Decoding the informative content of brain activation maps: state of the art, challenges and future directions

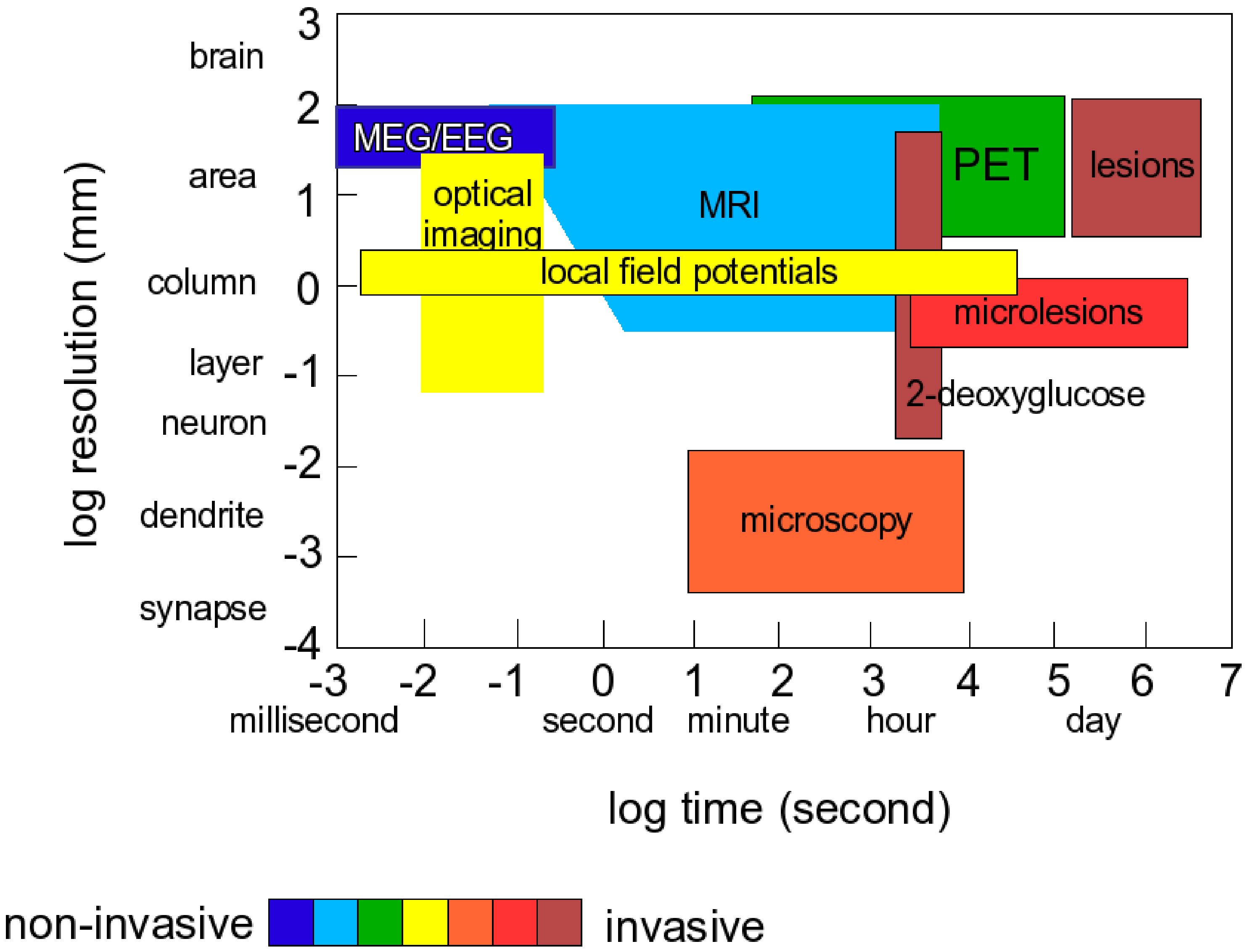
Bertrand Thirion,
INRIA Saclay-Île-de-France, Parietal team
http://parietal.saclay.inria.fr
bertrand.thirion@inria.fr

INVENTEURS DU MONDE NUMÉRIQUE

Outline

- Machine Learning in Neuroimaging
    - Overview
    - Common technical challenges
- Some learning problems in neuroimaging:
  - Medical diagnosis/study of between subject-variability
  - Brain reading
  - Brain connectivity mapping

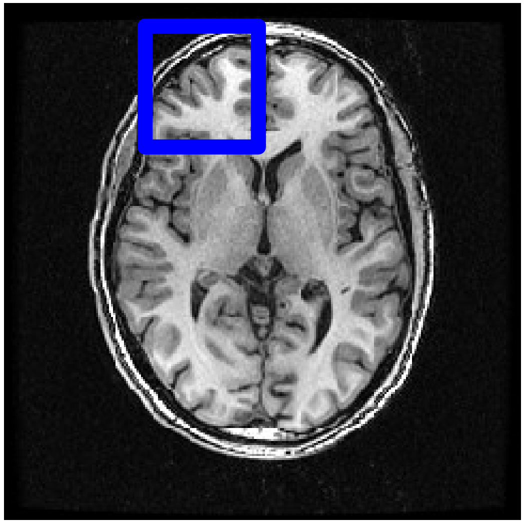
NeuroImaging: modalities and aims

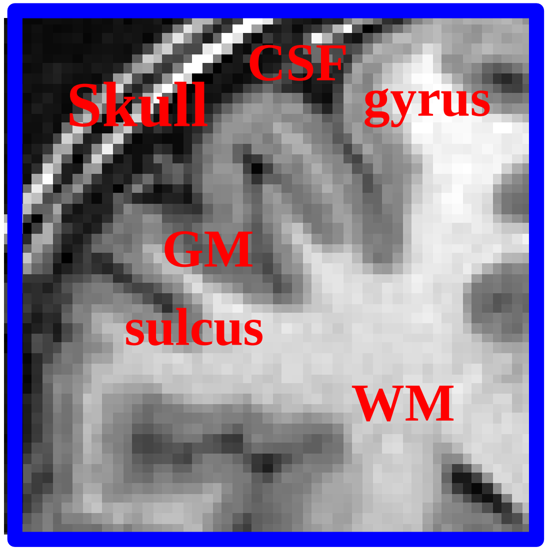


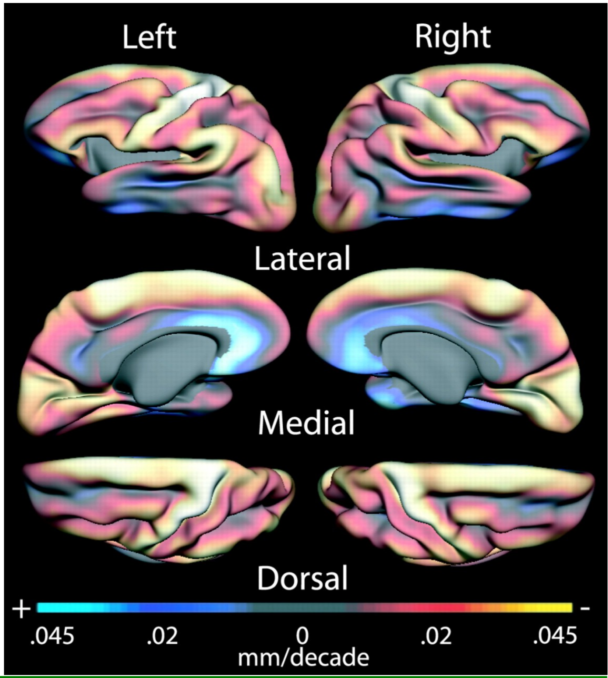
- 'Functional' (time resolved) modalities: fMRI, EEG, MEG
- vs 'anatomical' (spatially resolved) modalities: T1-MRI, DW-MRI

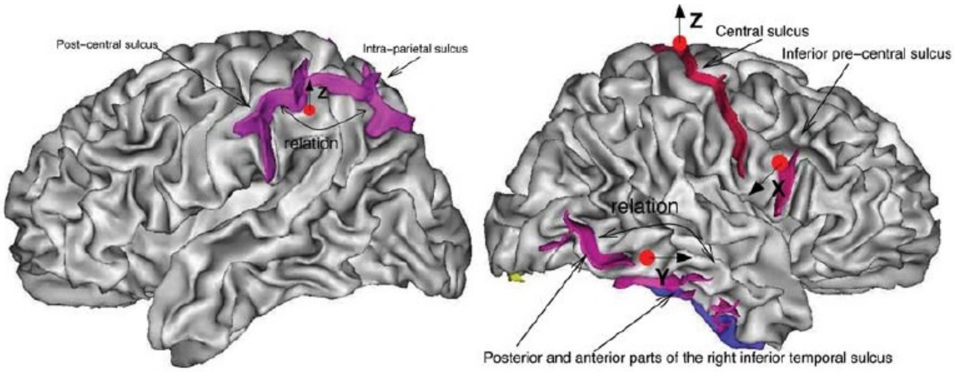
Neuroimaging modalities: T1 MRI

- T1 (1mm)^3 MRI yields
- Various measurements of brain structure
  - density of grey matter
  - Cortical thickness
  - Gyrification ratio
- Landmarks-based statistics
  - Sulcus shape/orientation
- 10^2 to 10^6 variables



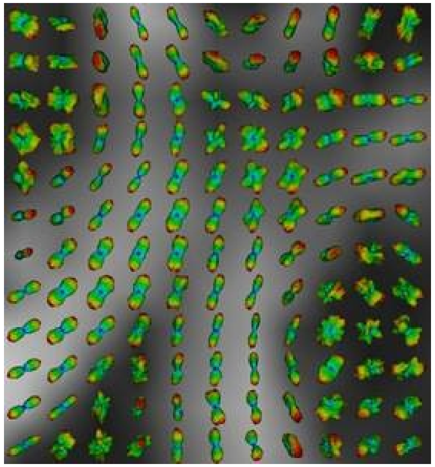


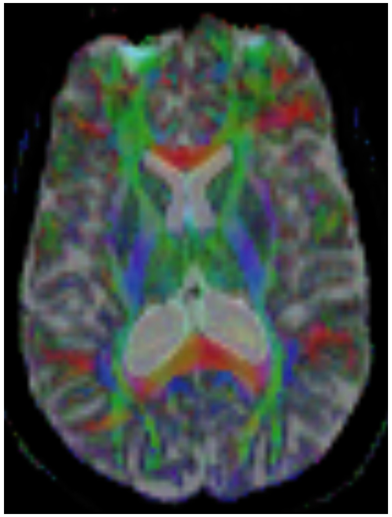


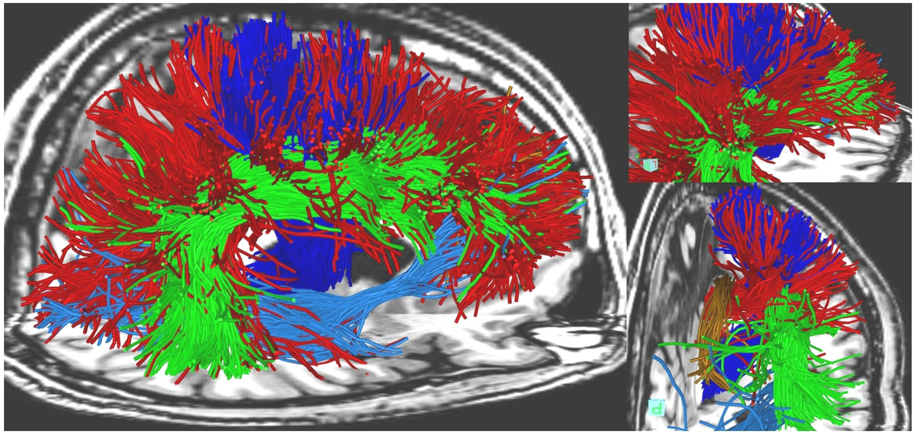


Neuroimaging modalities: DW-MRI

- Diffusion MRI: measurement of water diffusion in all directions in the white matter
- Resolution: $(2mm)^3$, 30-60 directions
- Yields the local direction of fiber bundles that connect brain regions
- *fibers/bundles* can be reconstructed through tractography algorithms
- Statistical measurement on bundles (counting, fractional anisotropy, direction)

