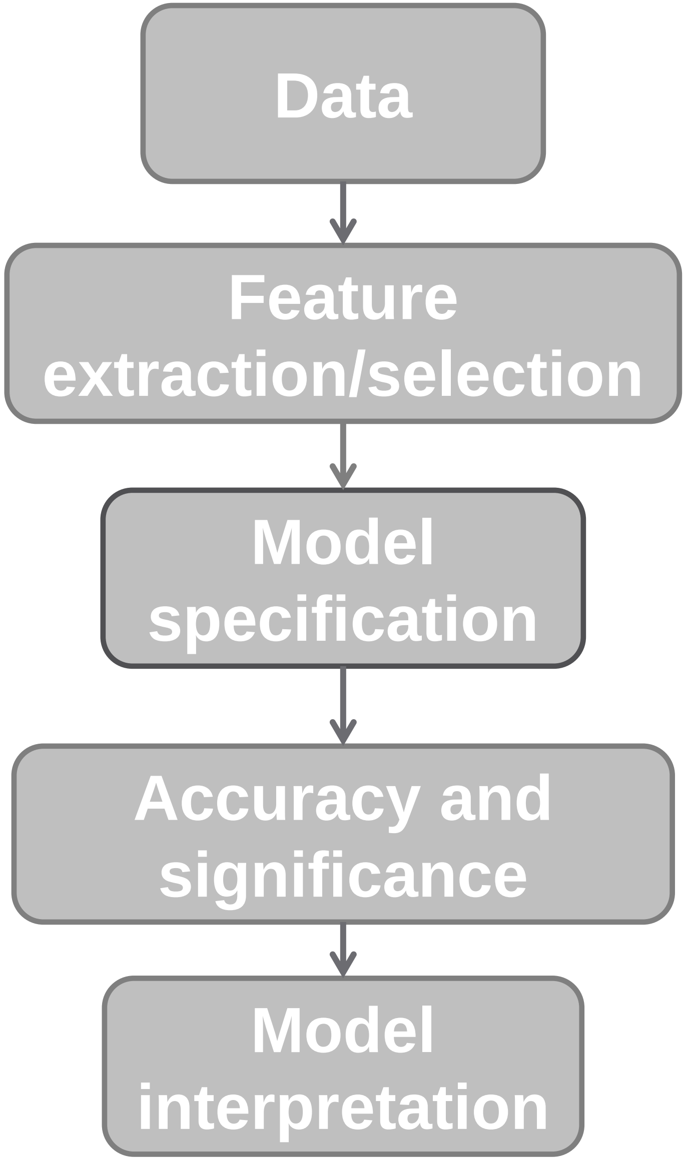
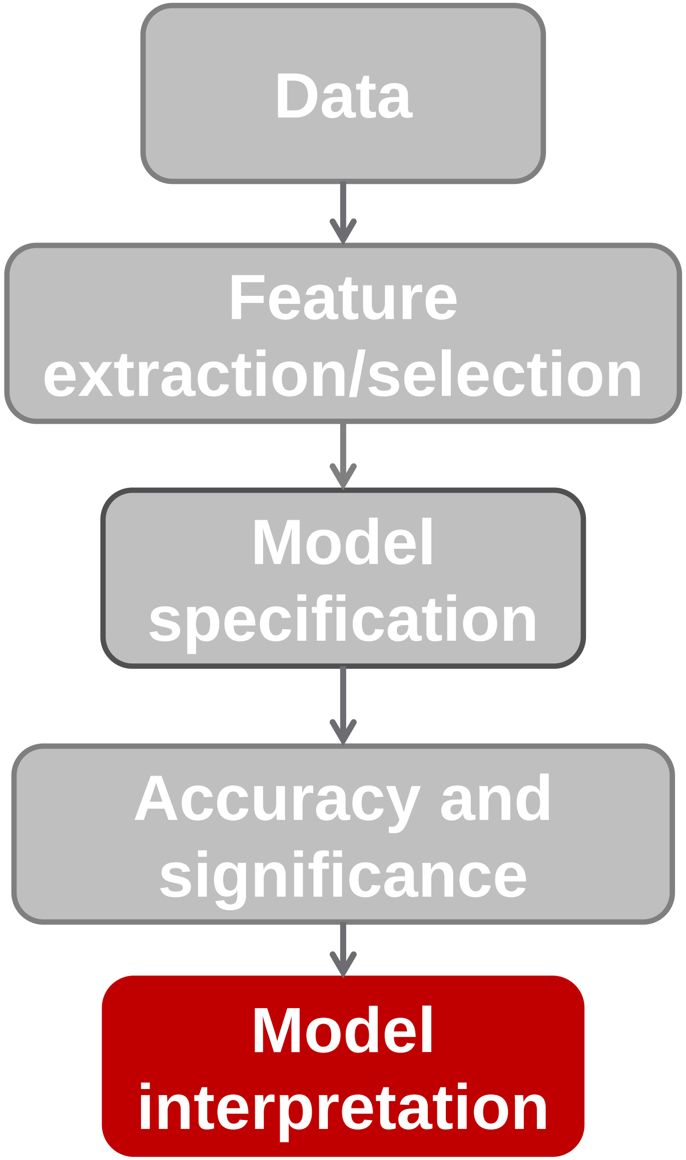
Pattern Recognition for NeuroImaging (PR4NI)

