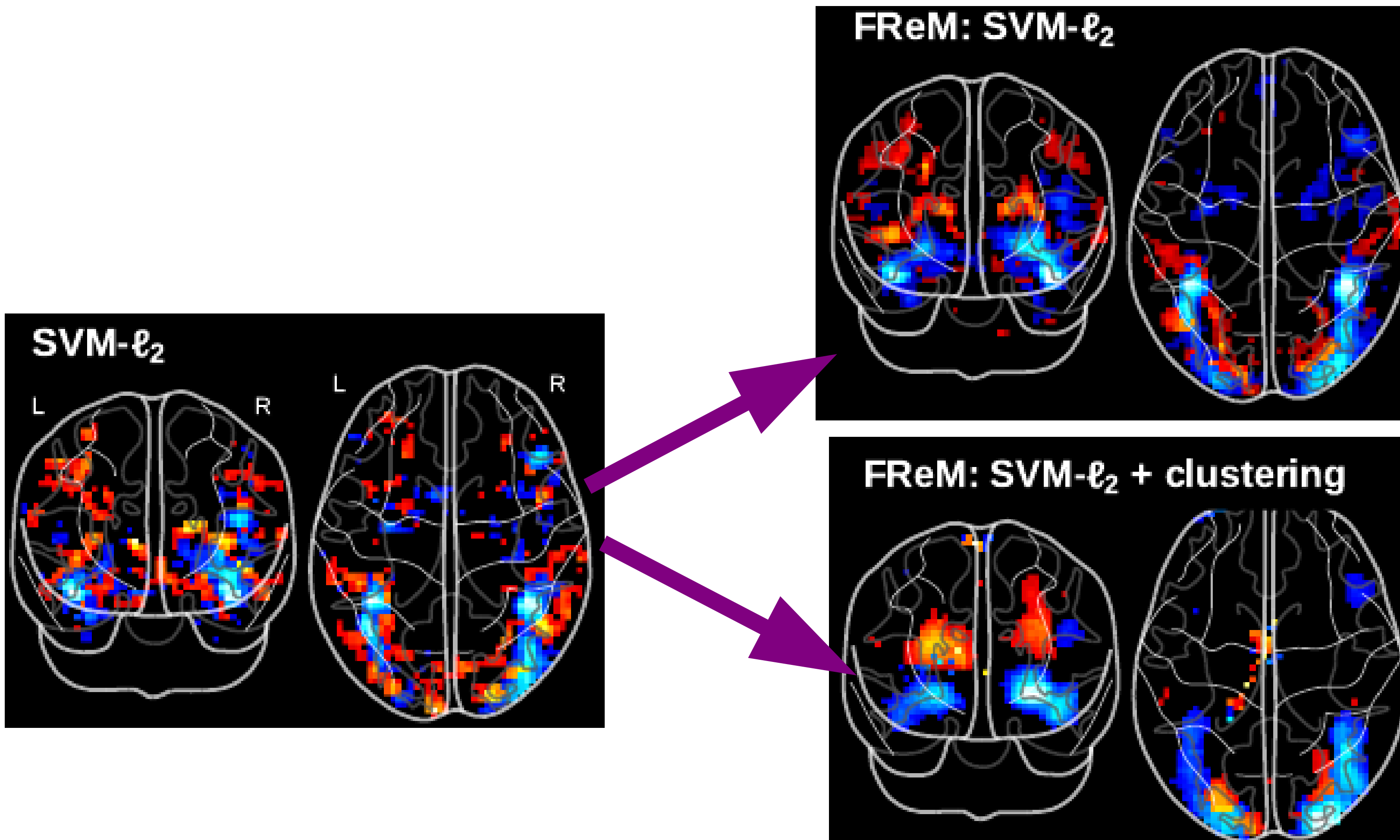
Bagging of sparse clustered models



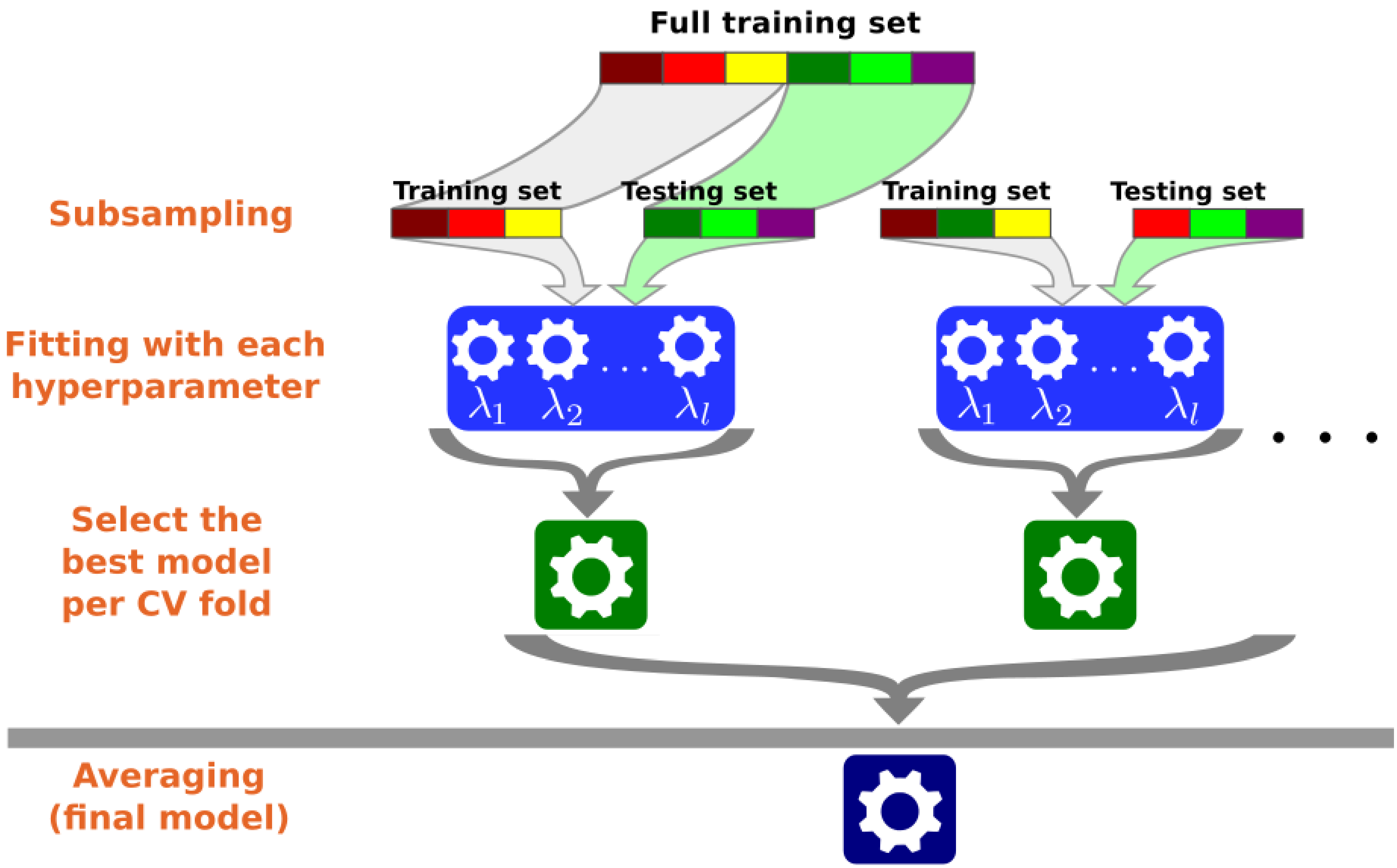
Computationally efficient structure

"fast regularized ensembles of models"

