# Spatial regularization and sparsity for multi-subject brain activity decoding

Bertrand Thirion,
INRIA Saclay-Île-de-France, Parietal team
http://parietal.saclay.inria.fr
bertrand.thirion@inria.fr

INVENTEURS DU MONDE NUMÉRIQUE

# Outline

- Machine learning techniques for brain activity decoding in functional neuroimaging
- Contribution 1: Tree-based decoding
- Contribution 2: Total Variation regularization for penalized regression

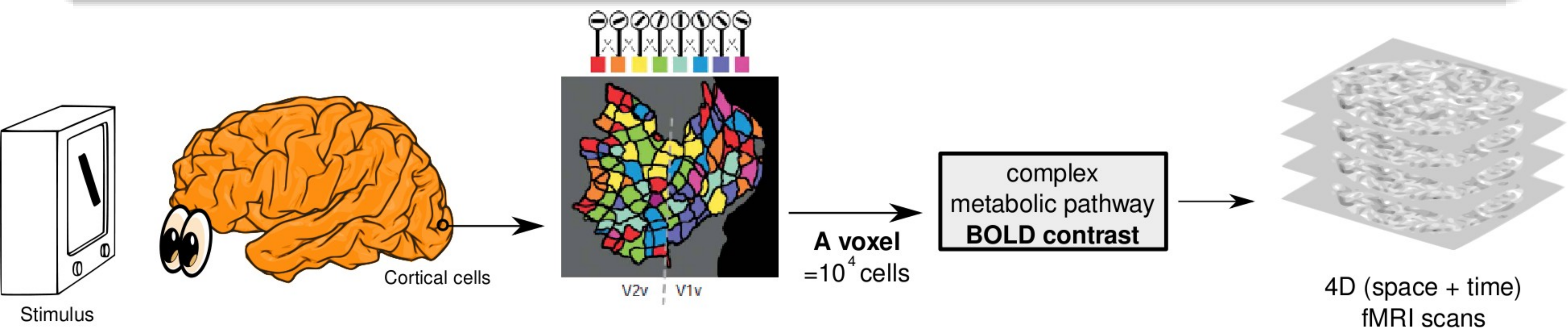
# Functional MRI for brain activity decoding

Functional neuroimaging
→ reveal brain physiological activity and its spatial distribution

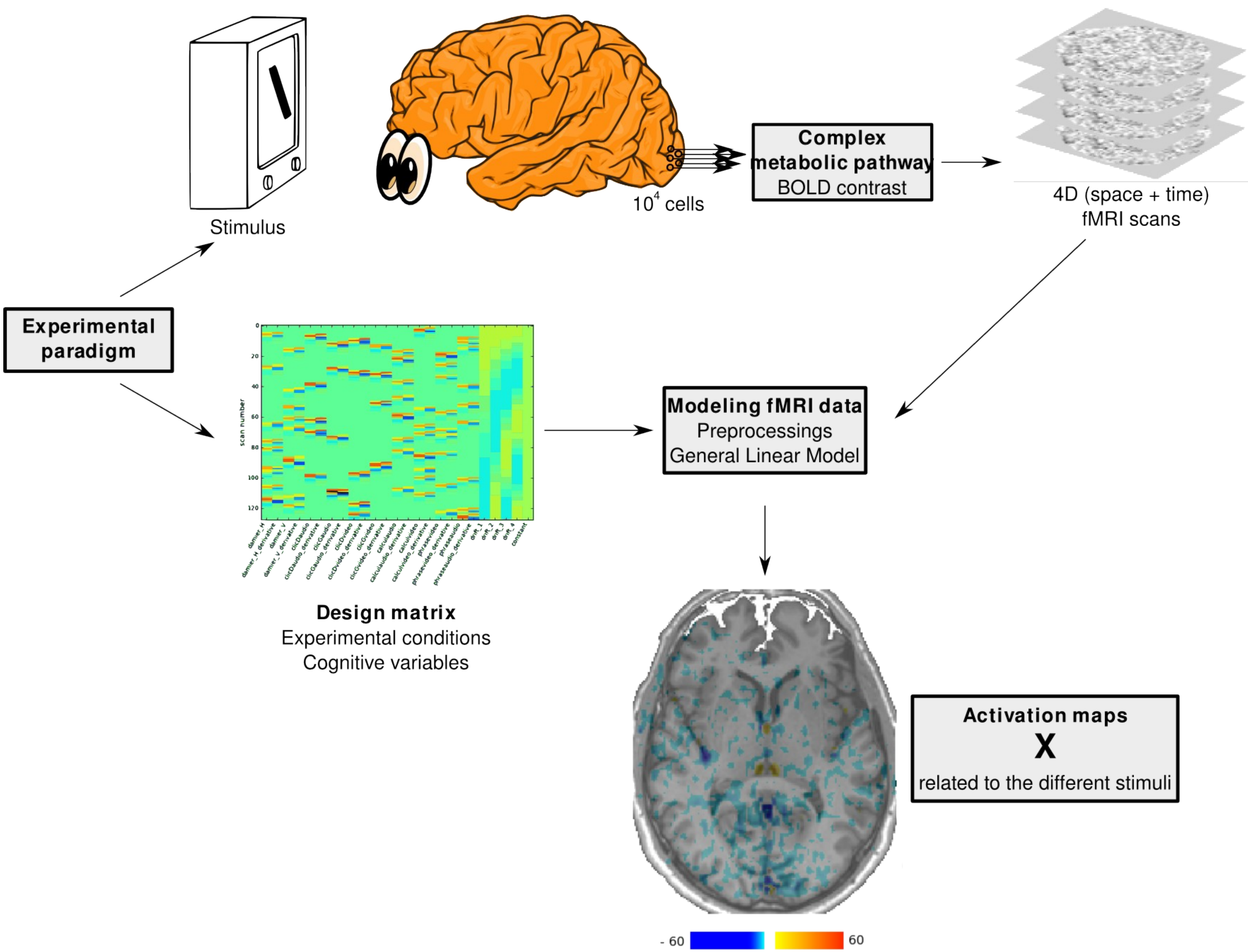
**Functional Magnetic Resonance Imaging - fMRI**

- ✔ non-invasive.
- ✔ good spatial resolution → voxel (volumetric pixel) $\sim 2 \times 2 \times 2$mm.
- ✔ *Blood Oxygenation Level-dependent (BOLD)* contrast