# Interpreting predictive models in terms of anatomically labelled regions

J. Schrouff
Stanford University, CA, USA


Data
Feature extraction/selection
Model specification
Accuracy and significance
Model interpretation




Where in the brain is the information of interest?

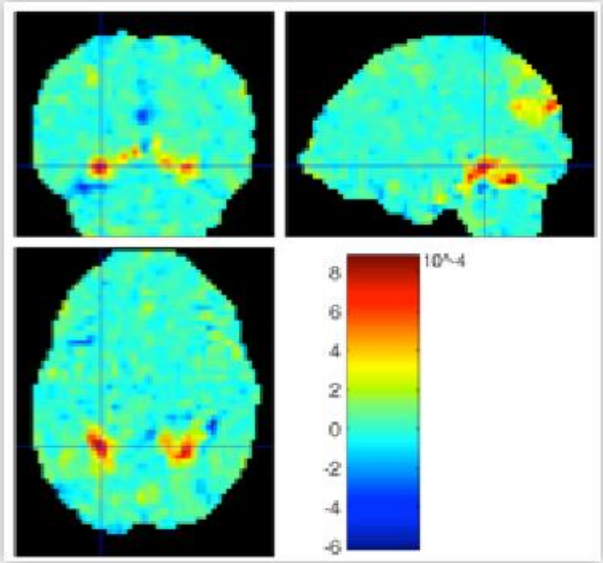
i.e. Which brain areas contribute most to the model's prediction?

Computing “weight maps”:

- From linear (kernel) methods
- Attributing one weight per feature

- For an SVM model:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i^* \ \mathbf{x_i}$$

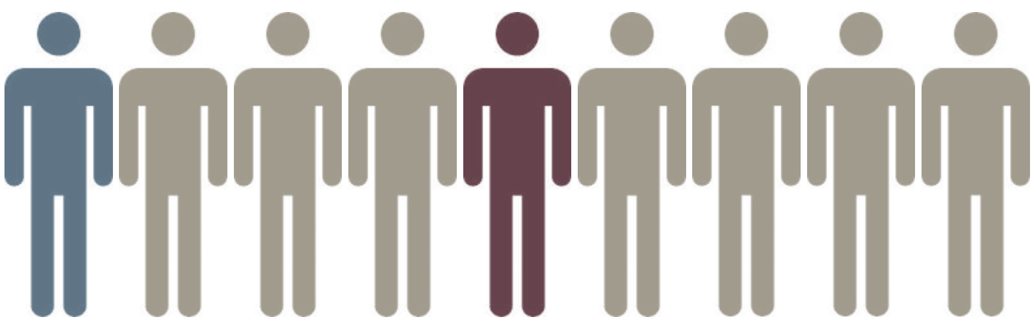


Thresholding weight maps:

*Can we choose to look only at the largest contributions?*

- Linear combination
- As in an election
- Each vote counts

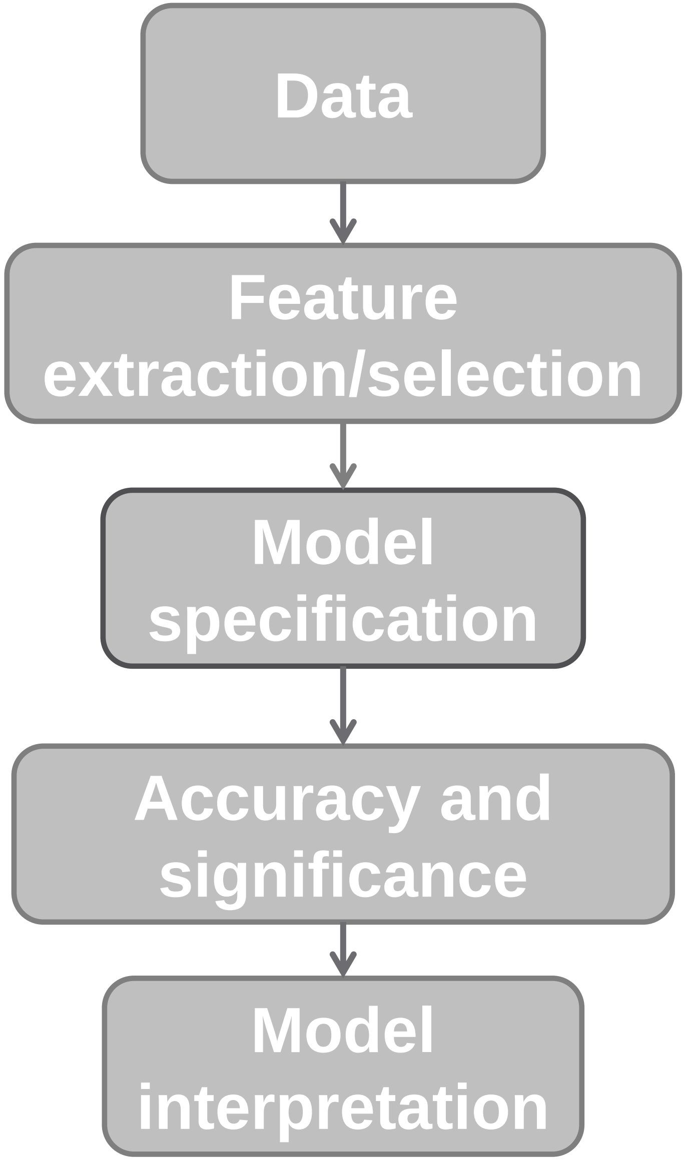
$$f(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x_i} \rangle + b$$



10 1 1 1 8 1 1 1 1

Different strategies:

1. A priori knowledge: ROI selection
2. Searchlight
3. Feature selection
4. Sparse algorithms
5. Multiple Kernel Learning (MKL)
6. A posteriori knowledge: weight summarization




Different strategies:

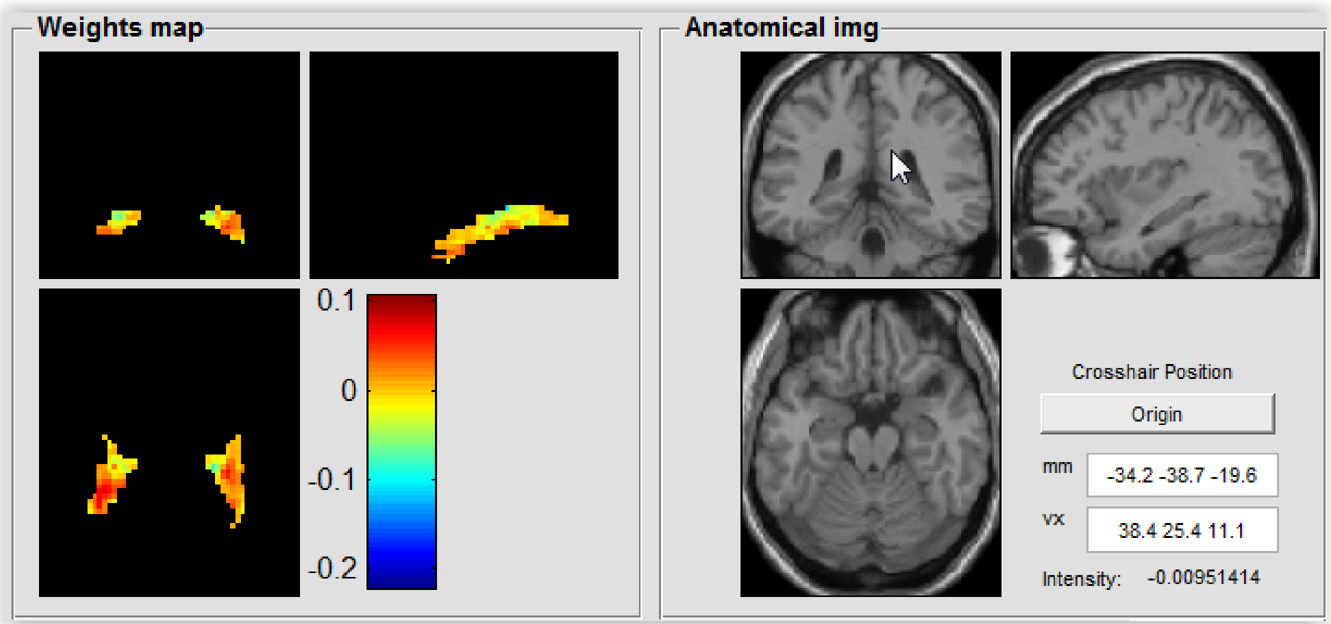
1. A priori knowledge: ROI selection
2. Searchlight
3. Feature selection