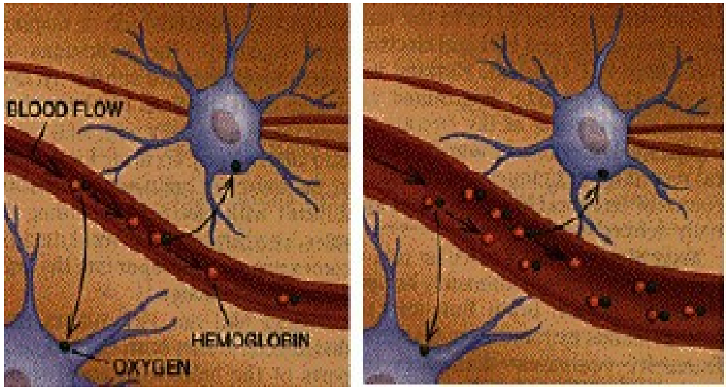


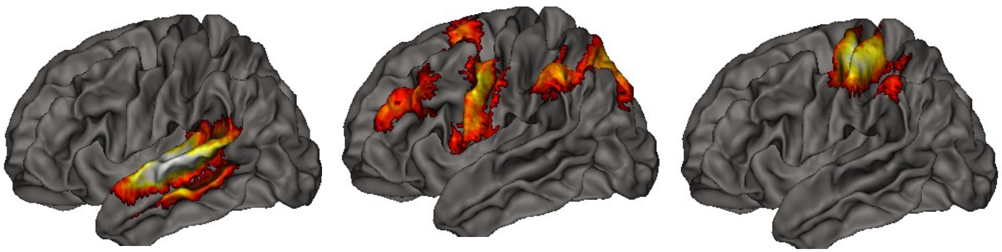


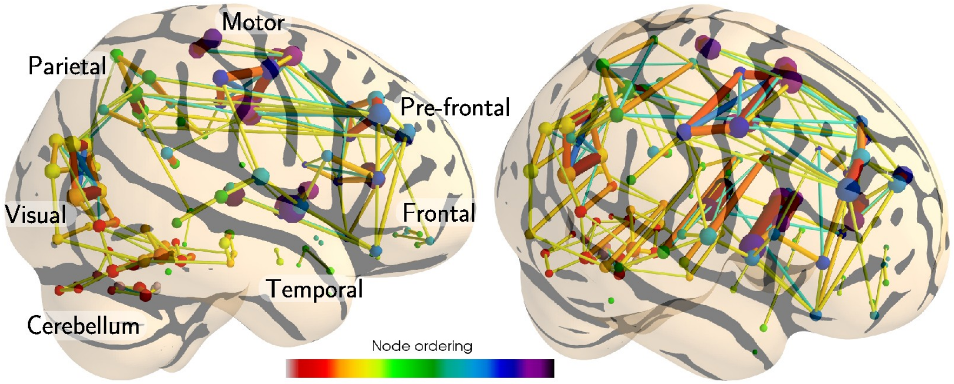


NeuroImaging modalities: fMRI

- BOLD signal: measures blood oxygenation in regions where synaptic activity occurs
  - Used to detect functionally specialized regions
  - But indirect measurement
  - Not a true quantitative measurement
- Can also be used to characterize network structure from brain signals
- 10^2 to 10^6 observations
- Resolution $(2\text{-}3mm)^3$, TR = 2-3s







NeuroImaging: modalities and aims

- Provide some biomarkers for diagnosis/prognosis, study of risk factors for various brain diseases
  - Psychiatric diseases
  - Neuro-degenerative diseases,
  - Brain lesions (strokes...)
- Understand brain organization and related factors: brain mapping, connectivity, architecture, development, aging, relation to behavior, relation to genetics
- Study chronometry of brain processes (EEG, MEG)
- Build brain computer interfaces (EEG)

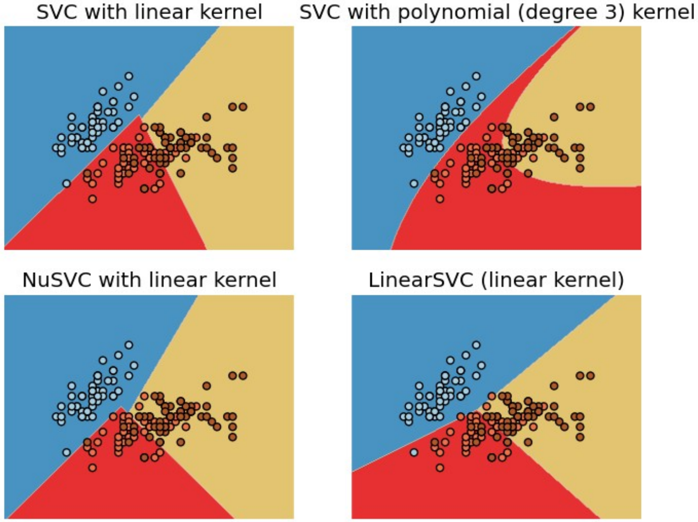
Technical challenges in MLNI

- Low SNR in the data
  - Only a fraction of the data is modeled (BOLD)
  - Presence of structured noise (noise is not i.i.d. Gaussian !) + non-stationarity in time and space
  - Few salient structures (resting-state fMRI...)
- Size of the data
  - 10^4 to 10^6 voxels in most settings
  - Compared to 10 to 10^2 samples available
- Related to the particular learning problems

Technical challenges in MLNI

- *Diagnosis/classification* problems
  - Needs accuracy mostly (+ robustness)
  - Suffers from curse of dimensionality, but this is well addressed in the literature: generic approaches perform well
  - But: not the main aim of most neuroimaging studies

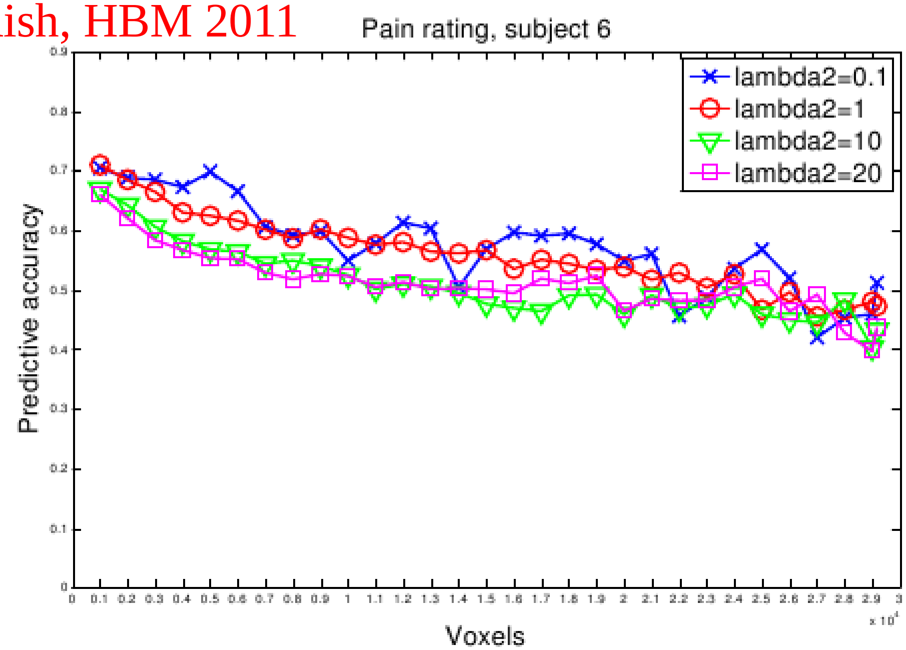
- Need a large set of tools to be compared against each other
- Need to take into account some priors on the data/true model (smoothness, sparsity)



Technical challenges in MLNI

- Recovery: retrieve the true model that accounts for the data
  - This is the main topic of all neuroimaging / brain mapping / decoding literature.
  - Suffers much more from feature dimensionality and correlation
  - Virtually in-addressed/unseen so far

I. Rish, HBM 2011



1. learn EN model for pain perception rating using first 120 TRs for training and next 120 TRs for testing.
2. Find 'best-predicting' 1000 voxels using EN, delete them, find next 1000 best-predicting, etc.

Does the predictive accuracy degrade sharply?

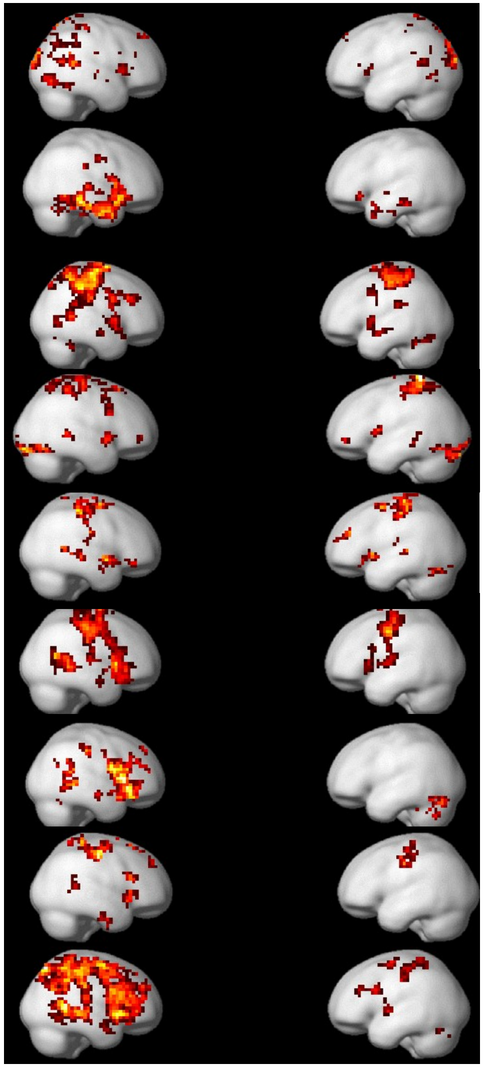
Surprisingly, the answer is 'NO'

Outline

- Machine Learning in Neuroimaging
    - Overview
    - Common technical challenges
- Some learning problems in neuroimaging:
  - Medical diagnosis/study of between subject-variability
  - Brain reading
  - Brain connectivity mapping

Study of between-subject variability

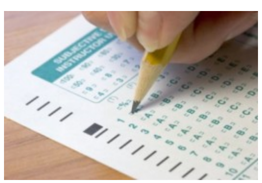
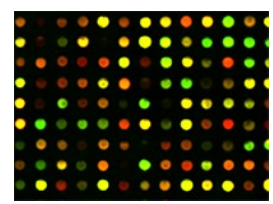
- Between-subject variability is a **prominent effect** in neuroimaging:
  - hard to characterize as such
  - how much of it can be explained using other data ?
- Brain diseases are extreme case of *normal variability*
- Data easier to acquire on *normal* populations
  - Confrontation to behavioral data
  - Confrontation to genetic data
- Perspective of individualized treatments