→ **measures a metabolic correlate of neural activity.**



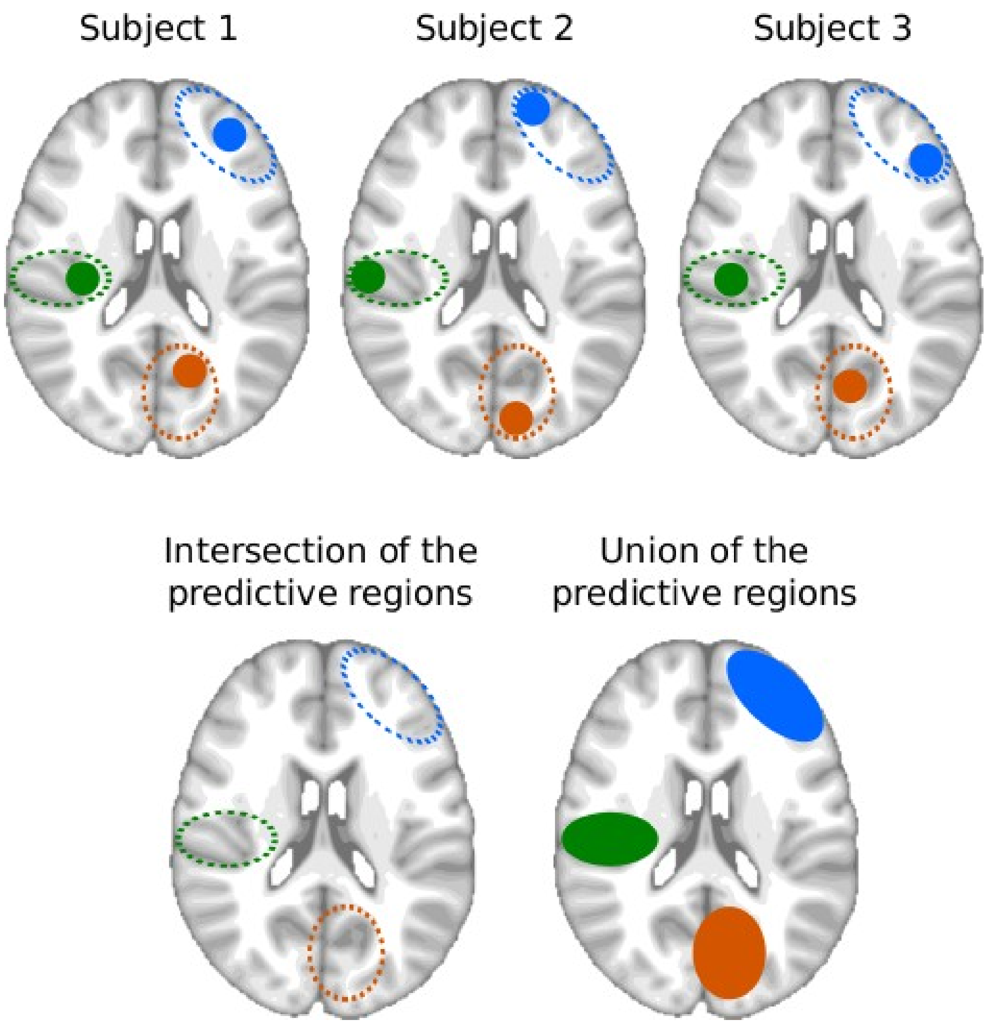
# Encoding of fMRI data



# Inter-subject variability

Inter-subject prediction → find predictive regions stable across subjects.
Inter-subject variability → lack of voxel-to-voxel correspondence

[Tucholka 2010]



# Prediction function

Predictive linear model

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = \mathbf{X}\,\mathbf{w} + b$$

$y \in R^n$ is the behavioral variable.
$X \in R^{n \times p}$ is the data matrix, i.e. the activations maps.
(w, b) are the parameters to be estimated.
**n** activation maps (samples), **p** voxels (features).

$y \in R^n \rightarrow$ regression setting :
f (X, w, b) = X w + b ,
$y \in \{-1, 1\}^n \rightarrow$ classification setting :
f (X, w, b) = sign(X w + b) ,
where “sign” denotes the sign function.

# Prediction functions in fMRI

- **Choosing the prediction function f (X, w, b)**
  - Kernel machines (SVC, SVR, RVM)
  - Discriminant analysis (LDA, QDA)
  - Regularized [logistic] regression (Lasso, Ridge, Elastic net)
- $p \gg n$ **Curse of dimensionality**

  Always possible to find a prediction function with perfect prediction on the data used for learning

  → learn noise or non-informative features of fMRI data.

  – cannot generalize to new samples

  → **Dimension Reduction/regularization** is mandatory.

# Dealing with the curse of dimensionality in fMRI

- **Feature selection** (e.g. Anova, RFE) :
  - Regions of interest → requires strong prior knowledge.
  - Univariate methods → selected features can be redundant.
  - Multivariate methods → combinatorial explosion, computational cost.

    [Mitchell et al. 2004], [De Martino et al. 2008]
- **Regularization** (e.g. Lasso, Elastic net) :
  - performs jointly feature selection and parameter estimation → majority of the features have zero/close to zero loadings.

    [Yamashita et al. 2004], [Carroll et al. 2010]
- **Feature agglomeration** :
  - agglomeration : construction of intermediate structures → based on the local redundancy of information.

    [Filzmoser et al. 1999], [Flandin et al. 2003]

# Evaluation of the decoding

**Prediction accuracy**

Explained variance ζ :

$$\zeta(y^t, \hat{y^t}) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^t - \hat{y_i}^t\right)^2}{\mathrm{var}(y^t)}$$

$$\kappa(y^t, \hat{y^t}) = \frac{1}{N}\sum_{i=1}^{N}\delta(y_i^t - \hat{y_i}^t)$$

→ assess the quantity of information shared by the pattern of voxels.

**Structure of the resulting maps of weights:** reflect our hypothesis on the spatial layout of the neural coding ?

**Common hypothesis :**

→ **sparse** : few relevant voxels/regions implied in the cognitive task.

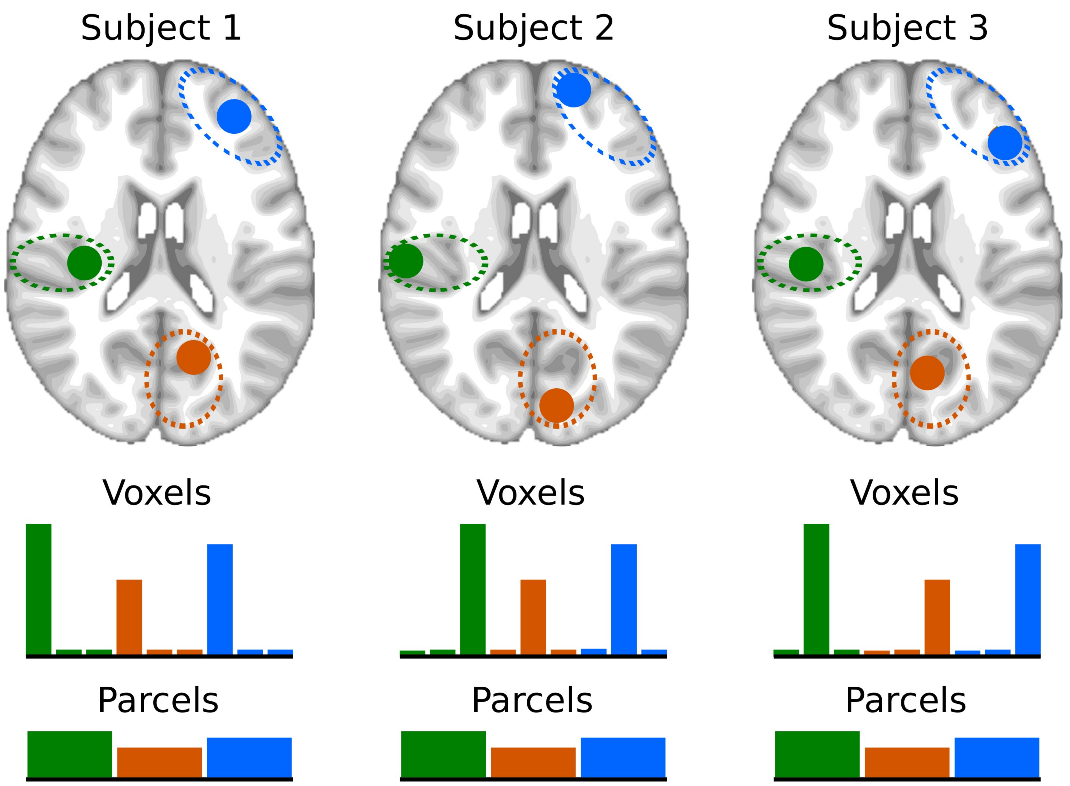
→ **compact structure** : relevant features grouped into connected clusters.