4. Sparse algorithms
5. Multiple Kernel Learning (MKL)

6. A posteriori knowledge: weight summarization

1. A priori knowledge: ROI selection

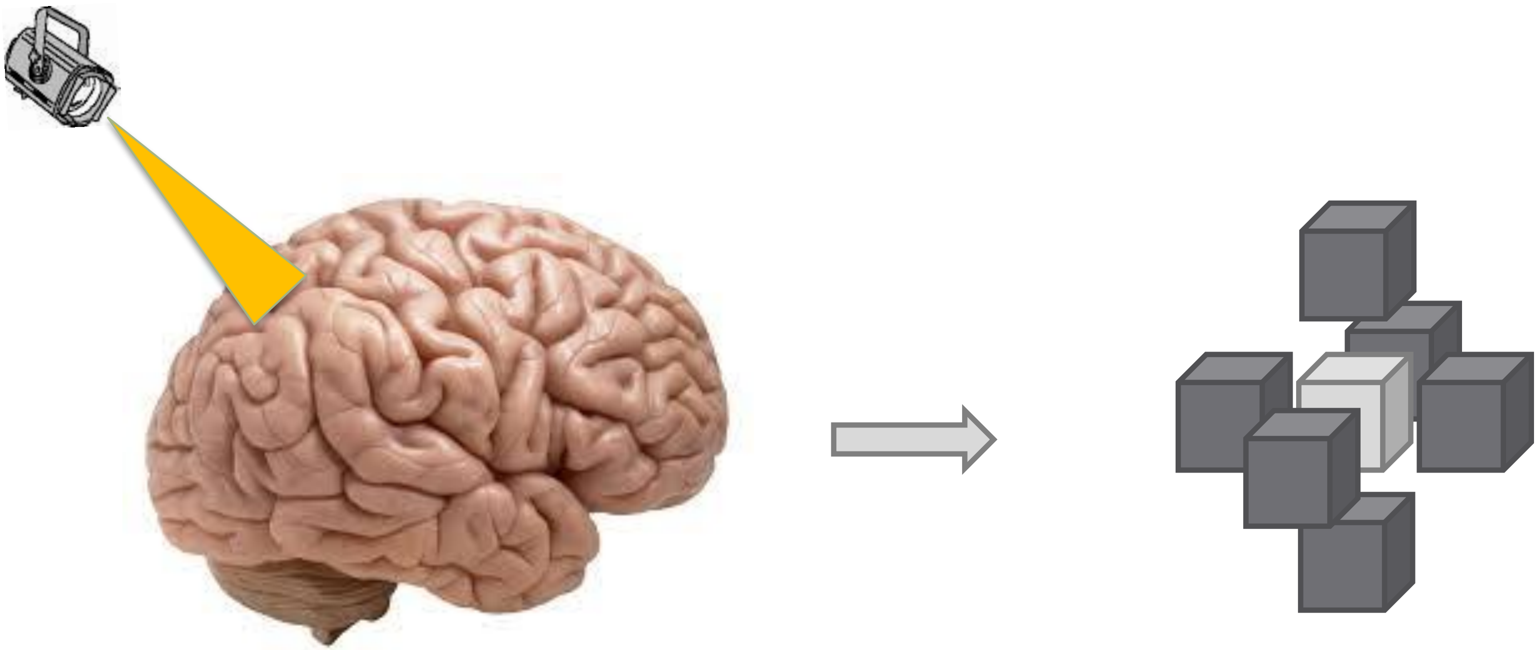
e.g. Choosing voxels/features based on literature or on previous univariate results (independent dataset).



1. A priori knowledge: ROI selection

| + | - |
|---|---|
| Good accuracy? | Need an anatomical hypothesis |
| Anatomically/ functionally relevant | Cannot threshold obtained map |
| | Arbitrary choice of region number and size |
| | Multiple models = multiple comparisons |

## 2. Searchlight (Kriegeskorte et al., 2006)



## 2. Searchlight (Kriegeskorte et al., 2006)

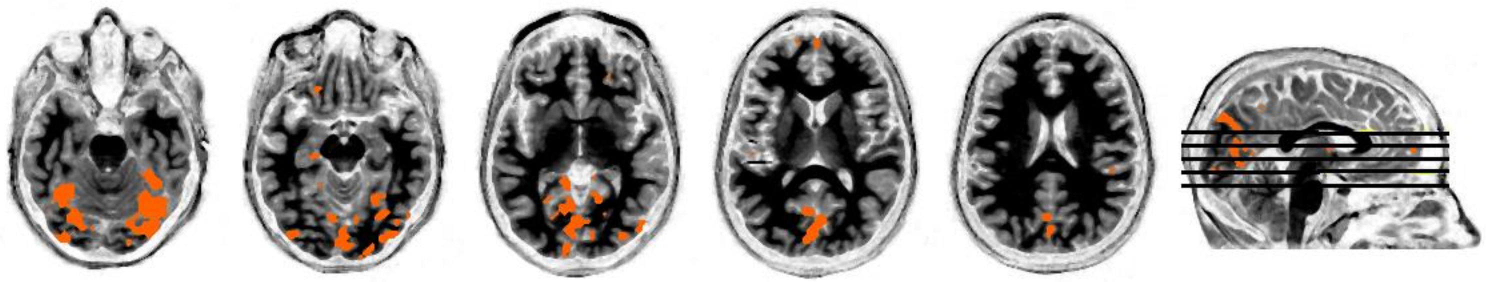
| + | - |
|---|---|
| Threshold for each neighborhood | Not anatomically/functionally relevant (spheres) |
| | Locally multivariate |
| | Arbitrary choice of neighborhood size |
| | Multiple models = multiple comparisons |
| | Time-consuming |

## 3. Feature selection

Rank the features according to a specific criterion and select a subset.