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

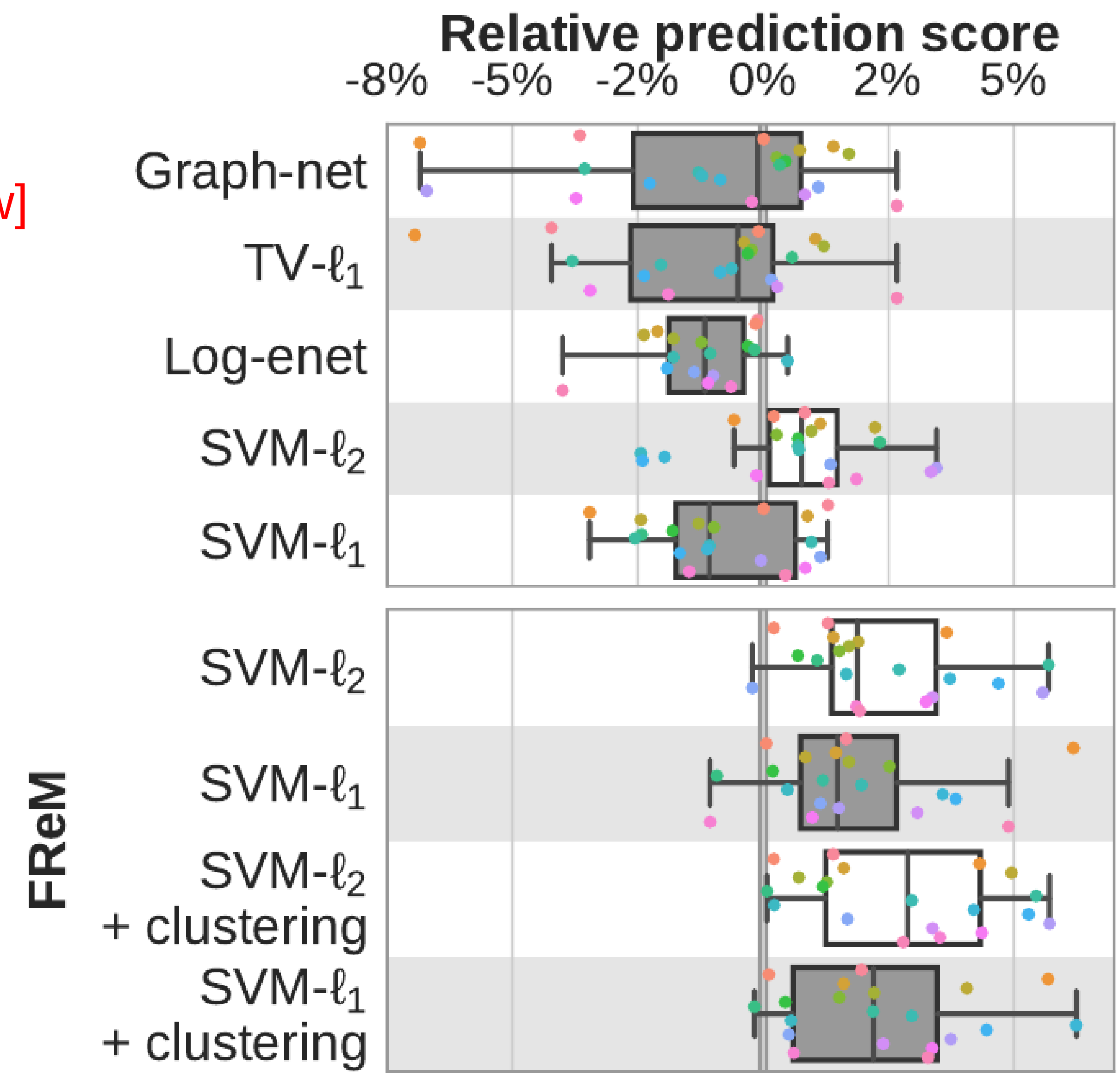


Computationally efficient structure



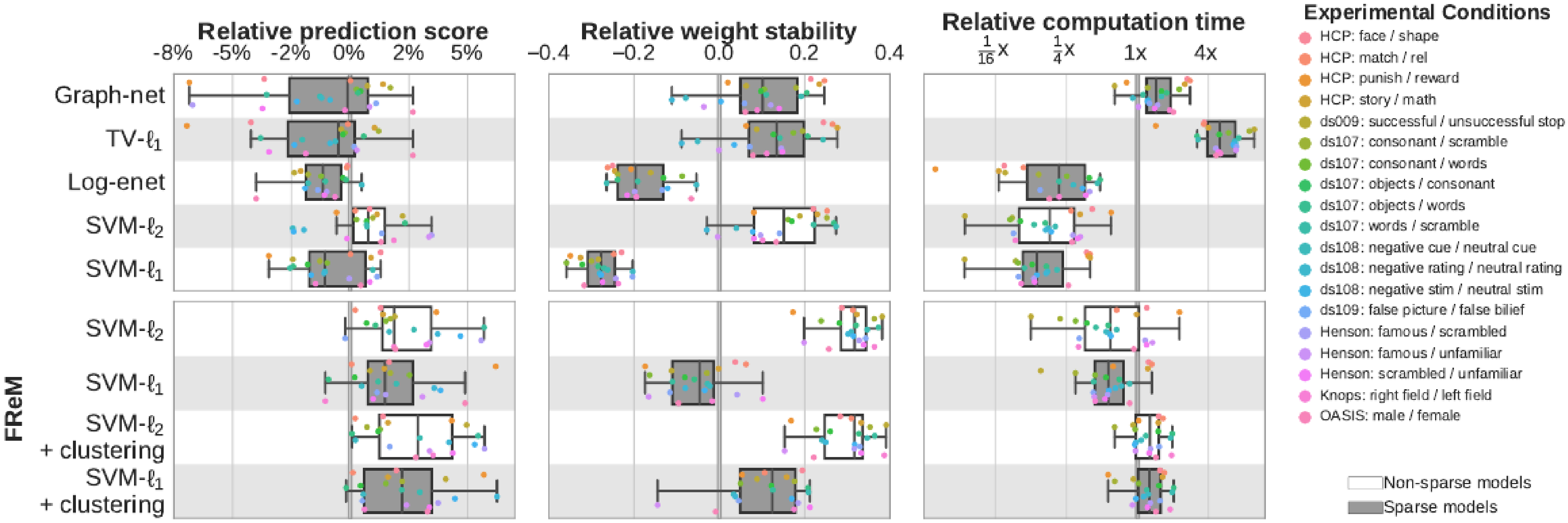