# Outline

- Machine learning techniques for brain activity decoding in functional neuroimaging
- Contribution 1: Tree-based decoding
- Contribution 2: Total Variation regularization for penalized regression

# Feature agglomeration

- Parcels: sets of connected voxels.
- Thought to correspond to meaningful structures in the brain (~cortical areas) [Filzmoser et al. 1999, Thirion et al. 2006, Golland et al. 2007]




- Reduce the dimensionality of the problem by averaging or grouping: $10^5$ voxels → $10^2$ parcels
- Cope with inter-subject variability.

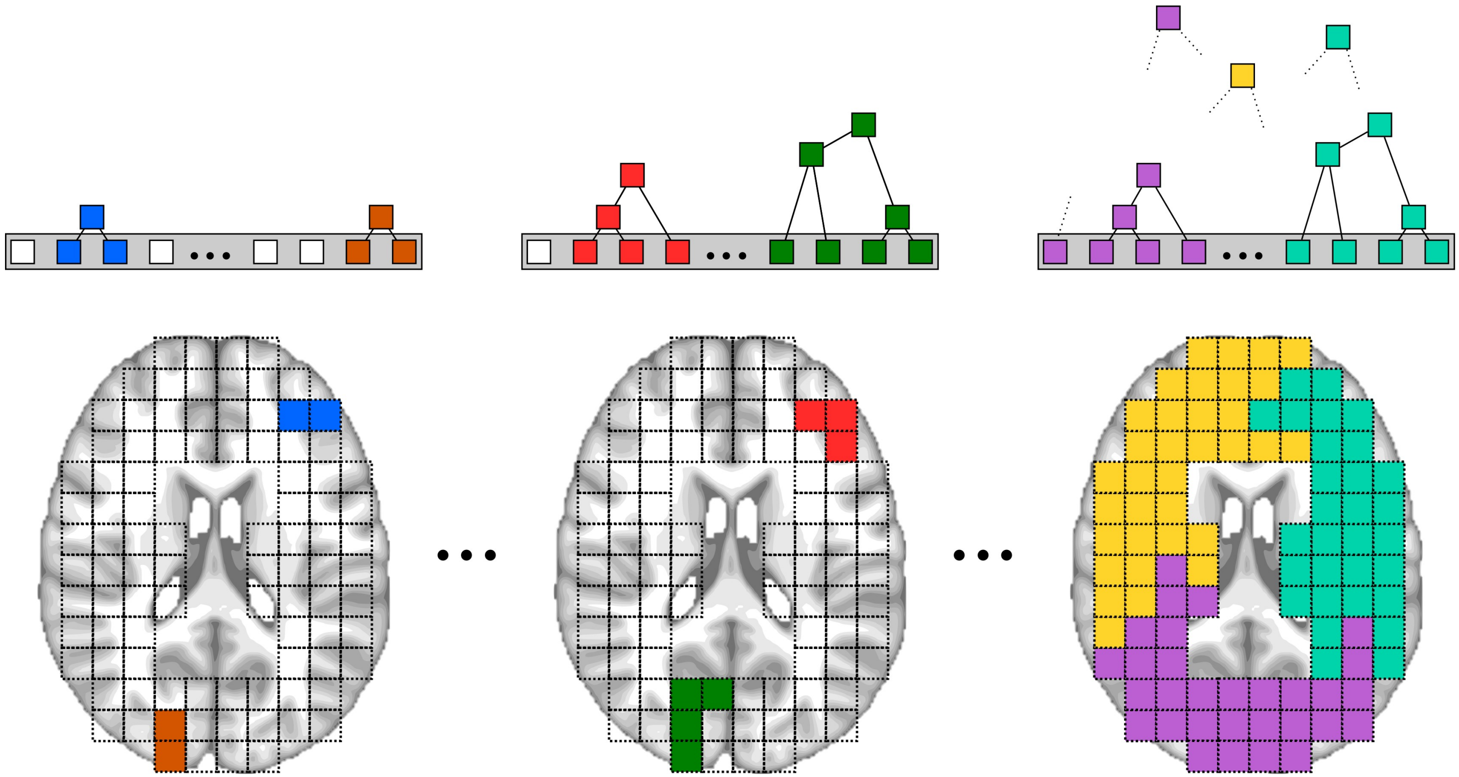
# Creating the parcels

**Hierarchical clustering → multi-scale approach**

Ward's algorithm - [J. H. Ward. 1963]

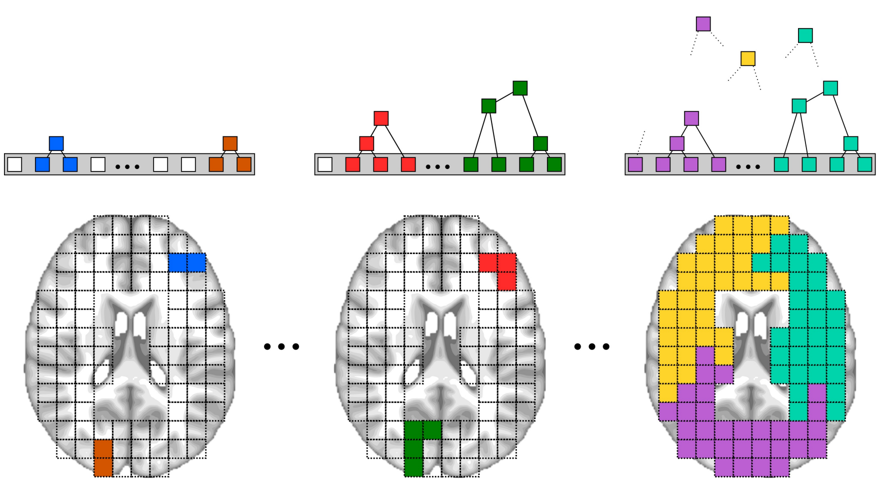
Minimizes the variance of the resulting parcels.

In our implementation, we add spatial connectivity constraints.



# Structured sparsity for fMRI data

- Structure:
- Hierarchical clustering of the brain volume
- Variance minimization (Ward's clustering)
- With connectivity constraints
- Nested/multi-scale



- Sparsity: group lasso on the clusters of the tree

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 = \sum_{g \in \mathcal{G}} \Big[ \sum_{j \in g} \mathbf{w}_j^2 \Big]^{1/2}$$

- Acts as the $l_1$-norm on the vector $(\|\mathbf{w}_g\|_2)_{g \in \mathcal{G}}$
- If one node is set to 0 , its descendants are also set to 0
- Consider large parcels before small parcels → robustness to spatial variability

# Optimization of the model

- Use of proximal methods for speed-up
  - Extension of gradient-based methods for non-smooth criteria [Nesterov, 2007]
  - Algorithm described in [Jenatton et al., ICML 2010]
    - Initial problem $\min_{w\in\mathbb{R}^p} \|Y - Xw\|^2 + \lambda\Omega(w) = \ell(w) + \lambda\Omega(w)$
    - Proximal $\min_{w\in\mathbb{R}^p} \ell(\hat{w}) + (w-\hat{w})^T \nabla\ell(\hat{w}) + \lambda\Omega(w) + \frac{L}{2}\|w-\hat{w}\|^2$
    - Which yields $\min_{w\in\mathbb{R}^p} \frac{1}{2}\|w - (\hat{w} - \frac{1}{L}\nabla\ell(\hat{w}))\|^2 + \frac{\lambda}{L}\Omega(w)$
    - And boils down to $\mathrm{prox}_{\lambda,\Omega}(v) = \min_{w\in\mathbb{R}^p} \|w - v\|^2 + \lambda\Omega(w)$
  - Computation of the proximal is efficient in the dual space

# Real fMRI dataset on representation of objects

4 different objects.

3 different sizes.