Example:

- Univariate t-test
- Recursive Feature Elimination (RFE, De Martino, 2008) based on SVM weights



## 3. Feature selection

| + | - |
|---|---|
| Data-driven | Not anatomically/functionally relevant (voxel sparse) |
| | Time-consuming: nested CV |
| | Depends on selected technique: flaws in feature selection are reflected in model. → user choice must be informed. |
| | Different strategies = different patterns with same accuracy. |

4. Sparse algorithms

Algorithms with regularization constraints such that the weight of some features is perfectly null.

Examples:

- LASSO (Tibshirani, 1996)
- Elastic-net (Zou and Hastie, 2005)

## 4. Sparse algorithms

| + | - |
|---|---|
| Data-driven | Not anatomically/functionally relevant (voxel sparse) |
| Embedded in model | Stability of selected voxels (Baldassarre, et al., 2012): different regularization terms = different patterns for same accuracy. |

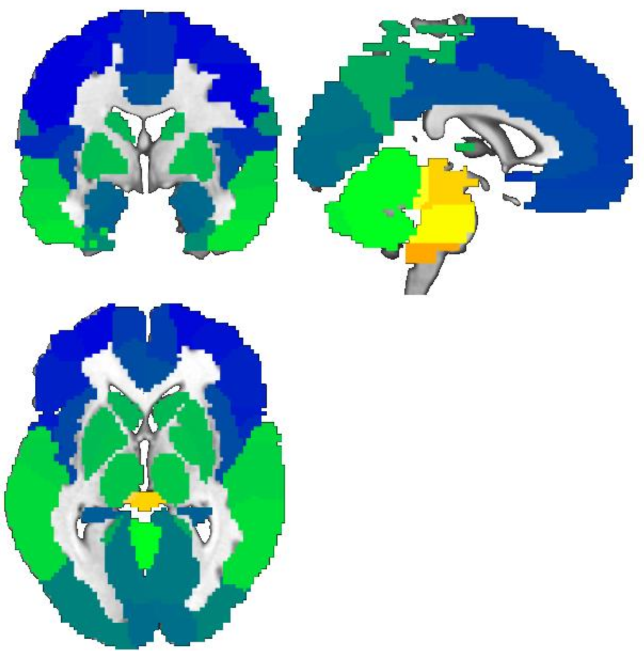
## 4. Sparse algorithms

| + | - |
|---|---|
| Data-driven | Not anatomically/functionally relevant |
| Embedded in model | Stability of selected voxels (Baldassarre, et al., 2012): different regularization terms = different patterns for same accuracy. |

Note: stability selection (Rondina et al., 2014).

## 5. Multiple Kernel Learning (MKL)

Build one kernel per anatomically defined region as input of a (sparse) MKL algorithm (Filippone, 2012).



$$K(\boldsymbol{x},\boldsymbol{x}') = \sum_{m=1}^{M} d_m K_m(\boldsymbol{x},\boldsymbol{x}')$$

with $d_m \geq 0, \ \sum_{m=1}^{M} d_m = 1$

## 5. Multiple Kernel Learning (MKL)

Decision function of an MKL problem:

$$f(\mathbf{x}_i) = \sum_m \langle \mathbf{w}_m, \mathbf{x}_\mathbf{i} \rangle + b$$

Considered algorithm (Rakotomamonjy et al., 2008):

$$\min \frac{1}{2} \sum_m \frac{1}{d_m} \|\mathbf{w}_m\|^2 + C \sum_i \xi_i$$

$$s.t. \quad y_i \left( \sum_m \langle \mathbf{w}_m, \mathbf{x}_\mathbf{i} \rangle + b \right) \geq 1 - \xi_i \qquad \forall i$$

$$\xi_i \geq 0 \qquad \forall i, \sum_m d_m = 1, \qquad d_m \geq 0 \quad \forall m$$