Computationally efficient structure

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



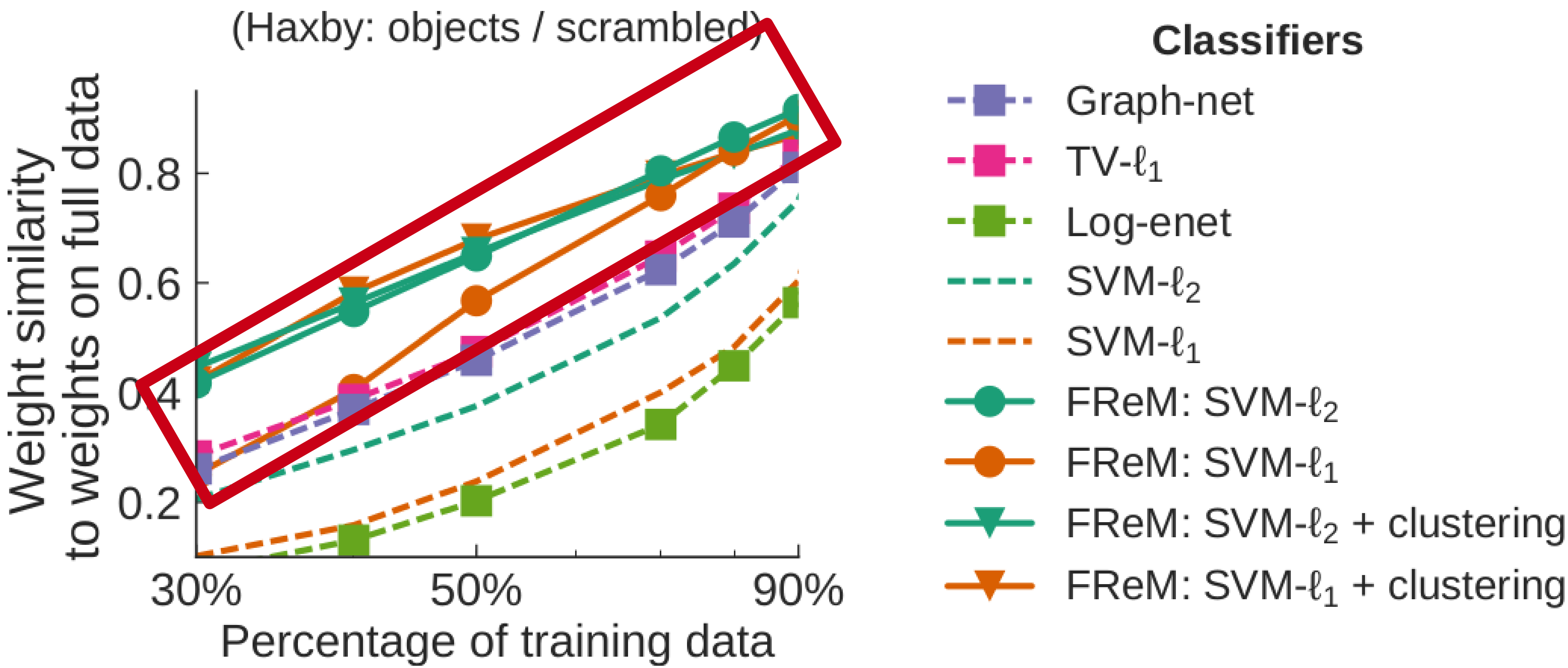
"fast regularized ensembles of models"