10 subjects, 6 sessions, 12 images/session. 70000 voxels.
**Inter-subject experiment** : 1 image/subject/condition → 120 images.
[Eger et al. - 2008]

# Results on real data

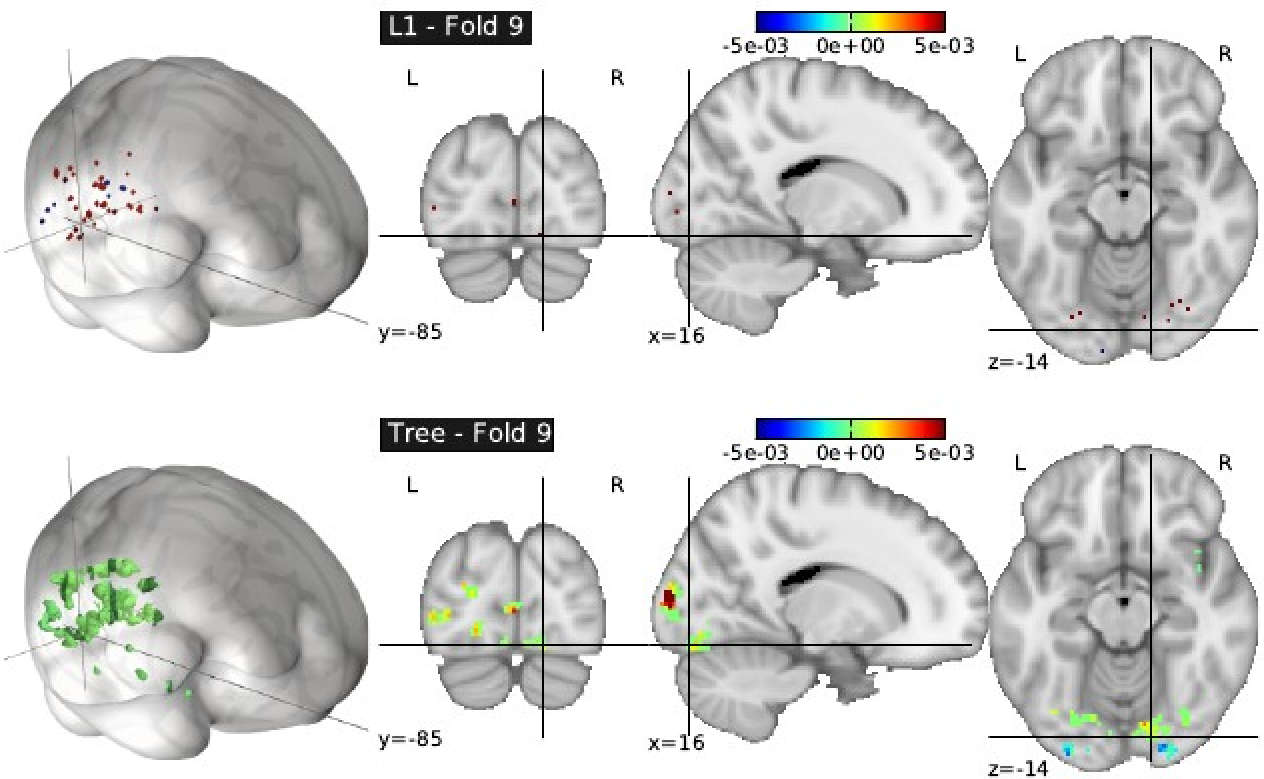
| Regularization function | mean error | std error | p-value w.r.t. Hierarchical Tree $\ell_2$ | median % non-zero coef. |
|---|---|---|---|---|
| Ridge - $\ell_2$ | 8.3 | 4.6 | 0.096 | 100.00 |
| Lasso - $\ell_1$ | 12.1 | 6.6 | 0.013* | 0.10 |
| Adaptive Lasso | 11.3 | 8.8 | 0.05* | 0.10 |
| $\ell_1$ (Tree weights) | 8.4 | 4.7 | 0.03* | 0.02 |
| Hierarchical Tree $\ell_2$ | 7.1 | 4.0 | - | 9.36 |

(Wilcoxon two-sample paired signed rank test)

- In the regression task, hierarchical tree $l_2$, yields significantly better prediction than the alternatives
- The sparsest models do not perform so well
- Not too sensitive to choice of λ

# Results on real data (2)

- Spatial maps: sparse, but with some compactness (spatial grouping / clustering)
- Easier to describe/report than Lasso maps
- Results in more robustness to spatial variability and more reproducible



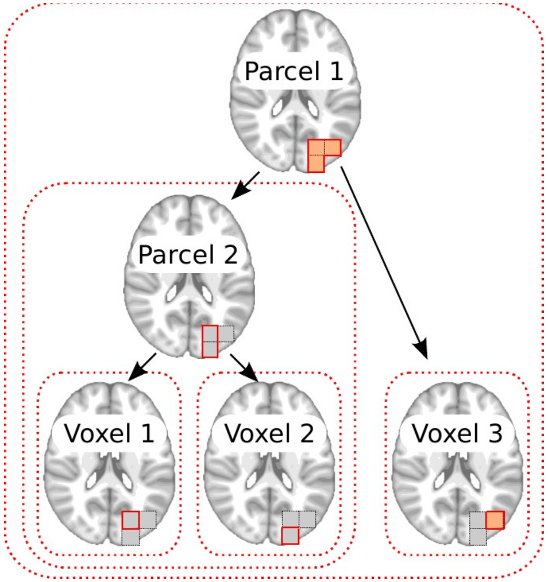
# Discussion

- Discover the spatial model that provides a maximal amount of information on the target variable
  - Find also the proper scale
- Convex criterion: an optimal solution is obtained
- The model favors large clusters against smaller ones
  - Built-in model selection
  - More robustness to inter-subject spatial variability
  - More reproducibility
- Yet a greedy approach [with no theoretical guarantee] is almost as sensitive and more efficient.

# Perspectives

- Multi-task version
- Other multi-subject datasets (diagnosis) – the method is well-suited to deal with between-subject variability
- Can also work on any dataset with multi-scale structure
- Efficiency/optimality tradeoff ?

R.Jenatton Rodolphe, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, B. Thirion. Multi-scale Mining of fMRI data with Hierarchical Structured Sparsity. PRNI 2011




V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin and B. Thirion. *A supervised clustering approach for fMRI-based inference of brain states*. Pattern Recognition - Special Issue on ’Brain Decoding’, in press.

# Outline

- Machine learning techniques for brain activity decoding in functional neuroimaging
- Contribution 1: Tree-based decoding
- Contribution 2: Total Variation regularization for penalized regression

# Regularization framework

**Constrain the values of w to select few parameters which explain well the data.**

Use of **penalized regression** → Minimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\operatorname{argmin}} \; \ell(\mathbf{y}, \mathbf{X}\mathbf{w}) + \lambda J(\mathbf{w}) \quad , \quad \lambda \geq 0$$

- λJ(w) is the **penalization term**.
- ℓ(y, Xw) is the *loss function*, usually $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ for regression.
- λ ≥ 0 balances the loss function and the penalty.
- Perform feature selection and parameter estimation *jointly*.

Usually: J is a $L_1$ or $L_2$ norm (ridge, lasso, elastic net)

# Total Variation (TV) regularization

Penalization J(w) based on the **$l_1$ norm of the gradient of the image**

$$J(\mathbf{w}) = TV(\mathbf{w}) = \int_{\omega \in \Omega} \|\nabla \mathbf{w}\| d\omega$$

[L. Rudin, S. Osher, and E. Fatemi - 1992], [A. Chambolle - 2004]

gives an estimate of w with a **sparse block structure**

→ take into account the spatial structure of the data.

extracts regions with piecewise constant weights

→ well suited for brain mapping.

requires computation of the gradient and divergence over a mask of the brain with correct border conditions.

# TV-based prediction

First use of TV for prediction task.