Study of between-subject variability

- Sometimes handled as unsupervised problems: describe the density of the data based on observations (manifold learning, mixture modeling)
- The major challenge here is to discover statistical associations between complex, high-dimensional variables (regression)

phenotype ← image   image ← genetic

$p($  $|$  $)$ $p($  $|$  $)$

Image→Phenotype   Gene→Image

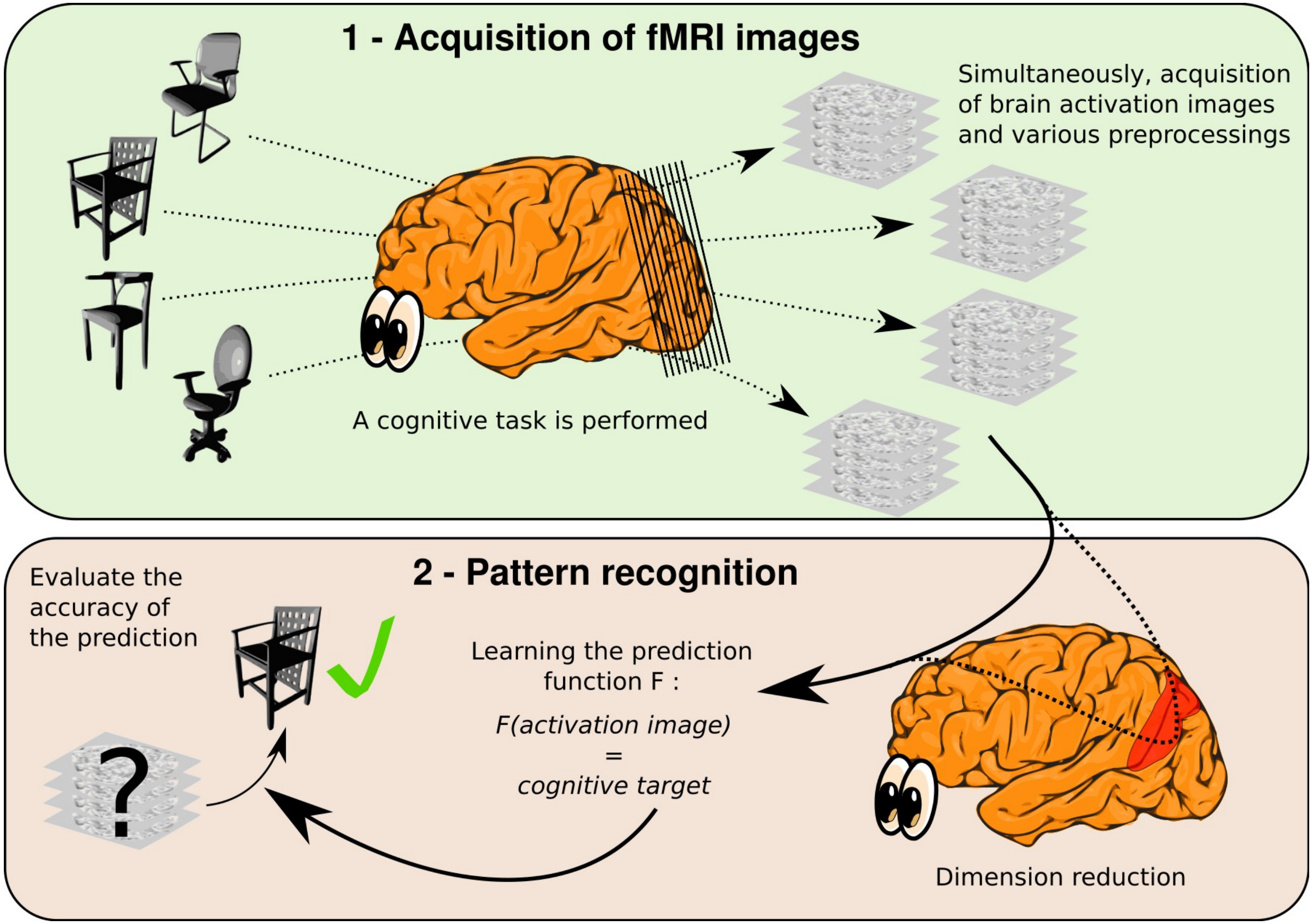
Imaging as an **intermediate (endo)phenotype**

- HPC
- Multiple comparison
- recovery

“Brain reading”

- Definition: Use of functional neuroimaging data to infer the subject's behaviour – typically the brain response related to a certain stimulus
- Similar to BCI -to some extent-
    - without time constraints
    - More emphasis on model correctness
- Popular due to its sensitivity to detect small-amplitude but distributed brain responses
- Rationale: population coding

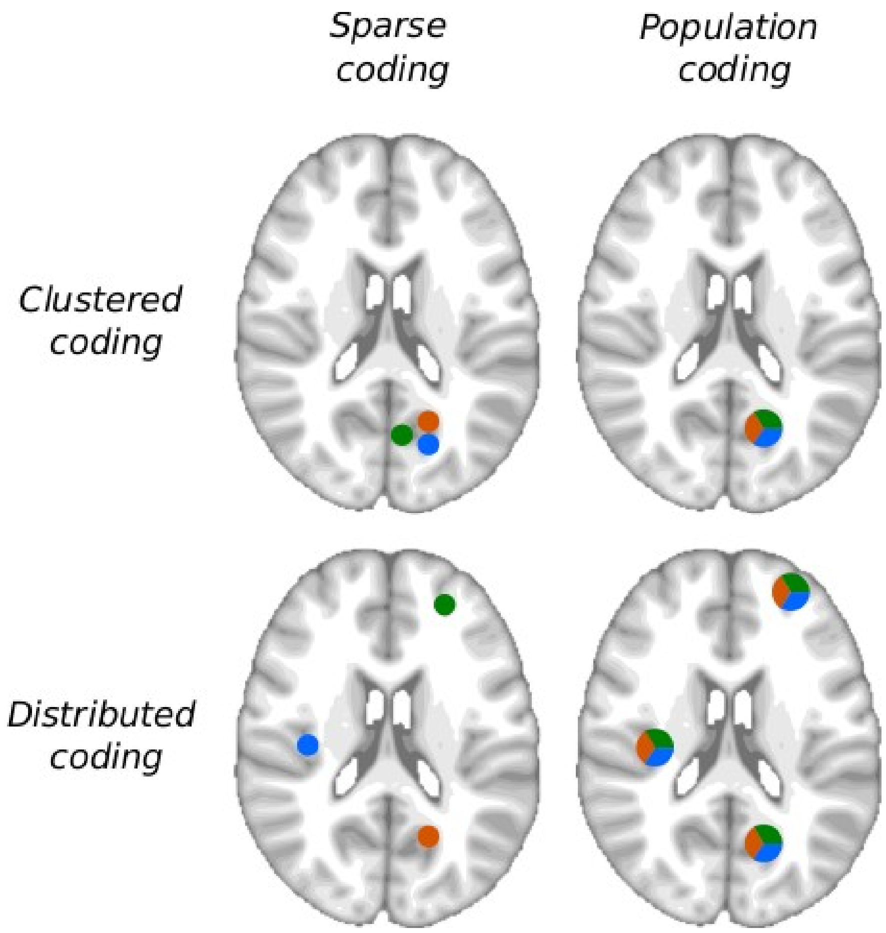
Brain reading / Reverse inference



Aims at predicting a cognitive variable → decoding brain activity
[Dehaene et al. 1998, Cox et al. 2003]

Brain reading: population coding

Different spatial models of the functional organization of neural networks

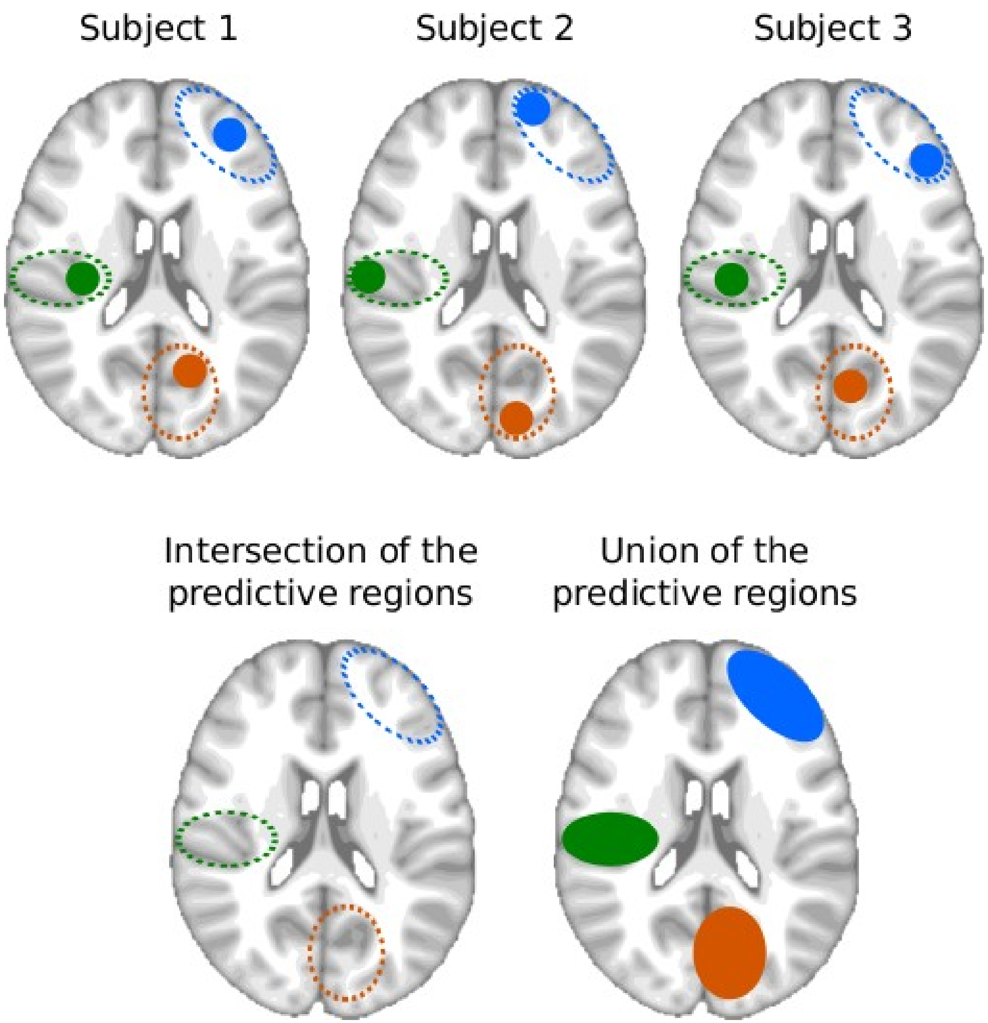


- Not a unique kind of pattern for the spatial organization of the neural code.
- This is further confounded by between-subject variability

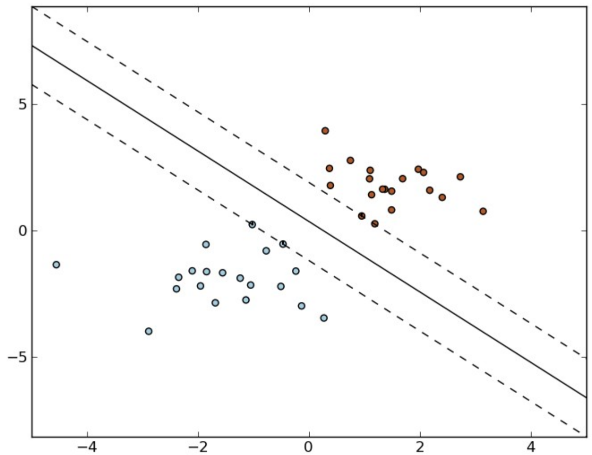
Inter-subject variability

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

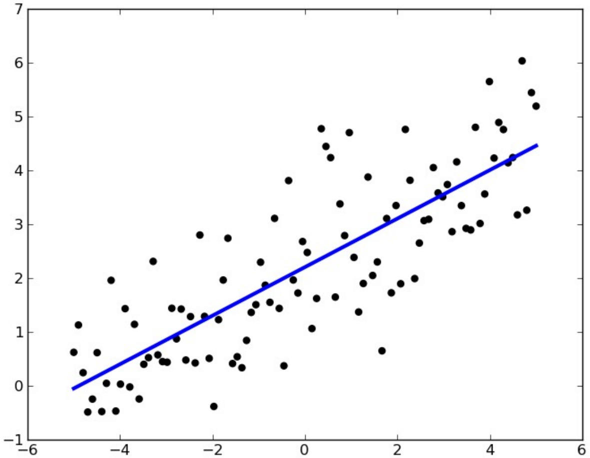
[Tucholka 2010]



Prediction function



$y = f(X, w, b) = X w + b$ or $\mathrm{sign}(X w + b)$



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

$p \gg n$
Curse of dimensionality
Risk of overfit

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 loading.