Computationally efficient structure



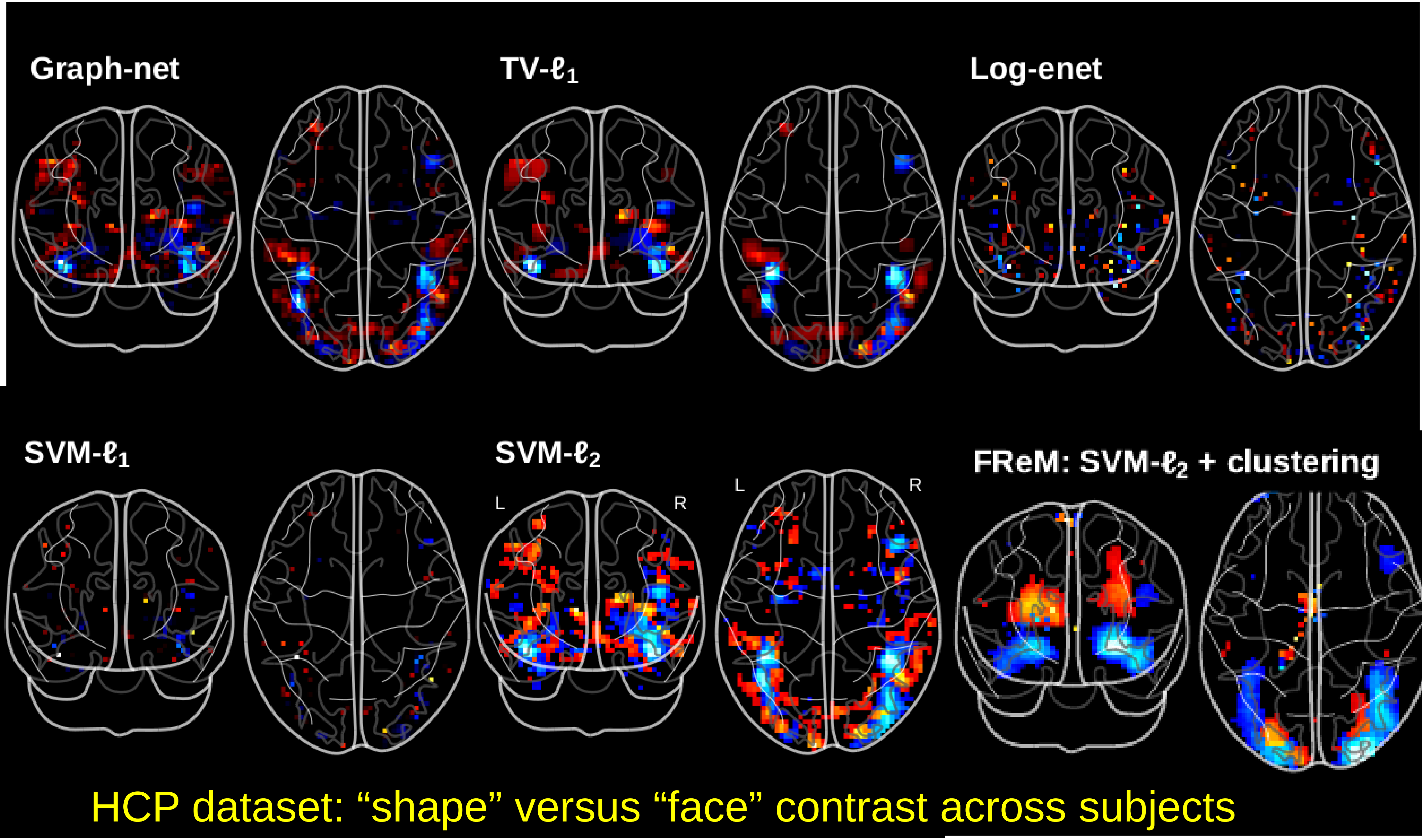
[Hoyos Idrobo et al PRNI 2015, Neuroimage 2017, PAMI under review]

Computationally efficient structure



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

Benchmark


Graph-net
TV-ℓ1
Log-enet
SVM-ℓ1
SVM-ℓ2
L
R
L
R
FReM: SVM-ℓ2 + clustering
HCP dataset: “shape” versus “face” contrast across subjects

Outline

- Massive online dictionary learning
- 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
  - Desparsified Lasso

- Adaptation to brain imaging ($p \gg n$)

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:

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

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

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

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]):

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

Desparsified Lasso: which λ

[zhang & zhang 2014 Series B Stat Meth]

$$\eta_j(\lambda) = \max_{k \neq j} |\boldsymbol{x}_k^T \boldsymbol{z}_j(\lambda)| / \|\boldsymbol{z}_j(\lambda)\|_2,$$

$$\tau_j(\lambda) = \|\boldsymbol{z}_j(\lambda)\|_2 / |\boldsymbol{x}_j^T \boldsymbol{z}_j(\lambda)|,$$

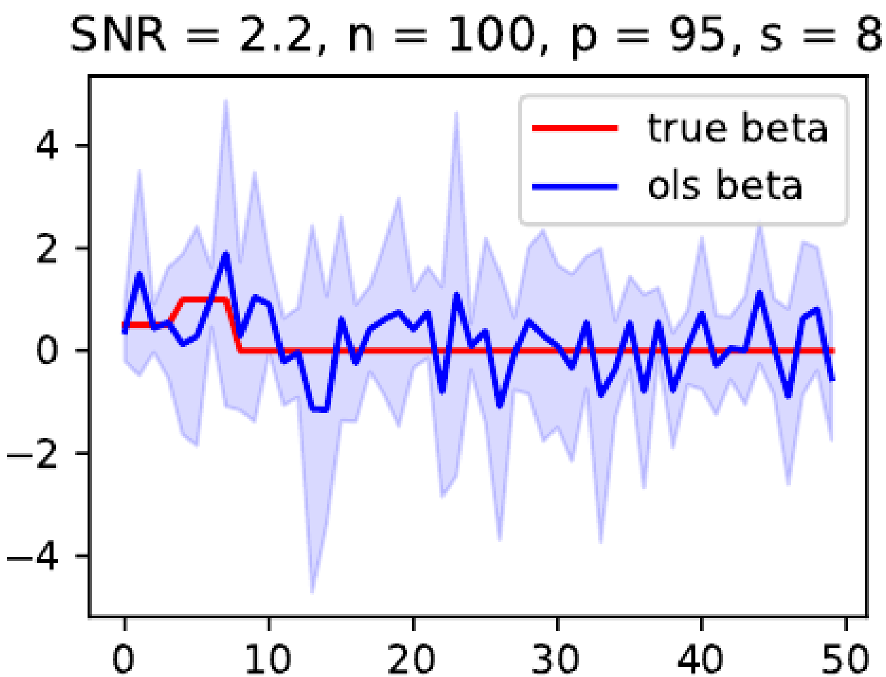
- For each j,
  - η_j should be as small as possible
    - keep λ small
  - τ_j should be as high as possible
    - λ not too small