5. Multiple Kernel Learning (MKL)

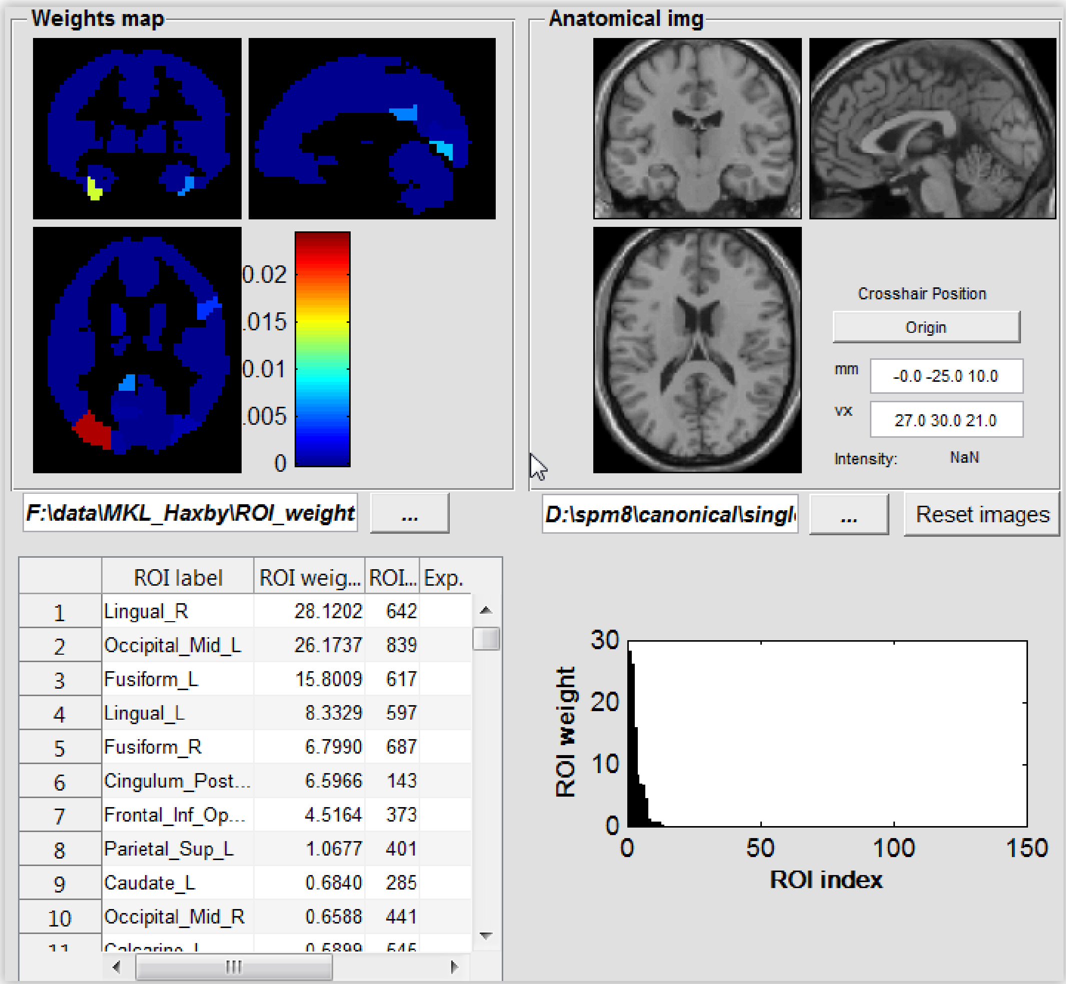
Weight map of an MKL model:

$$\mathbf{w}_{\boldsymbol{m}} = d_m \sum_{i=1}^{n} y_i \alpha_i \ \mathbf{x}_{\mathbf{i}}$$