    [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.

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 ϵ 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**

      $\mathbf{z}_{old} = \mathbf{z}$

      $\mathbf{z} =$

      $\Pi_K\left(\mathbf{z}_{aux} - \frac{1}{\lambda \tilde{L}}\mathrm{grad}(L\mathbf{u} + \lambda \mathrm{div}(\mathbf{z}_{aux}))\right)$

      $t_{old} = t$

      $t = (t + \sqrt{1 + 4t^2})/2$

      $\mathbf{z}_{aux} = \mathbf{z} + \frac{t_{old} - 1}{t}(\mathbf{z} - \mathbf{z}_{old})$

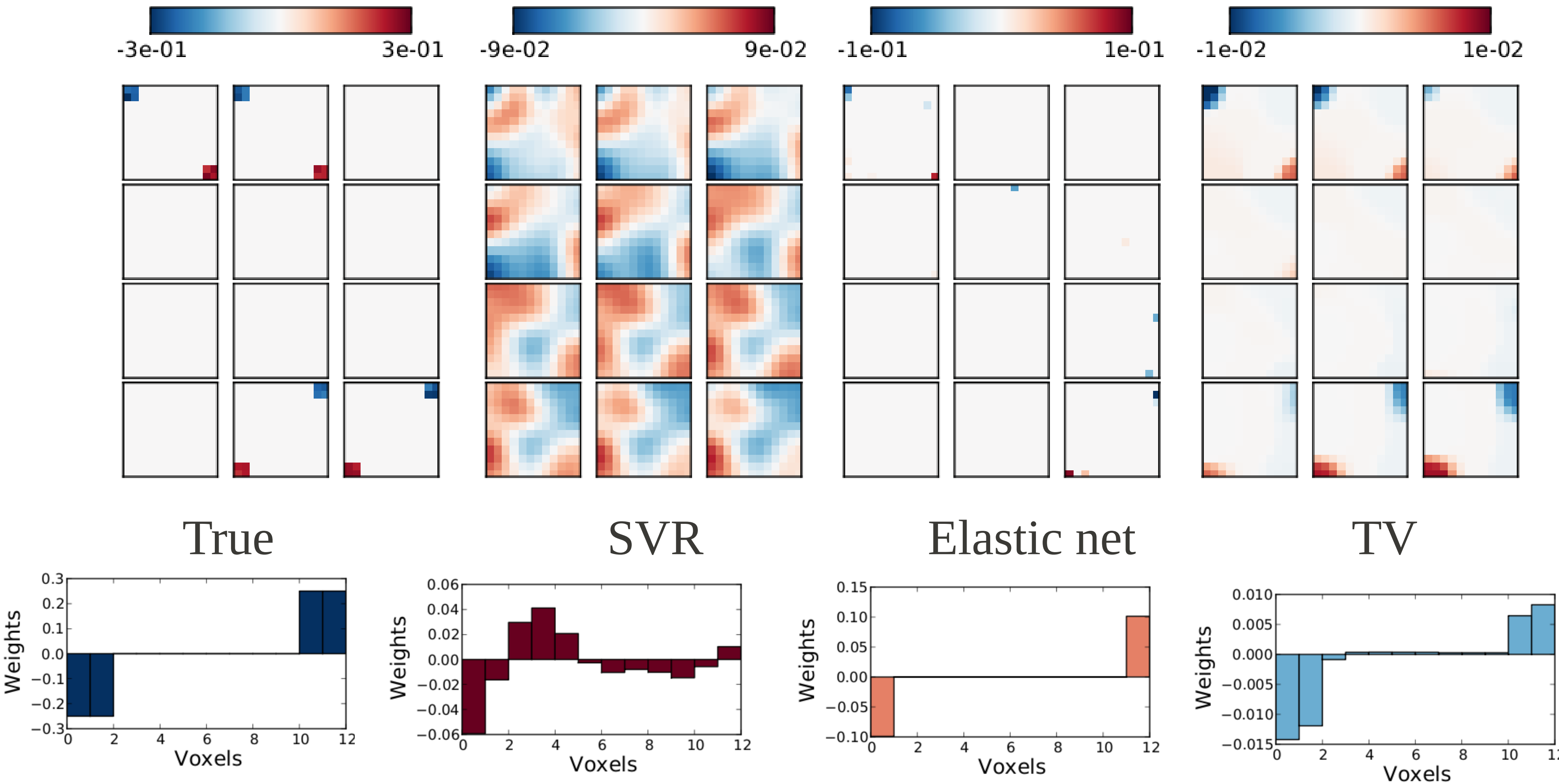
    end while

    $\mathbf{w} = \mathbf{u} + \lambda \mathrm{div}(\mathbf{z})$

  end for

  return $\mathbf{w}$

Intuition on simulated data



→ extract weights with a sparse block structure.

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]

Prediction accuracy on inter-subject analyzes

Regression analysis

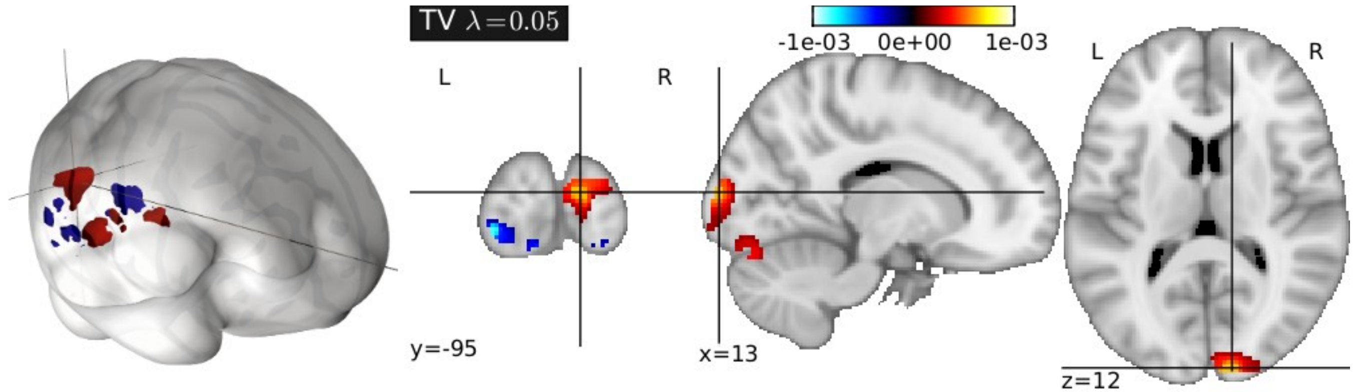
Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.77	0.11	0.97	0.58	0.0277 **
Elastic net	0.78	0.1	0.97	0.65	0.0405 **
TV $\lambda = 0.05$	0.84	0.07	0.97	0.72	-

Classification analysis

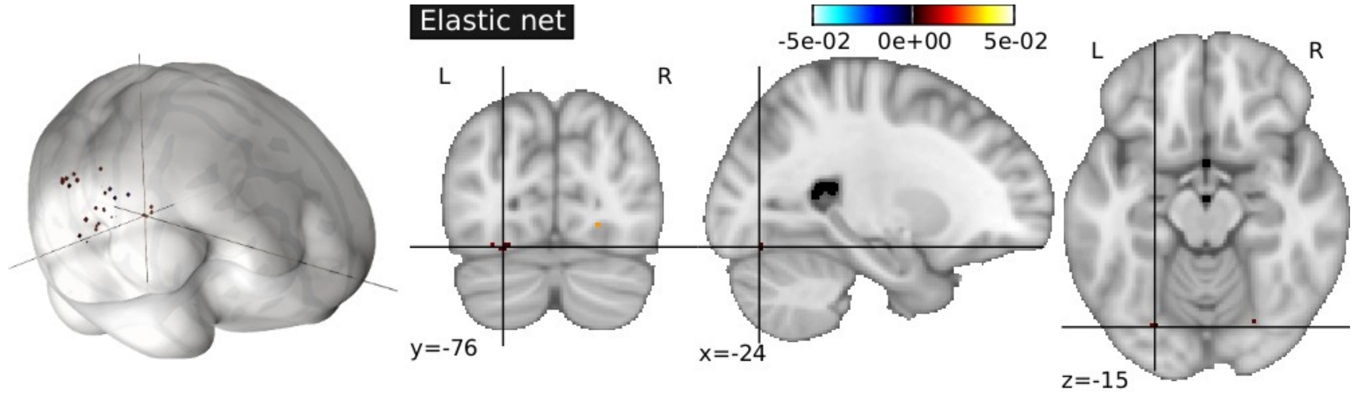
Methods	mean κ	std κ	max κ	min κ	p-value to SVC
SVC	48.33	15.72	75.0	25.0	-
SMLR	42.5	9.46	58.33	33.33	0.2419
TV $\lambda = 0.05$	45.83	14.55	66.67	25.0	0.7128