Evaluating η and τ for many λ's is expensive

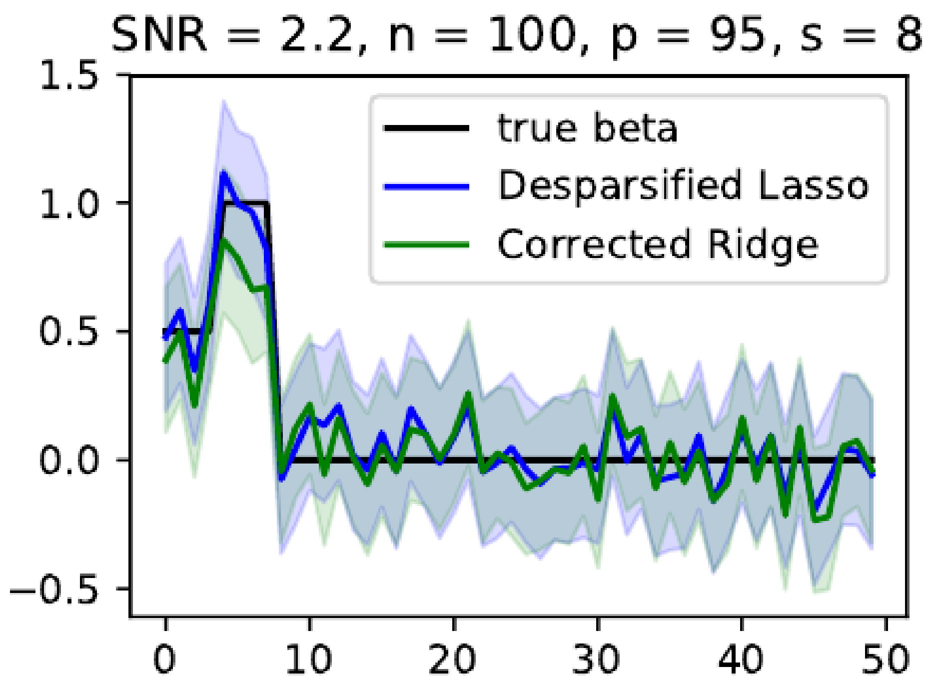
→ We choose $\lambda = .03\ \lambda_{max}$