Minimization problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\operatorname{argmin}} \; \ell(\mathbf{y}, \mathbf{X}\mathbf{w}) + \lambda TV(\mathbf{w}) \quad , \quad \lambda \geq 0$$

Regression → least-squares loss :

$$\ell(\mathbf{y}, \mathbf{X}\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Classification → logistic loss :

$$\ell(\mathbf{y}, \mathbf{X}\mathbf{w}) = \frac{\sum_{i=1}^{n} \log\left(1 + \exp^{-y_i(\mathbf{x_i}^T \mathbf{w})}\right)}{n}$$

TV(w) not differentiable but convex
→ optimization by iterative procedures (ISTA, FISTA).
[I. Daubechies, M. Defrise and C. De Mol - 2004], [A. Beck and M. Teboulle - 2009]

# Convex optimization for TV-based decoding

First order iterative procedures:

- FISTA procedure
  - → TV (ROF problem).
- ISTA procedure
  - → main minimization problem

Natural stopping criterion:

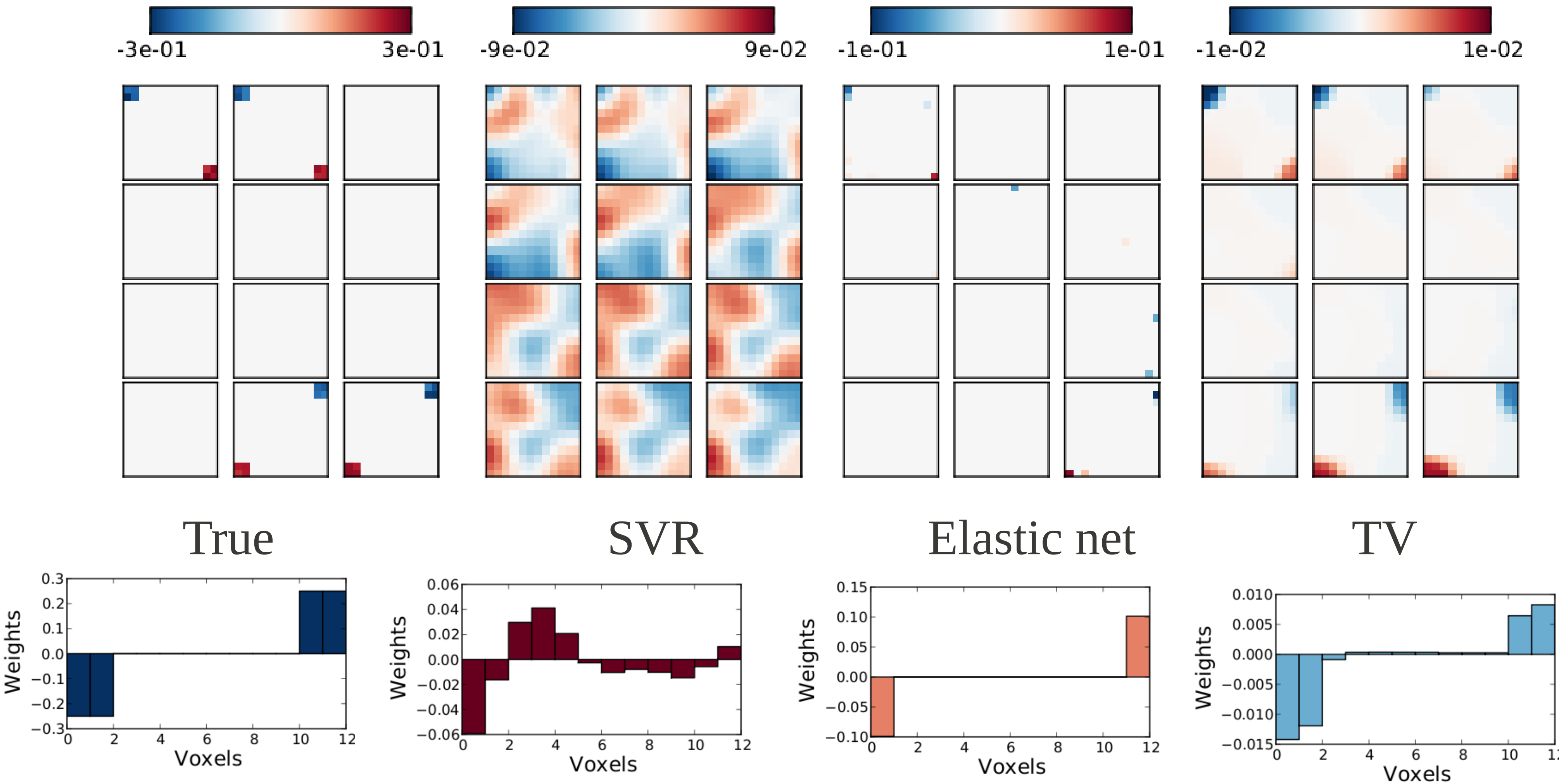
duality gap.

---

**Require:** Set maximum number of iterations $K$ (*ISTA*), and the threshold $\epsilon$ on the dual gap (*FISTA*).
**Require:** Initialize $\mathbf{z} \in \mathbb{R}(\Omega^3)$ with zeros.

**### ISTA loop ###**
**for** $k = 1 \ldots K$ **do**
  $\mathbf{u} = \mathbf{w} - \frac{1}{L}\nabla\mathcal{L}(\mathbf{w})$
  **### FISTA loop ###**
  Initialize $\mathbf{z}_{aux} = \mathbf{z}$, $t = 1$
  **while** $\delta_{gap}(\mathbf{u} + \lambda \mathrm{div}(\mathbf{z})) > \epsilon$ **do**
    $\mathbf{z}_{old} = \mathbf{z}$
    $\mathbf{z} = \Pi_K \left( \mathbf{z}_{aux} - \frac{1}{\lambda \tilde{L}} \mathrm{grad}(L\mathbf{u} + \lambda \mathrm{div}(\mathbf{z}_{aux})) \right)$
    $t_{old} = t$
    $t = (t + \sqrt{1 + 4t^2})/2$
    $\mathbf{z}_{aux} = \mathbf{z} + \frac{t_{old} - 1}{t}(\mathbf{z} - \mathbf{z}_{old})$
  **end while**
  $\mathbf{w} = \mathbf{u} + \lambda \mathrm{div}(\mathbf{z})$
**end for**
**return** $\mathbf{w}$

# Intuition on simulated data




→ extract weights with a sparse block structure.

# Prediction accuracy on inter-subject analyzes

Regression analysis

| Methods | mean $\zeta$ | std $\zeta$ | max $\zeta$ | min $\zeta$ | p-value to TV |
|---|---|---|---|---|---|
| SVR | 0.77 | 0.11 | 0.97 | 0.58 | 0.0277 ** |
| Elastic net | 0.78 | 0.1 | 0.97 | 0.65 | 0.0405 ** |
| TV $\lambda = 0.05$ | 0.84 | 0.07 | 0.97 | 0.72 | - |