TV → maps for brain mapping

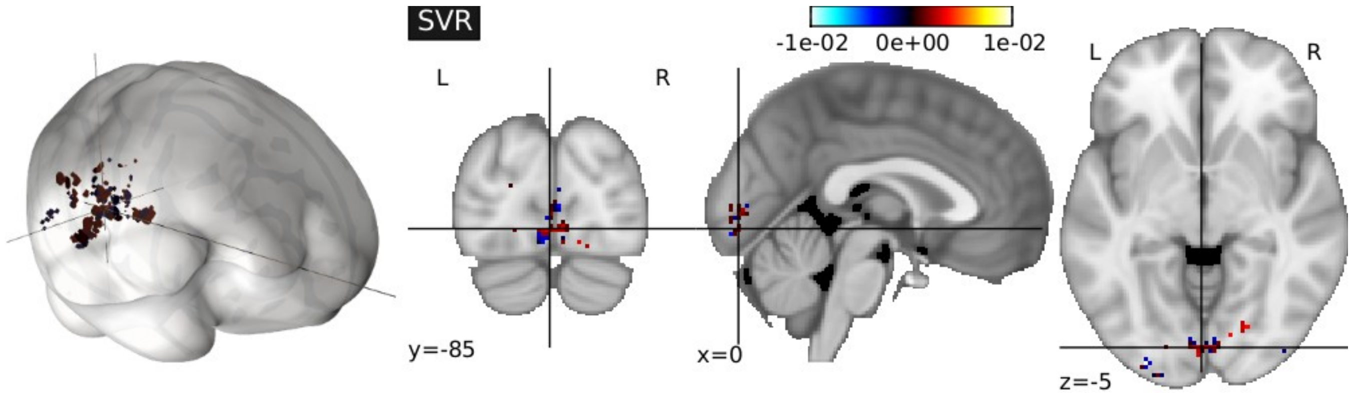
TV



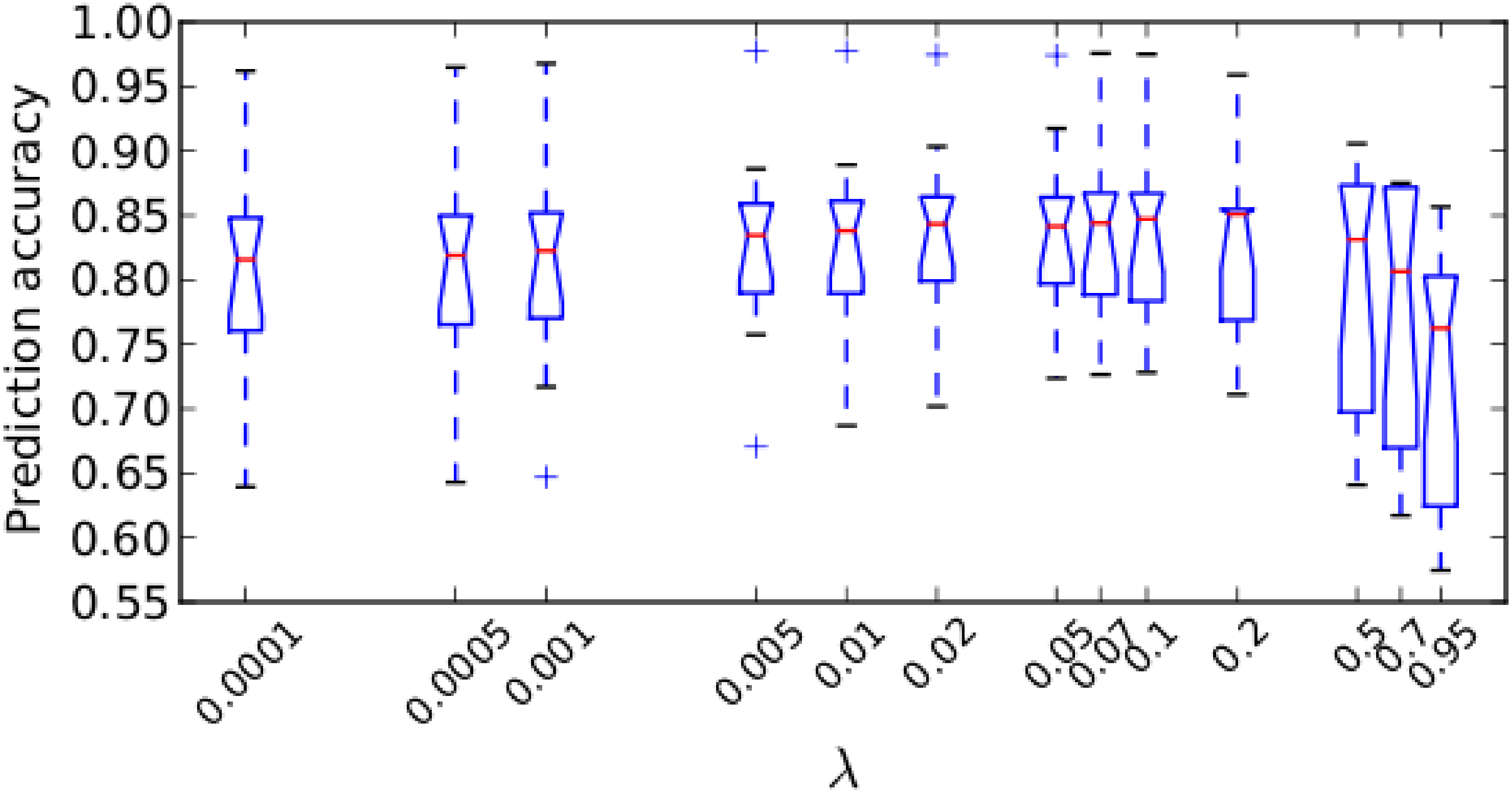
Elastic net



SVR



Influence of the regularization parameter λ



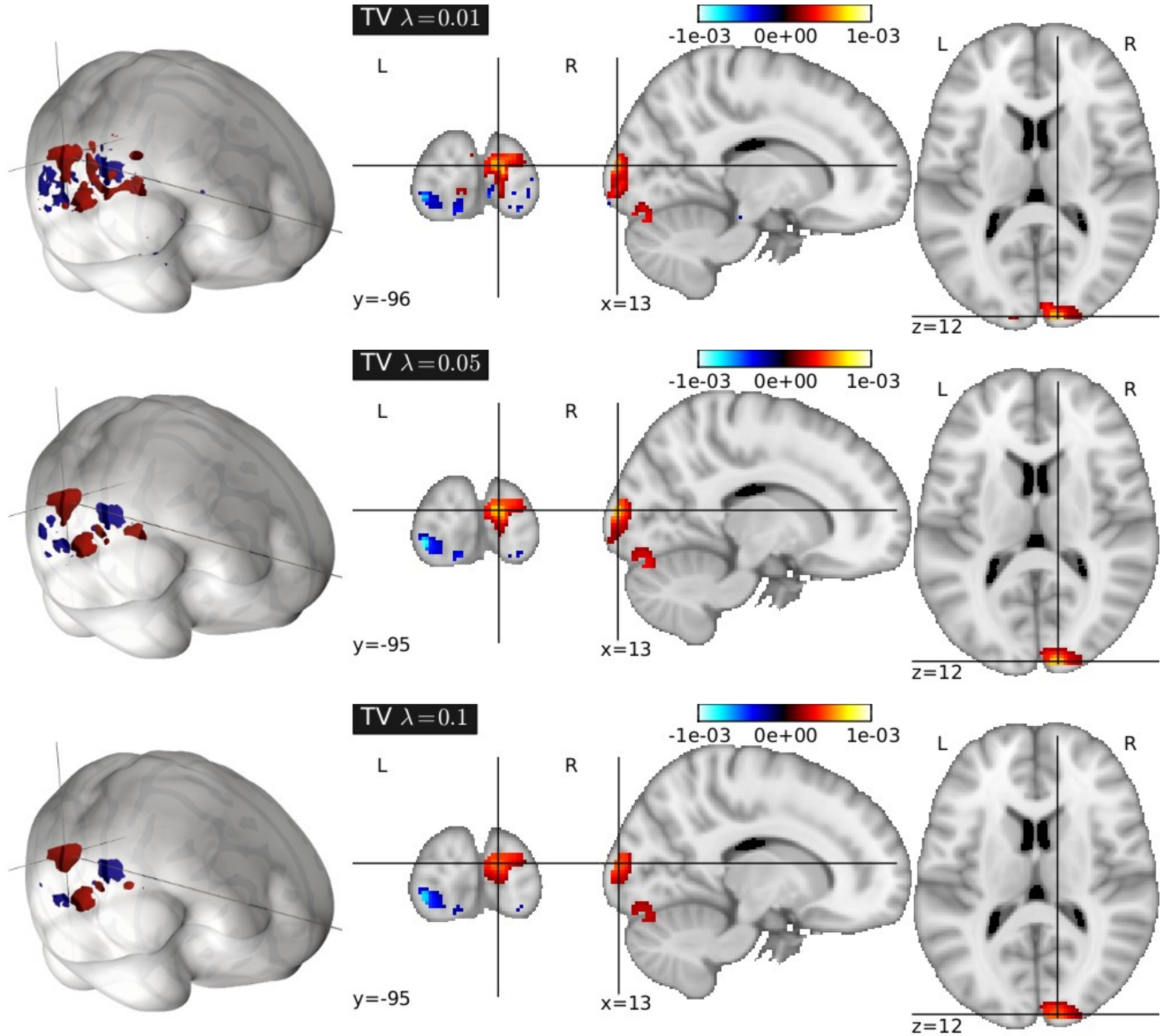
→ results are extremely stable with respect to λ.

Influence of the regularization parameter λ

λ = 0.01
ζ = 0.83

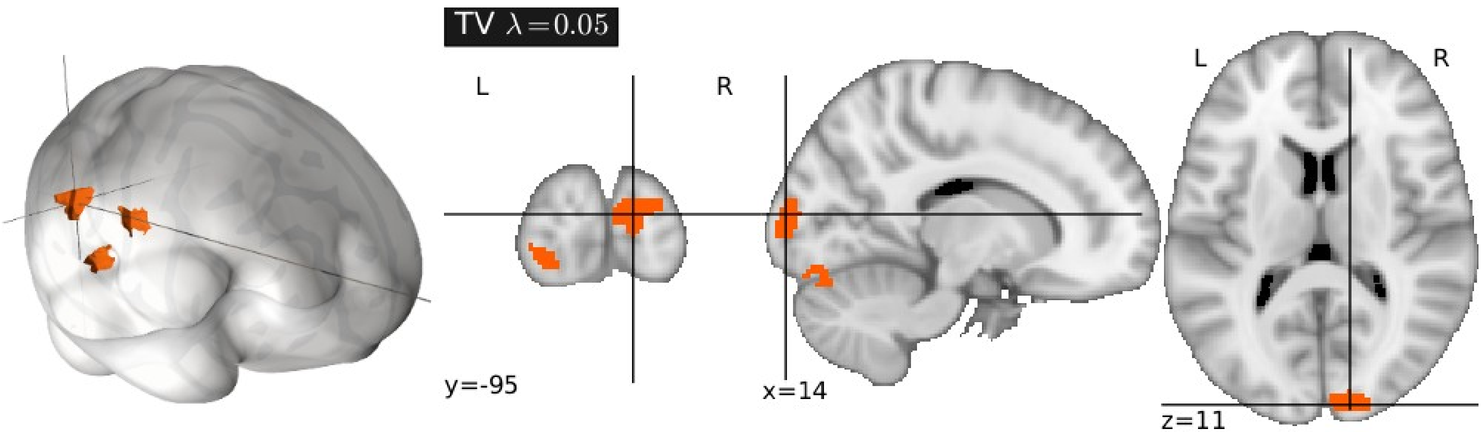
λ = 0.05
ζ = 0.84

λ = 0.1
ζ = 0.84

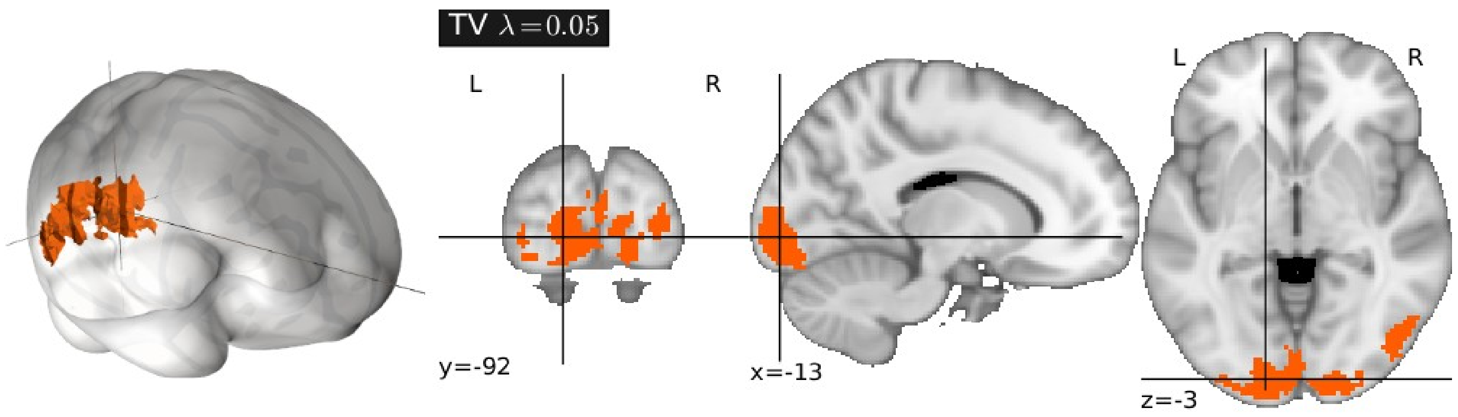


TV for fMRI-based decoding

Inter-subject regression analysis.