Preliminary assessment

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

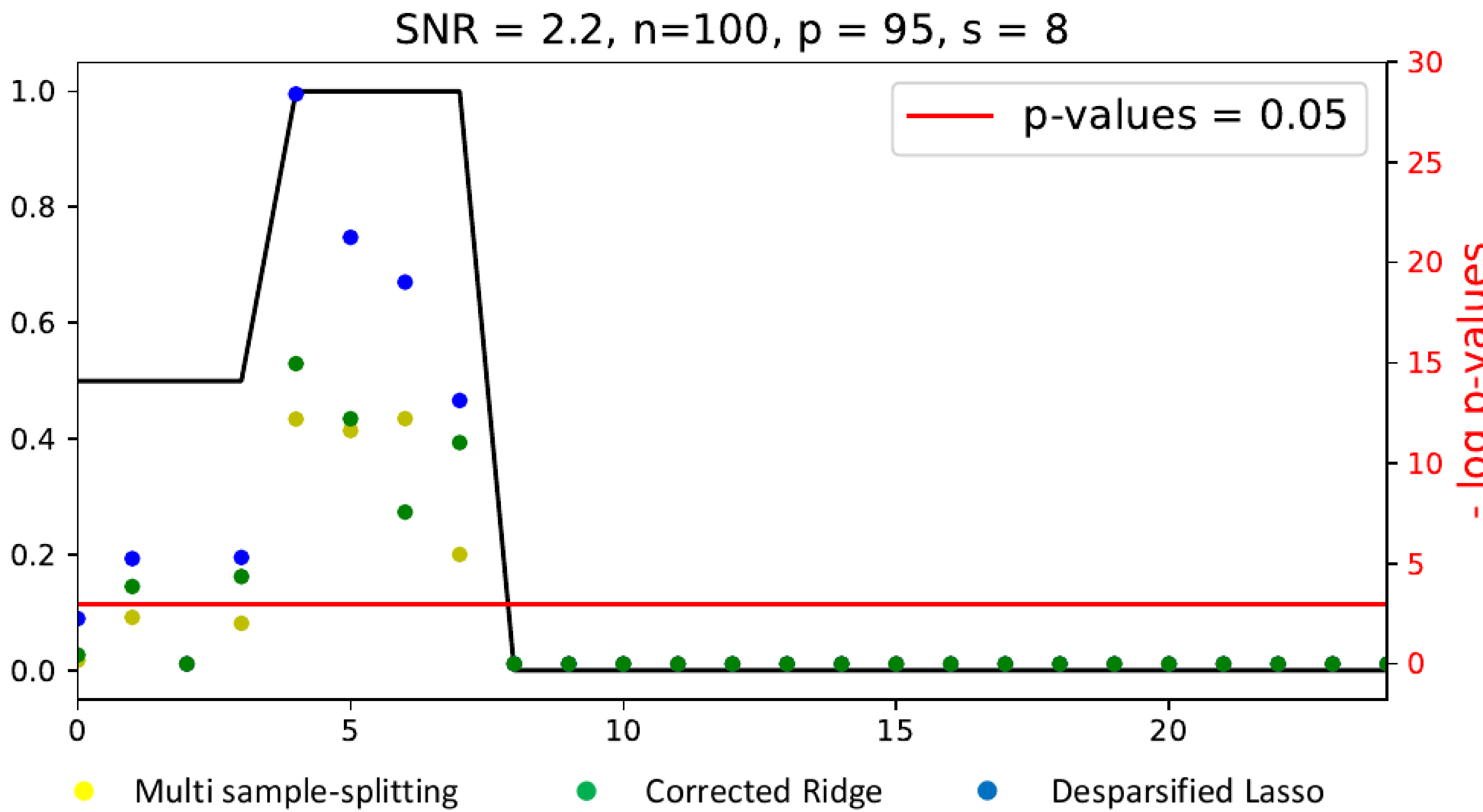


OLS regression when $p \approx n$

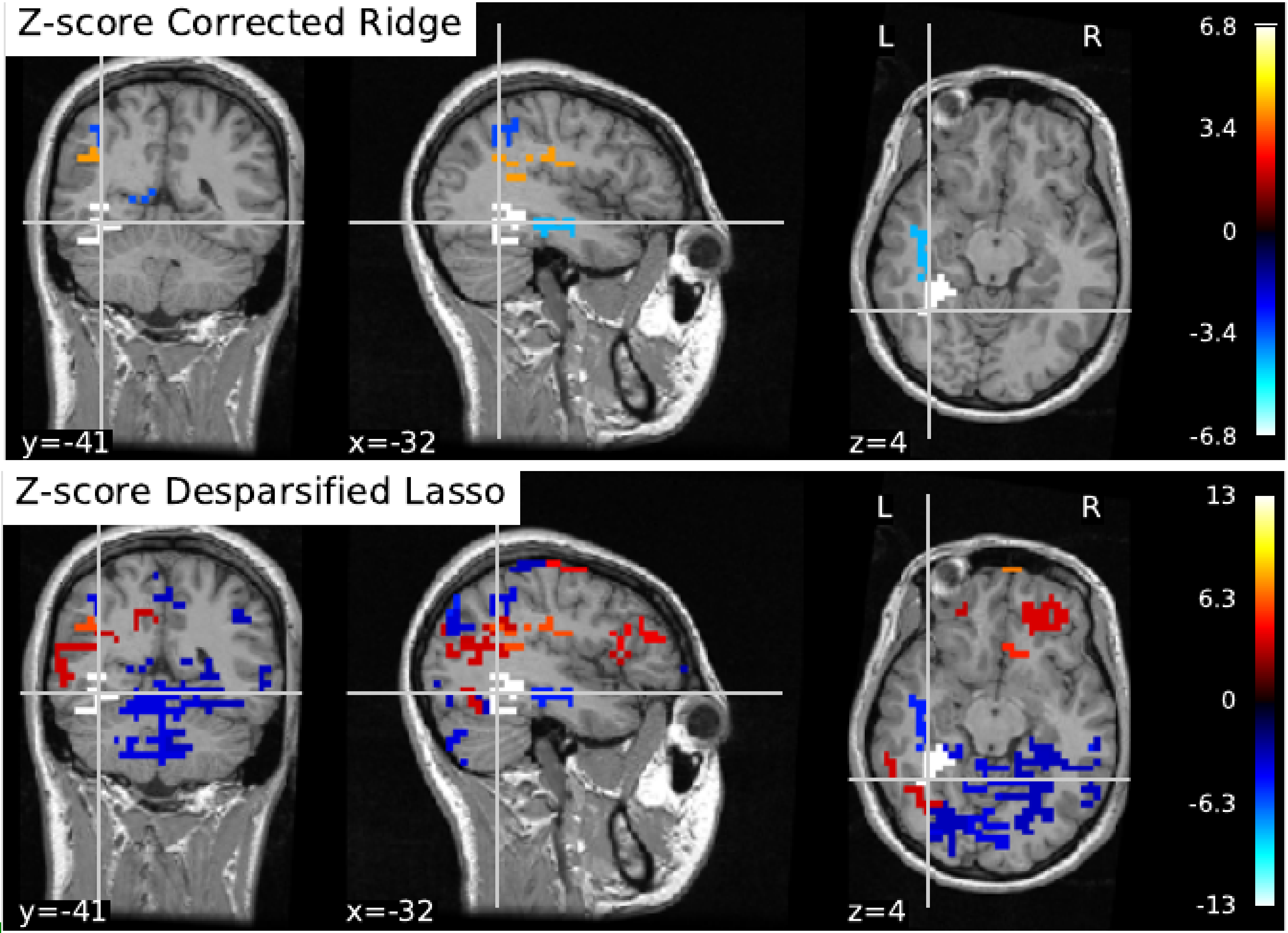


Corrected Ridge and Desparsified Lasso when $p \approx n$

Preliminary assessment

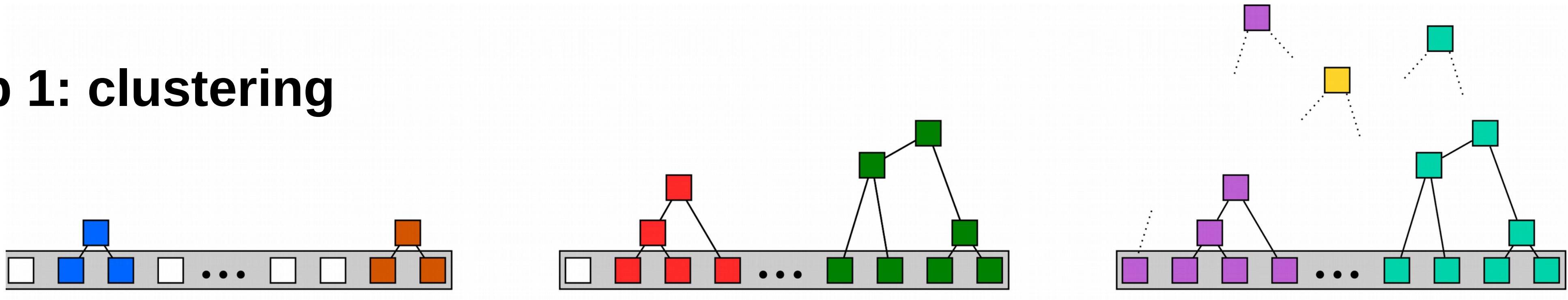


Preliminary assessment