Classification analysis

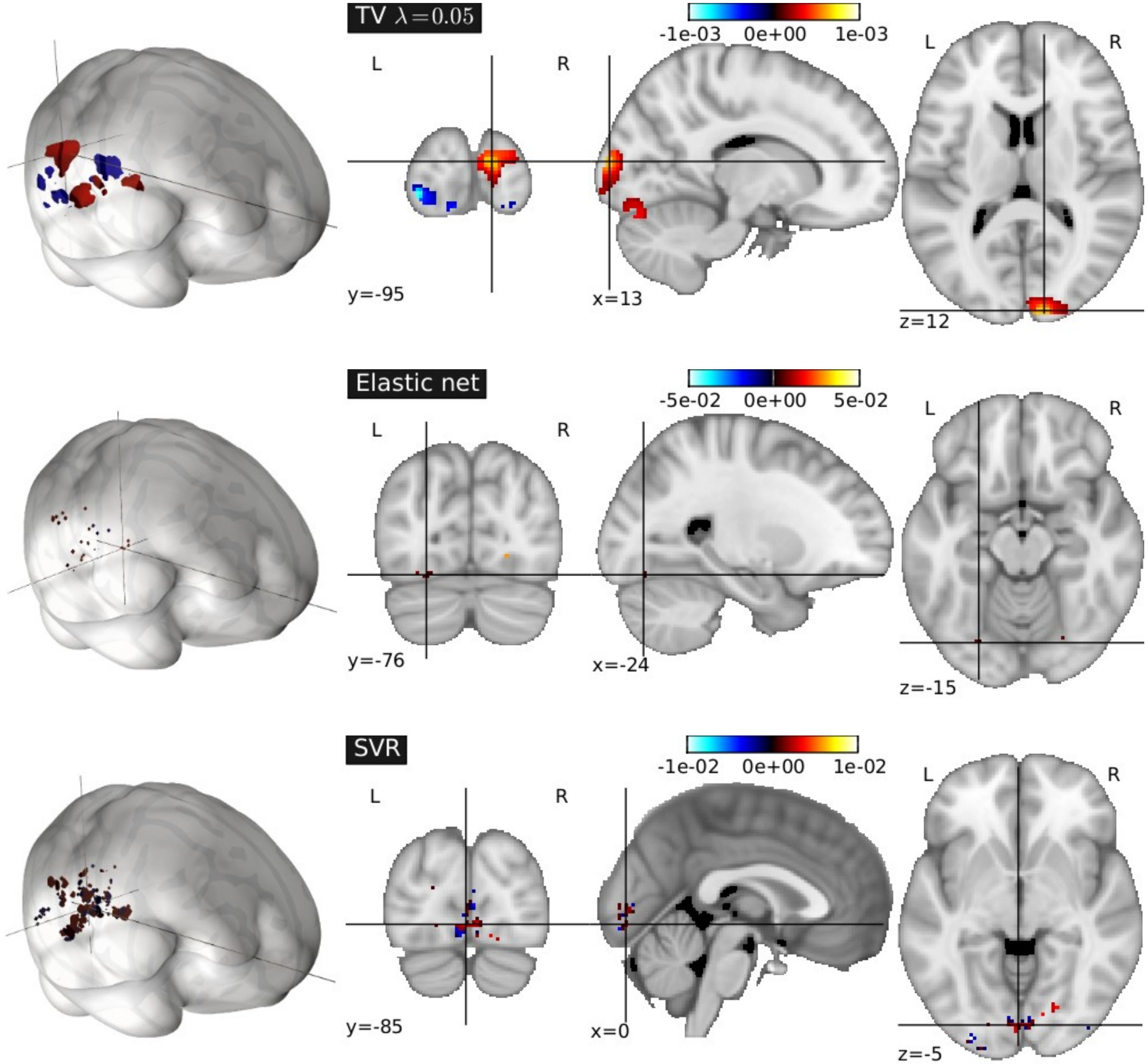
| Methods | mean $\kappa$ | std $\kappa$ | max $\kappa$ | min $\kappa$ | p-value to SVC |
|---|---|---|---|---|---|
| SVC | 48.33 | 15.72 | 75.0 | 25.0 | - |
| SMLR | 42.5 | 9.46 | 58.33 | 33.33 | 0.2419 |
| TV $\lambda = 0.05$ | 45.83 | 14.55 | 66.67 | 25.0 | 0.7128 |