Inter-subject classification analysis.



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

Conclusion on TV regularization

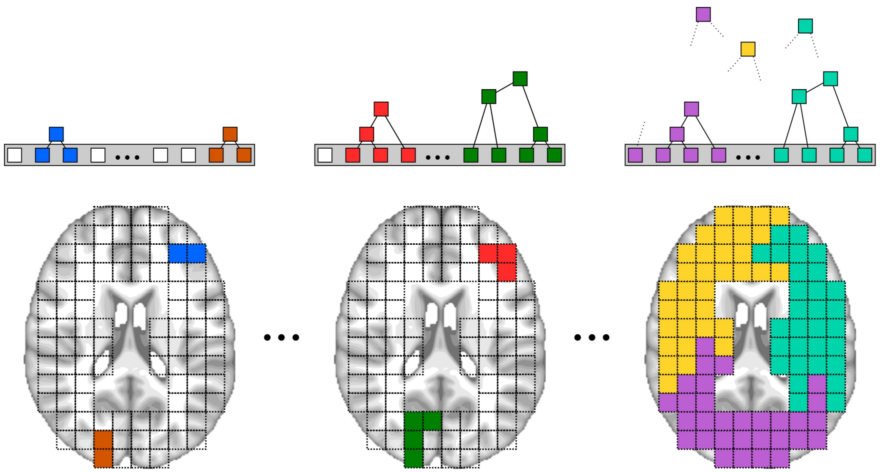
First use of TV for prediction problem (classification/regression).

- 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*. IEEE Transactions on Medical Imaging, 2011, 30 (7), pp. 1328 – 1340.

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

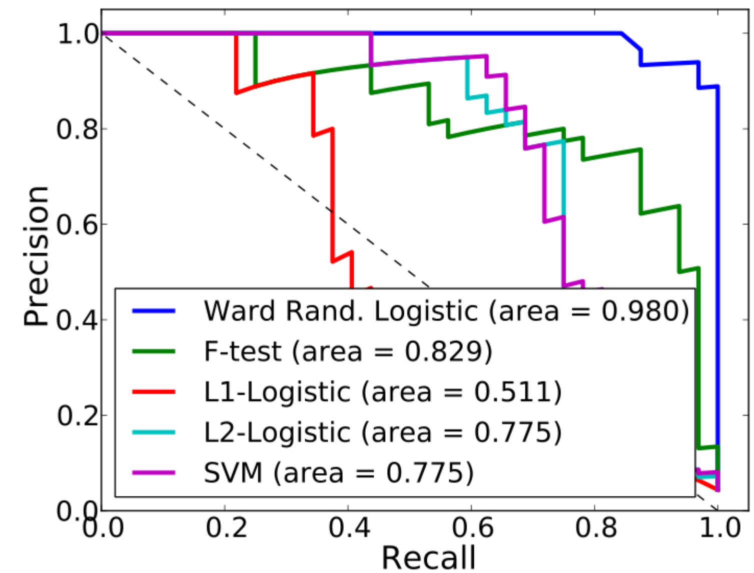
$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 = \sum_{g \in \mathcal{G}} \Big[\sum_{j \in g} \mathbf{w}_j^2\Big]^{1/2}$$

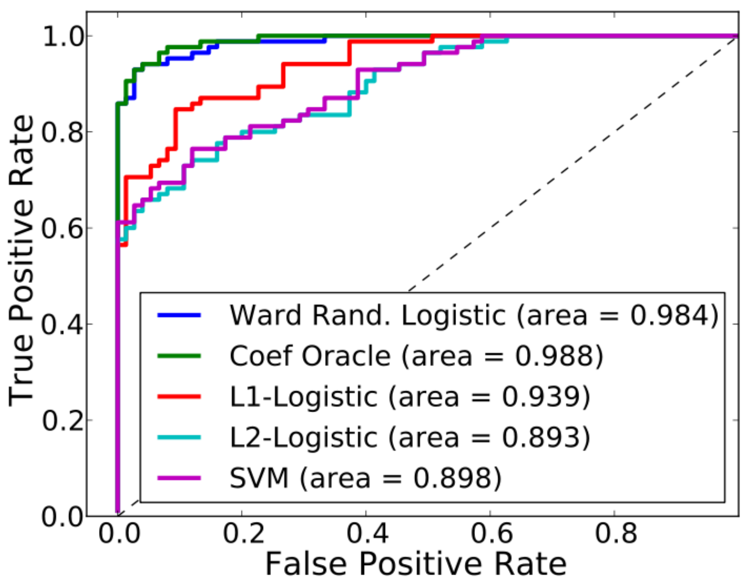
- Acts as the l_1-norm on the vector $(\|\mathbf{w}_g\|_2)_{g \in \mathcal{G}}$
- If one node is set to 0 , its descendants are also set to 0
- Consider large parcels before small parcels → robustness to spatial variability

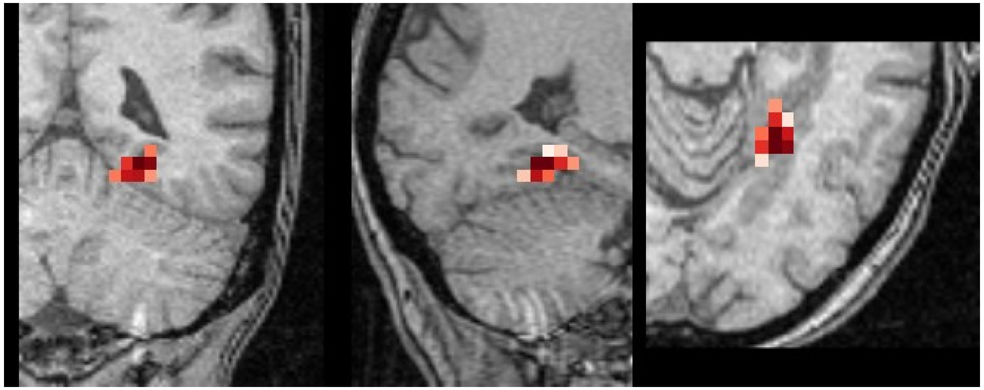
[Michel et al. Pattern Recognition 2011]
[Jenatton et al PRNI 2011, subm to SIAM imaging]

Dealing with the recovery issue

- Recovery: retrieve the true model that accounts for the data
  - Use of stability selection (randomized lasso on bootstrapped data)
  - adaptive brain parcellations (Ward's algorithm)
  - yields high accuracy and good recovery on simulations







Gramfort et al., MLINI 2011

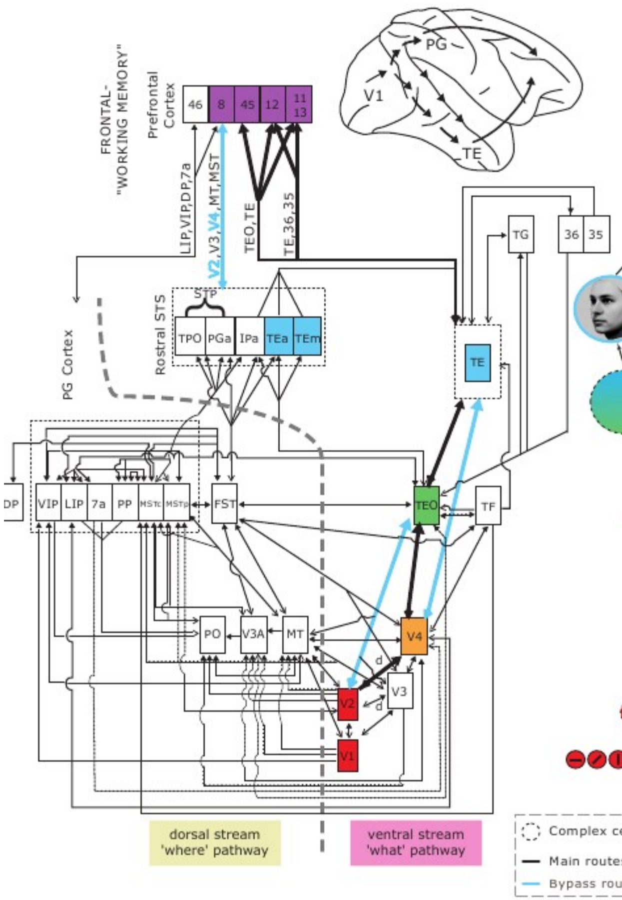
Brain reading / open issues

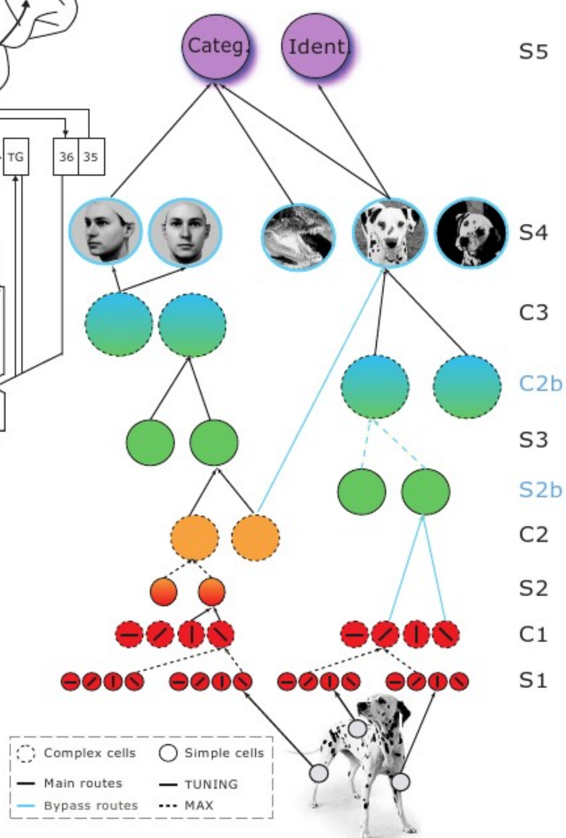
Do we want this....

```
### Compute the prediction accuracy for the different folds (i.e. session)
cv_scores = cross_val_score(anova_svc, X, y, cv=cv, n_jobs=-1,
                            verbose=1, iid=True)

### Return the corresponding mean prediction accuracy
classification_accuracy = np.sum(cv_scores) / float(n_samples)
```

```
>>> print "Classification accuracy: %f" % classification_accuracy, \
    " / Chance level: %f" % (1. / n_conditions)
Classification accuracy: 0.744213  / Chance level: 0.125000
```