Adaptation to brain imaging

Step 1: clustering



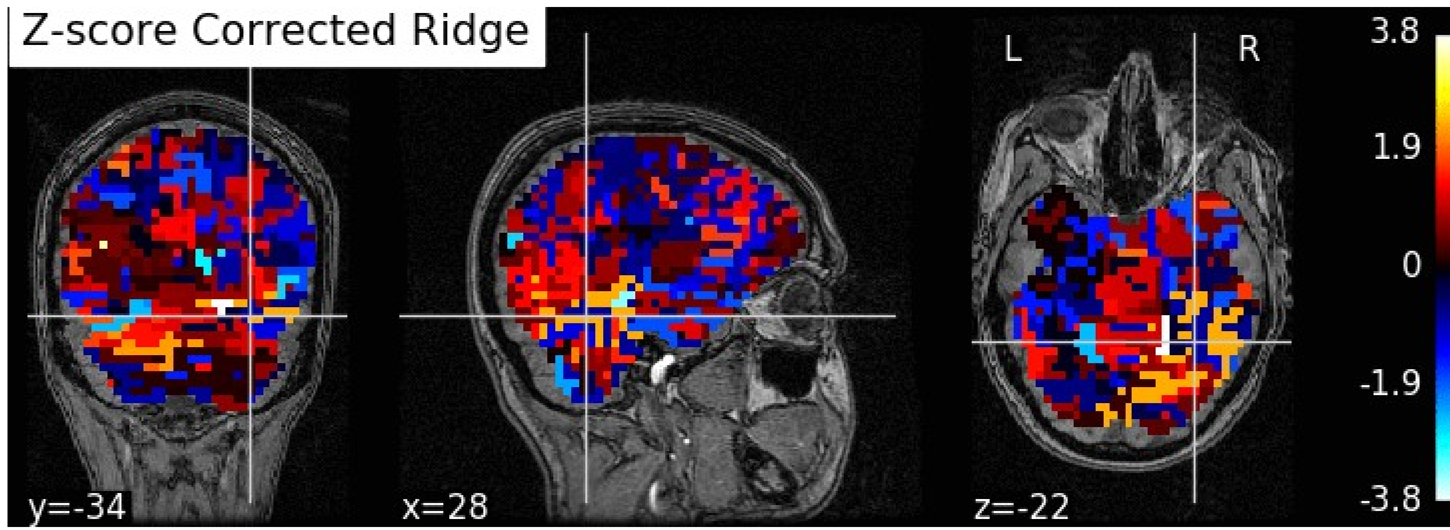
Step 2: inference on compressed representations

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

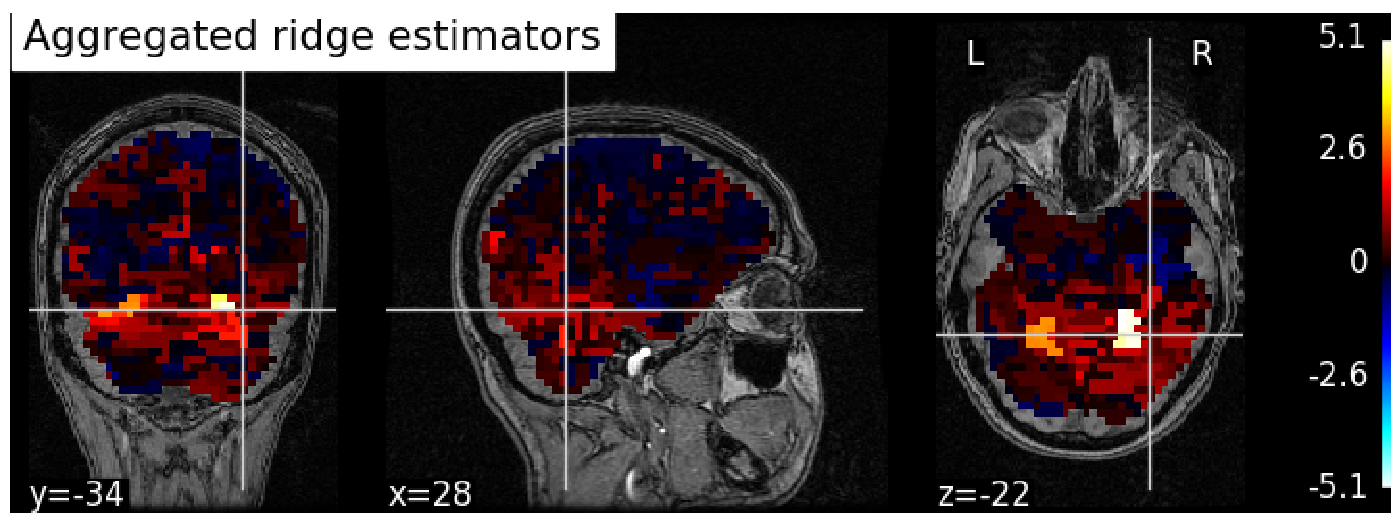
Step 3: repeat on different parcellations and aggregate the p-values
(FReM-like approach)