# TV → maps for brain mapping

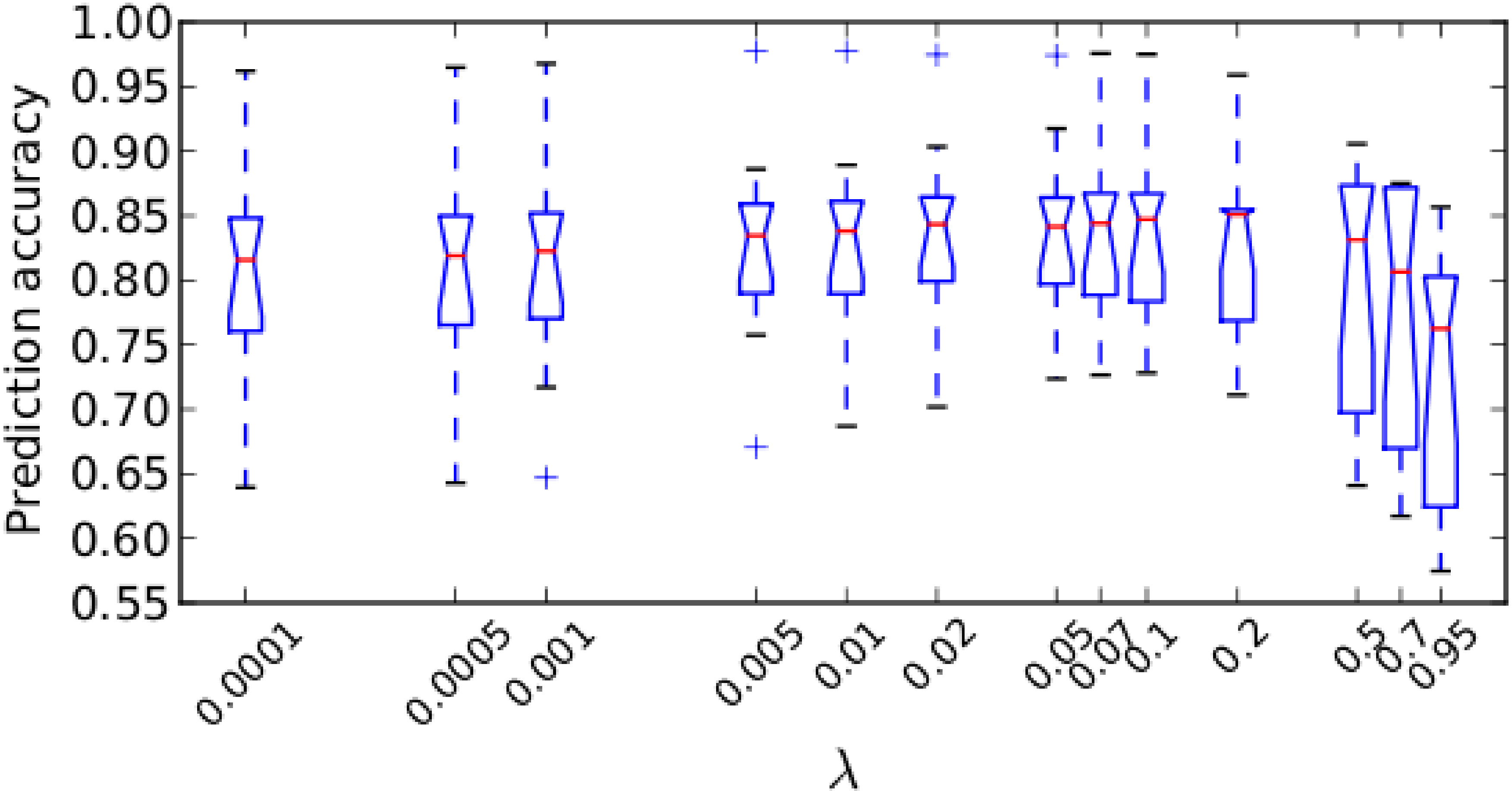
TV

Elastic net

SVR



# Influence of the regularization parameter λ



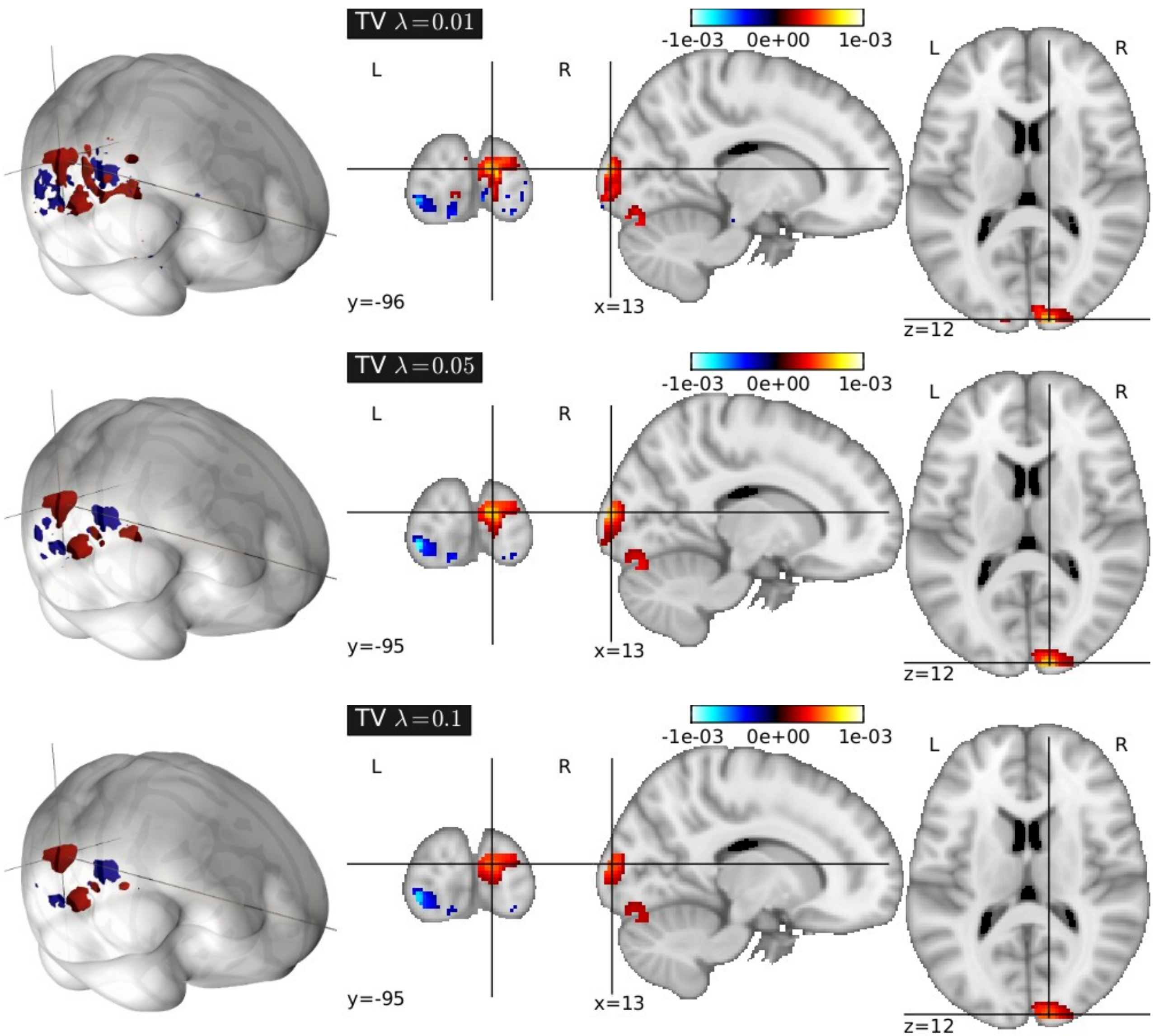

→ results are extremely stable with respect to λ.

# Influence of the regularization parameter λ

$\lambda = 0.01$
$\zeta = 0.83$

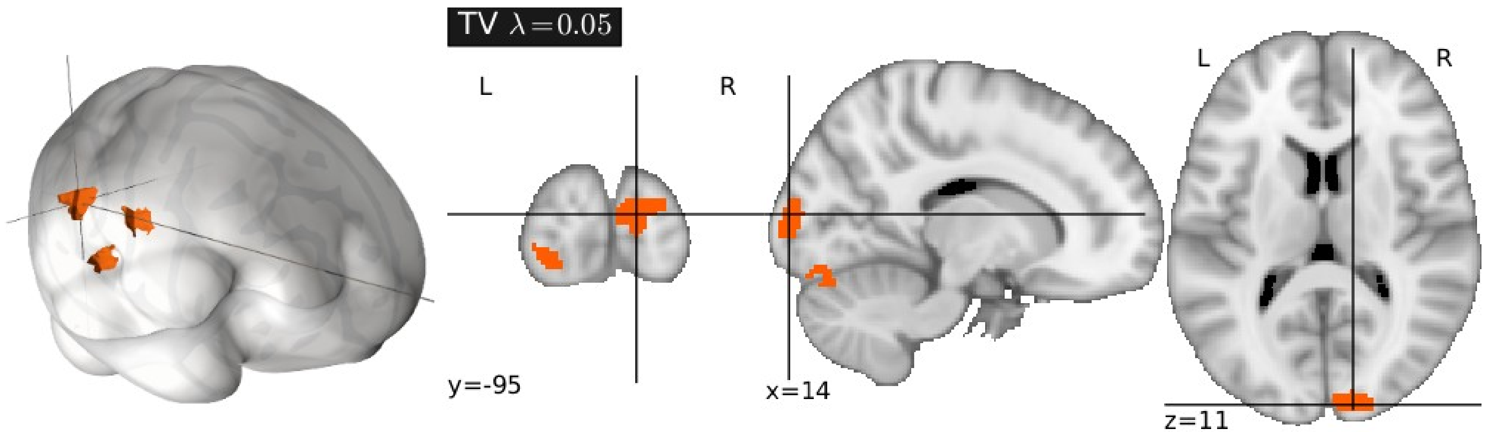
$\lambda = 0.05$
$\zeta = 0.84$

$\lambda = 0.1$
$\zeta = 0.84$

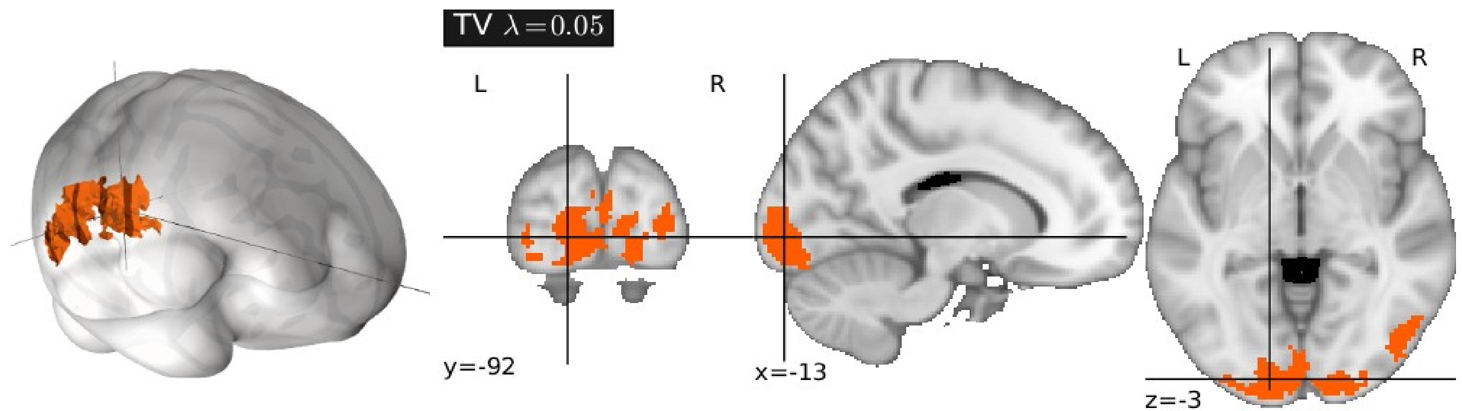


# TV for fMRI-based decoding

Inter-subject regression analysis.