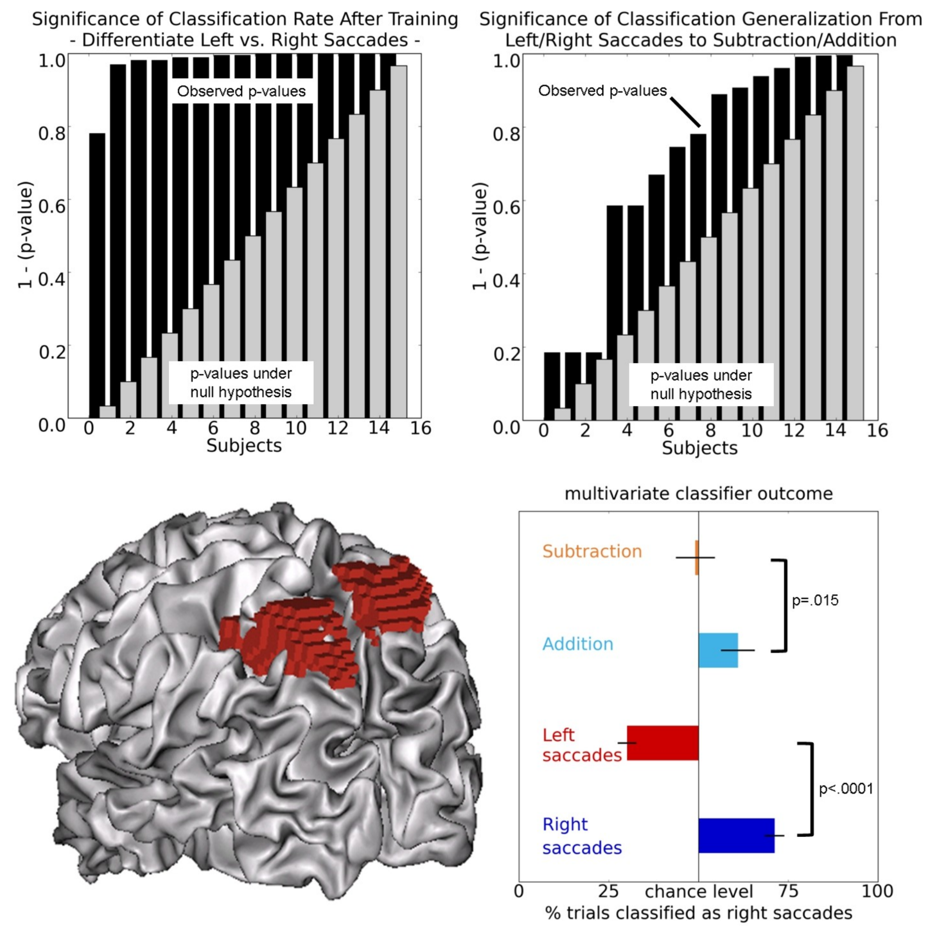




… or that ?

Brain reading: Transfer learning

- a classifier trained to discriminate left versus right saccades can also *decode* mental arithmetics:
- subtraction ⇔ left saccade
- addition ⇔ right saccade
- This generalization occurs only when based on two regions of the parietal cortex
- This shows that the same neural populations are involved in ocular saccades and arithmetics



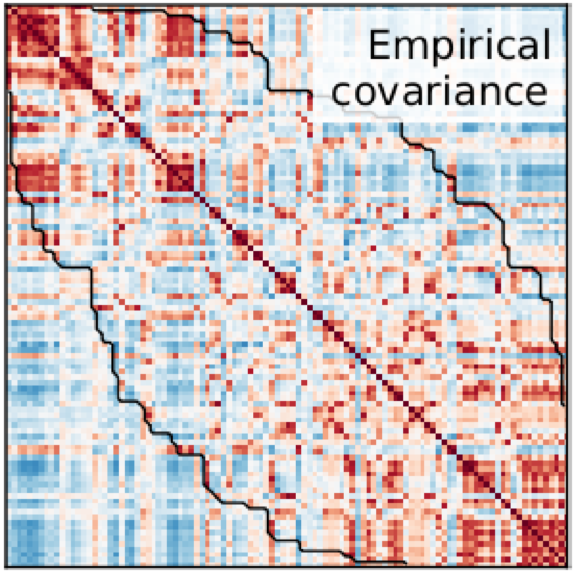
[Knops et al., *science* 2009]

Outline

- Machine Learning in Neuroimaging
    - Overview
    - Common technical challenges
- Some learning problems in neuroimaging:
  - Medical diagnosis/study of between subject-variability
  - Brain reading
  - Brain connectivity mapping

Functional connectivity mapping

- Definition: consists in deriving a quantitative measure of brain networks integration based on functional neuroimaging correlations

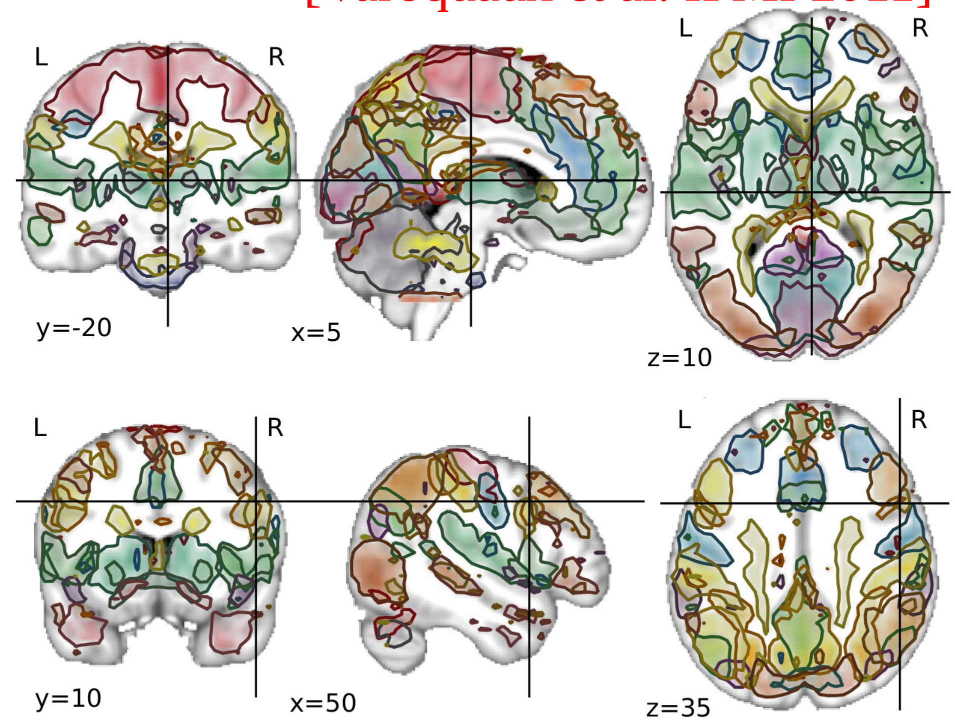


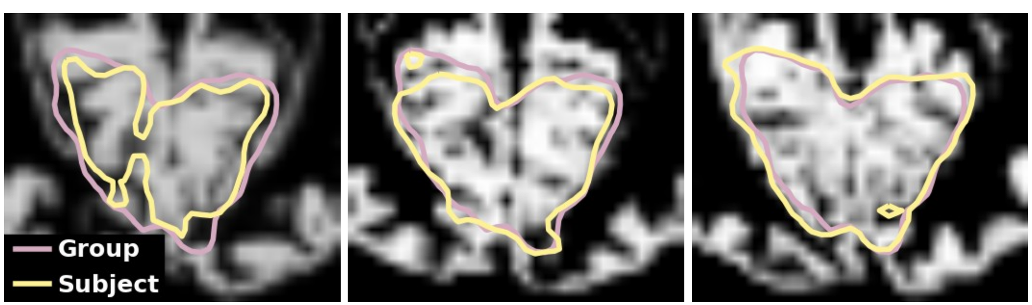
- Rationale
  - Popularity of resting-state fMRI.
  - Model-driven approach (SEM, DCM) do not scale well
- Learning problems
  - Segment regions based on observed correlations (common to many neuroimaging problems)
  - Inference of graphical models

Learning in FCM (1)

- Learn a spatial model (atlas) from the resting state data
  - ICA, clustering provide little guarantees on the result
  - Dictionary learning (SSPCA) can be used instead

[Varoquaux et al. IPMI 2011]





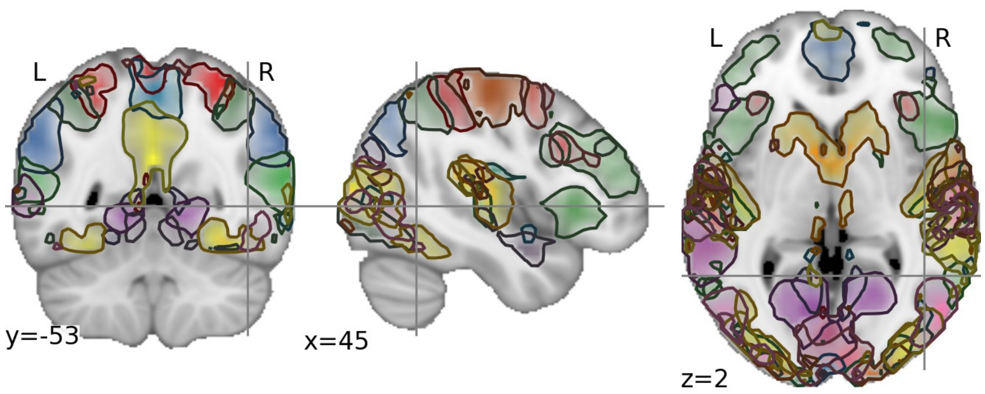
The population-level model adapts to individual configurations

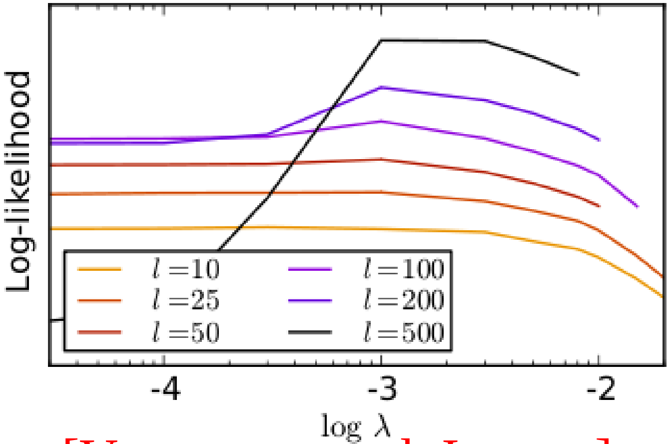
$$(\mathbf{U}^s, \mathbf{V}^s)_{s \in \{1 \dots S\},}, \mathbf{V} = \underset{\mathbf{U}^s, \mathbf{V}^s, \mathbf{V}}{\operatorname{argmin}} \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V}), \qquad \text{s.t. } \|\mathbf{u}_l^s\|_2^2 \leq 1$$

$$\text{with} \quad \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V}) = \sum_{s=1}^{S} \frac{1}{2} \left(\|\mathbf{Y}^s - \mathbf{U}^s \mathbf{V}^{sT}\|_{\text{Fro}}^2 + \mu \|\mathbf{V}^s - \mathbf{V}\|_{\text{Fro}}^2 \right) + \lambda \, \Omega(\mathbf{V}),$$

Toward large-scale brain atlases

- More generally learn brain functional atlases from the data...
  - requires lots of data
  - Could be the first serious attempt to map brain space to brain function
  - Requires learning methods that scale with huge datasets
    - online dictionary learning
  - Model selection is tricky