Some initial results

DL p-values from different clusterings

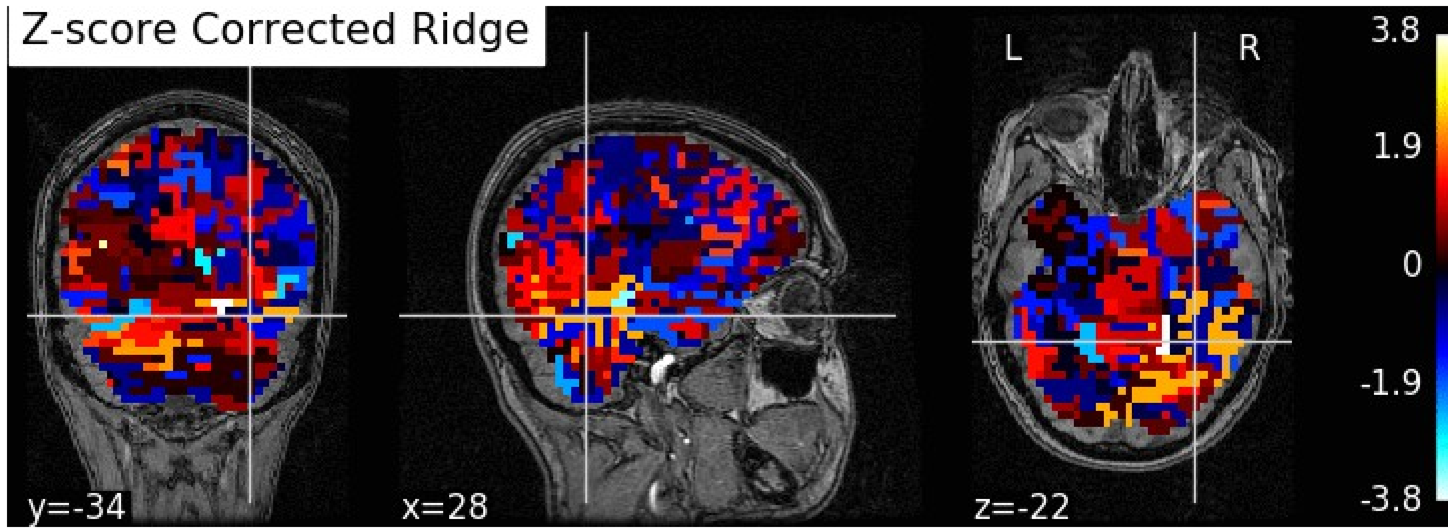


aggregation

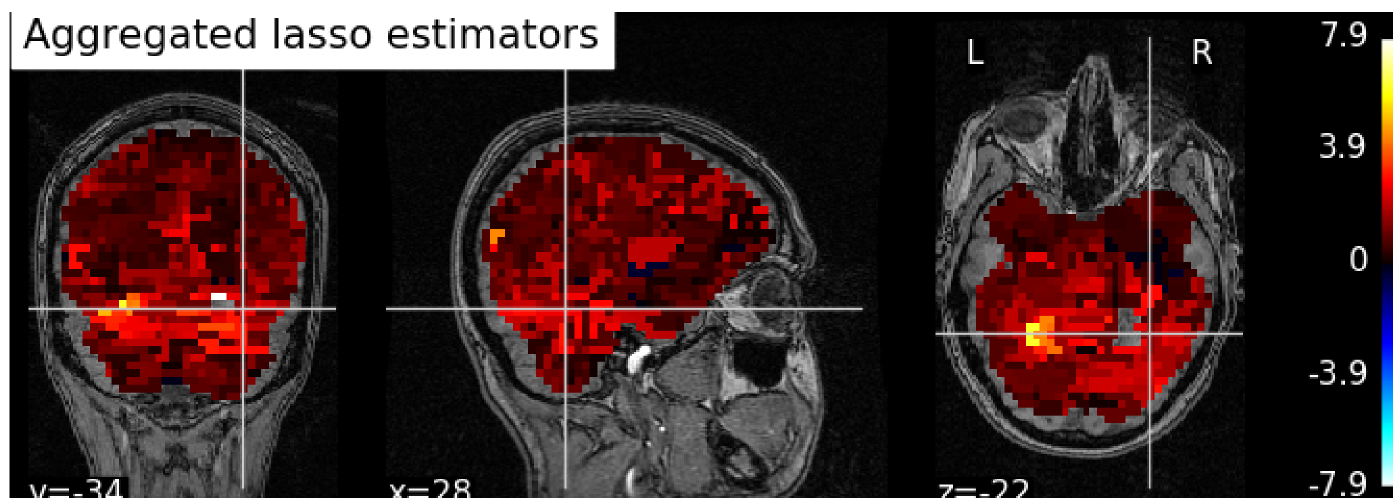


Some initial results

DL p-values from different clusterings



aggregation



Conclusion

- Large-p data bring challenges:
  - Computation cost
  - Overfit
  - Difficulty of statistical inference
- Solutions: online learning, subsampling, compression
- **Ensembling improves estimators**
- Open frontiers: statistical inference

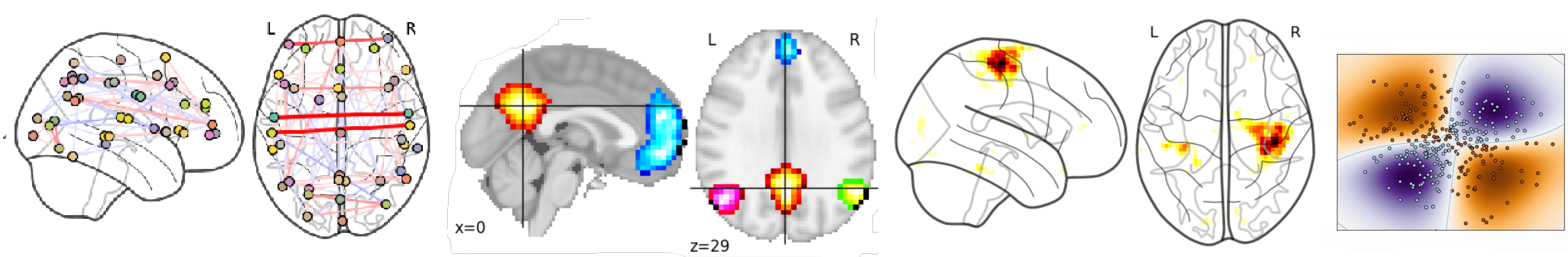
From good ideas to good practices: software

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,
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