Inter-subject classification analysis.




→ derive maps similar to classical inference, within the inverse inference framework.

# Conclusion on TV regularization

**First use of TV for prediction problem** (classification/regression).

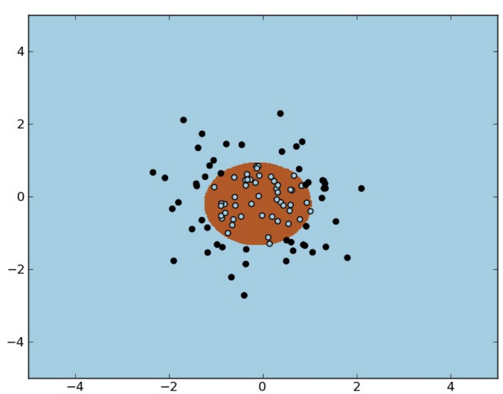
- TV approach allows to **take into account the spatial structure of the data** in the regularization.
  → yields better prediction accuracy than reference methods.
- TV d**eals with inter-subject variability.**
  → well suited for inter-subjects analysis.
- TV creates **cluster-like activation maps.**
  → provides interpretable maps for brain mapping.

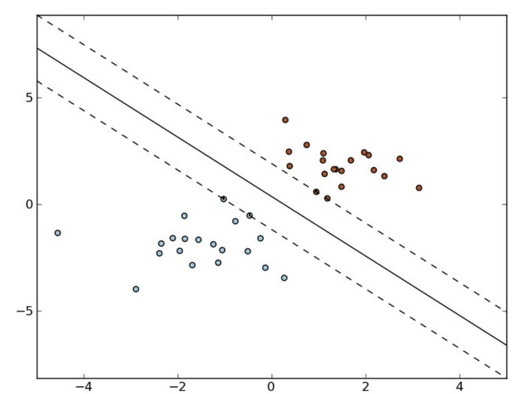

- V. Michel, A. Gramfort, G. Varoquaux and B. Thirion. *Total Variation regularization enhances regression-based brain activity prediction*. In 1st ICPR Workshop on Brain Decoding. 2010.
- V. Michel, A. Gramfort, G. Varoquaux, E. Eger and B. Thirion. *Total variation regularization for fMRI-based prediction of behaviour*. Submitted to IEEE Transactions on Medical Imaging. 2010.

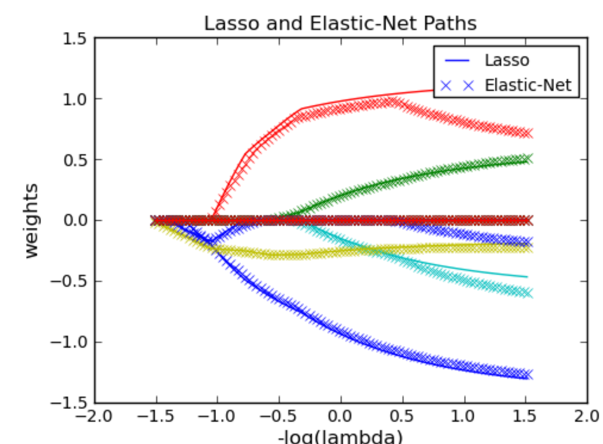
# scikit learn: open source kit for machine learning (in python)

scikits learn
machine learning in Python

- Started dec. 2009; mainly developed by F. Pedregosa (INRIA Parietal), but shared with a wide community
- Contains **standard** tools for machine learning: classifiers, regression, feature selection, clustering, dimension reduction
- Emphasis on **efficiency** (moderate computation time) and easy/intuitive use (doc + tests + examples)
- Not dedicated to neuroimaging (but many parts have been developed in view of neuroimaging applications) – see http://nisl.github.com/
- Freely available, open to contributions http://scikit-learn.org



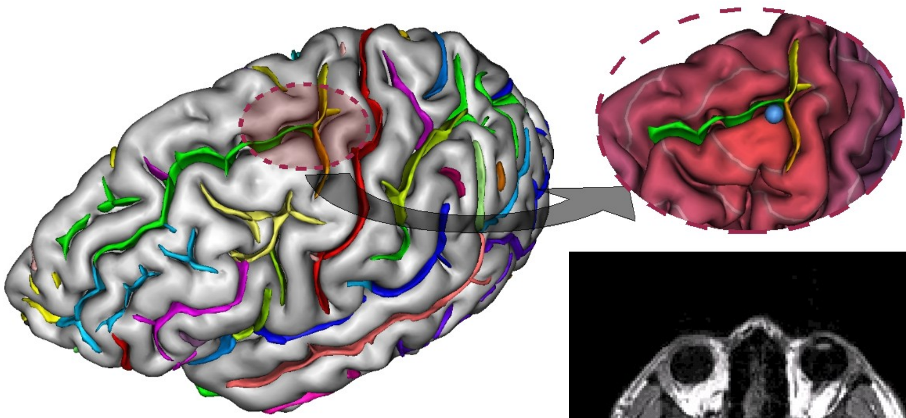


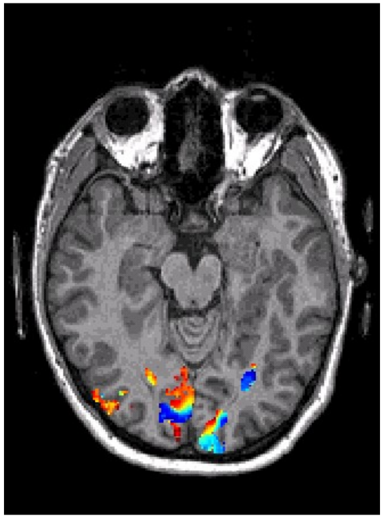


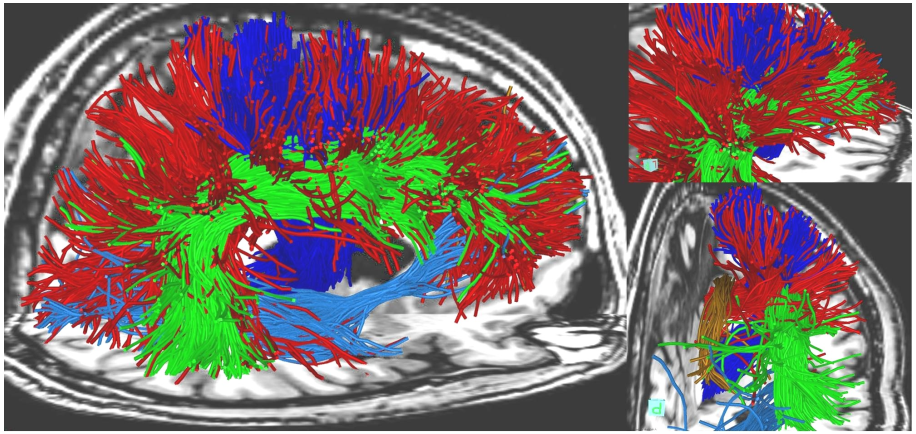
# Acknowledgements


- Many thanks to my co-workers: **V. Michel**, G. Varoquaux, **A. Gramfort**, F. Pedregosa, P. Fillard, J.B. Poline, V.Fritsch, V. Siless, S.Medina, R. Bricquet
- To INRIA colleagues: G.Celeux, C. Keribin, F. Bach, R. Jenatton, G. Obozinski
- To CEA/Neurospin & INSERM U562 colleagues: E.Eger, A. Kleinschmidt, S.Dehaene, J.F. Mangin

# Thank you for your attention







http://parietal.saclay.inria.fr