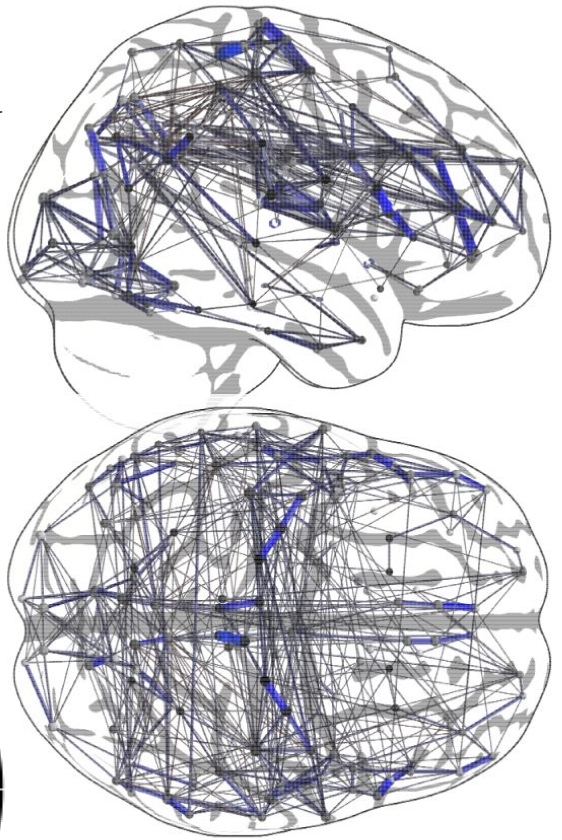


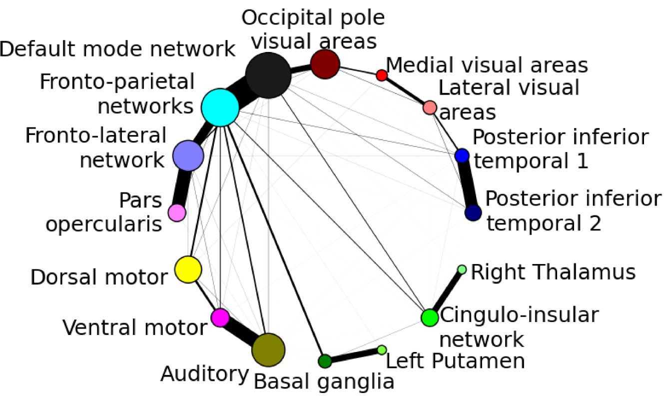


[Varoquaux et al. In prep]

Learning in FCM (2)

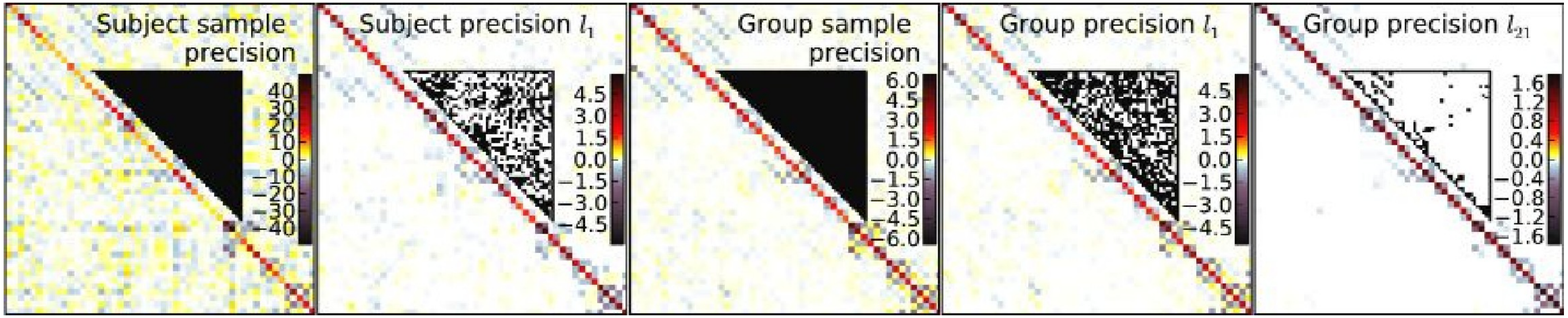
- Next: Given a set of regions, quantify properly their interactions/integration of the underlying networks





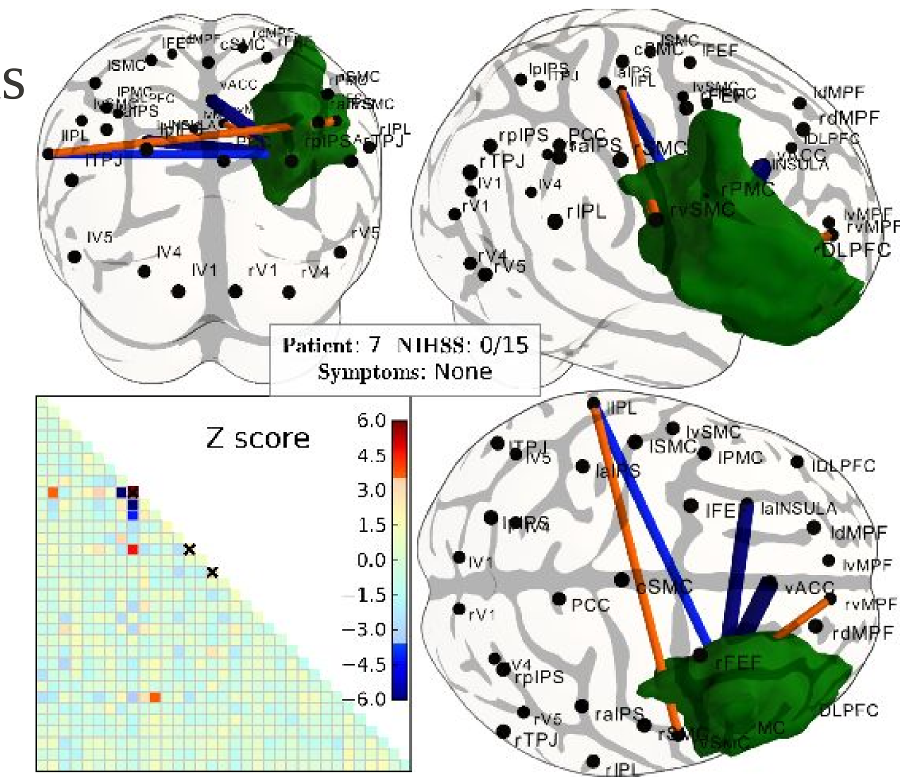
  - Learn covariance model between the set of regions (partial correlations)
  - Group- sparse- penalty

$$\left(\hat{\mathbf{K}}^{(s)}_{\ell_{21}}\right)_{s=1..S} = \operatorname{argmin}_{\mathbf{K}^{(s)}\succ 0}\left(\sum_{s=1}^{S}\left(\operatorname{tr}(\mathbf{K}^{(s)}\,\hat{\mathbf{\Sigma}}^{(s)}_{\text{sample}}) - \log\det\mathbf{K}^{(s)}\right) + \lambda\sum_{i\neq j}\|\mathbf{K}^{(\cdot)}_{ij}\|_2\right)$$



Learning in FCM (3)

- Do statistical inference on these objects:localize the differences in the graph structure between two populations
- Example: stroke patients
- Problem: covariance matrices live on a manifold; computing statistics (mean, variance) is challenging
- Our solution so far: linearize the variability model, assuming small differences



$$\mathbf{\Sigma}^s = \phi^{-1}_{\mathbf{\Sigma}^\star}(\mathbf{d\Sigma}^s) = \mathbf{\Sigma}^{\star\,\frac{1}{2}} \exp(\mathbf{d\Sigma}^s)\, \mathbf{\Sigma}^{\star\,\frac{1}{2}}$$

$$\mathbf{\Sigma}^s \simeq \mathbf{\Sigma}^{\star\,\frac{1}{2}} (\mathbf{I}_n + \mathbf{d\Sigma}^s)\, \mathbf{\Sigma}^{\star\,\frac{1}{2}}$$

Conclusion

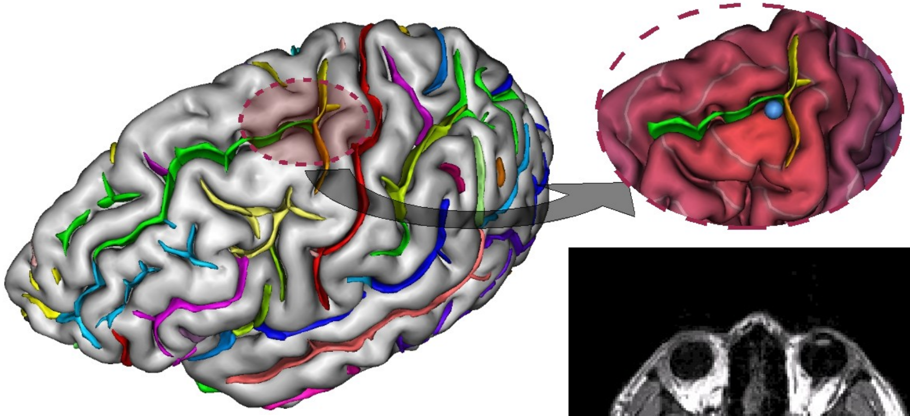
- Machine learning in Neuroimaging
  - standard challenges (but lack of data)
  - Need guarantees on the result (e.g. support recovery)
  - Neuroimaging people also need guidelines
- At INRIA
  - Fruitful & long-term collaborations with Select and Sierra
  - Other ongoing projects (MEG, BCI) → more impact
- Implementation matters:
  - the success of many methods is related to their availability (libsvm !)
  - Computation time is important in practice

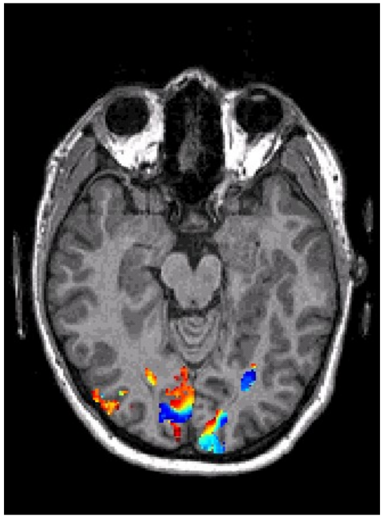
scikits learn
machine learning in Python

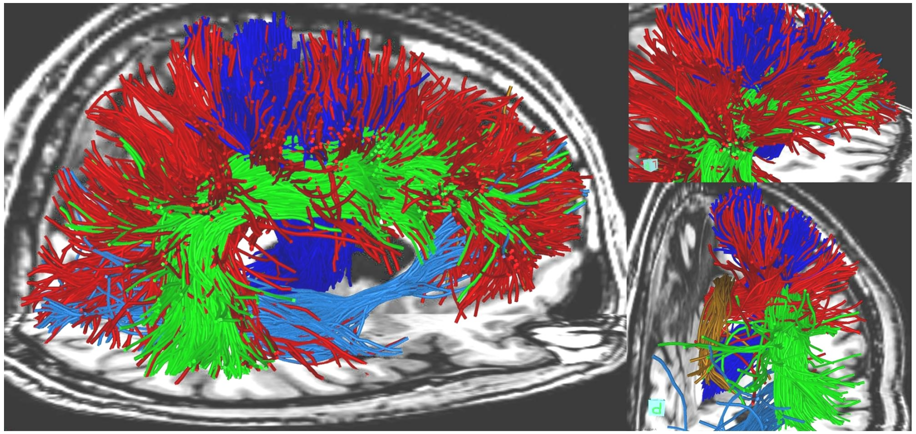
Acknowledgements

- 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