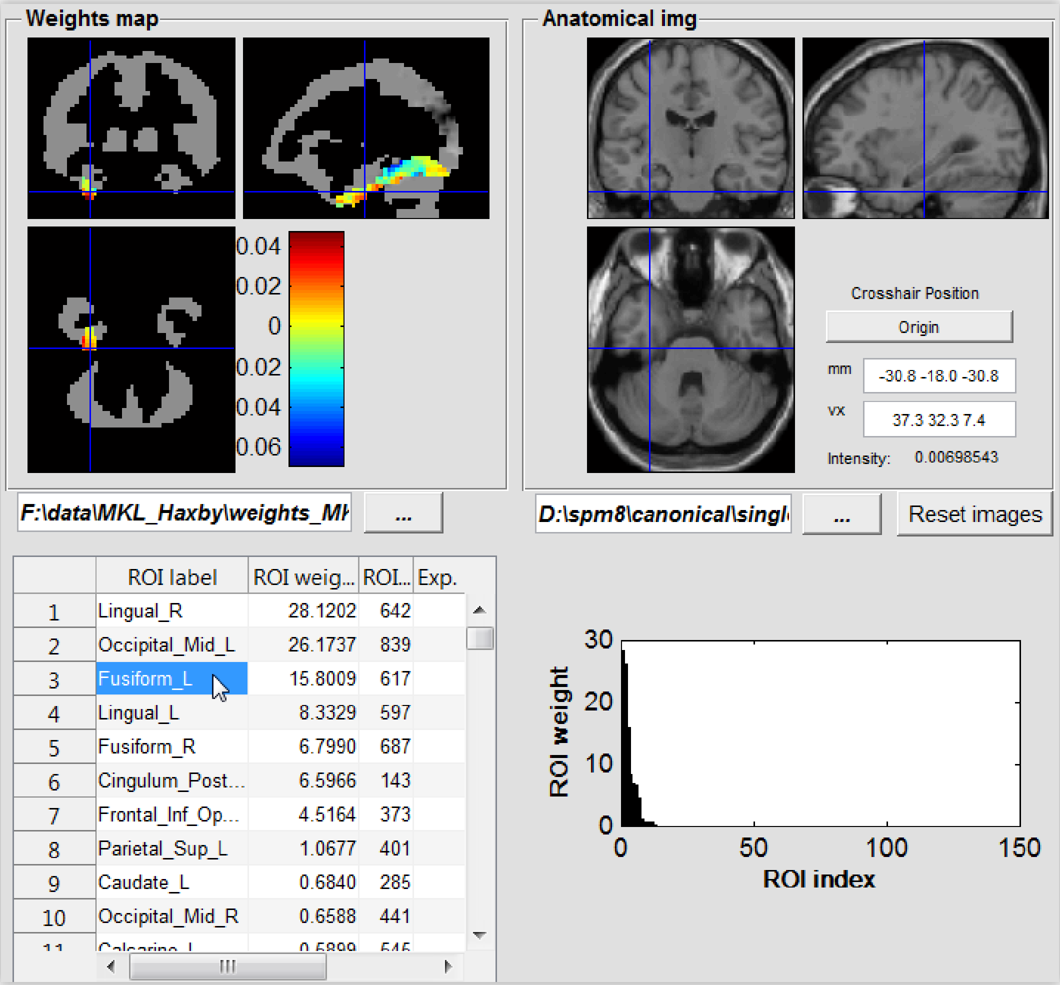
But also $d_m$

→ Can be considered as a hierarchical approach

# 5. Multiple Kernel Learning (MKL)



# 5. Multiple Kernel Learning (MKL)



## 5. Multiple Kernel Learning (MKL)

| + | - |
|---|---|
| Anatomically/ functionally relevant | Stability of the selected regions |
| Used the full pattern to build the model | Depends on atlas |
| Thresholded and sorted list of regions according to $d_m$ | |
| Hierarchical view of the model parameters | |

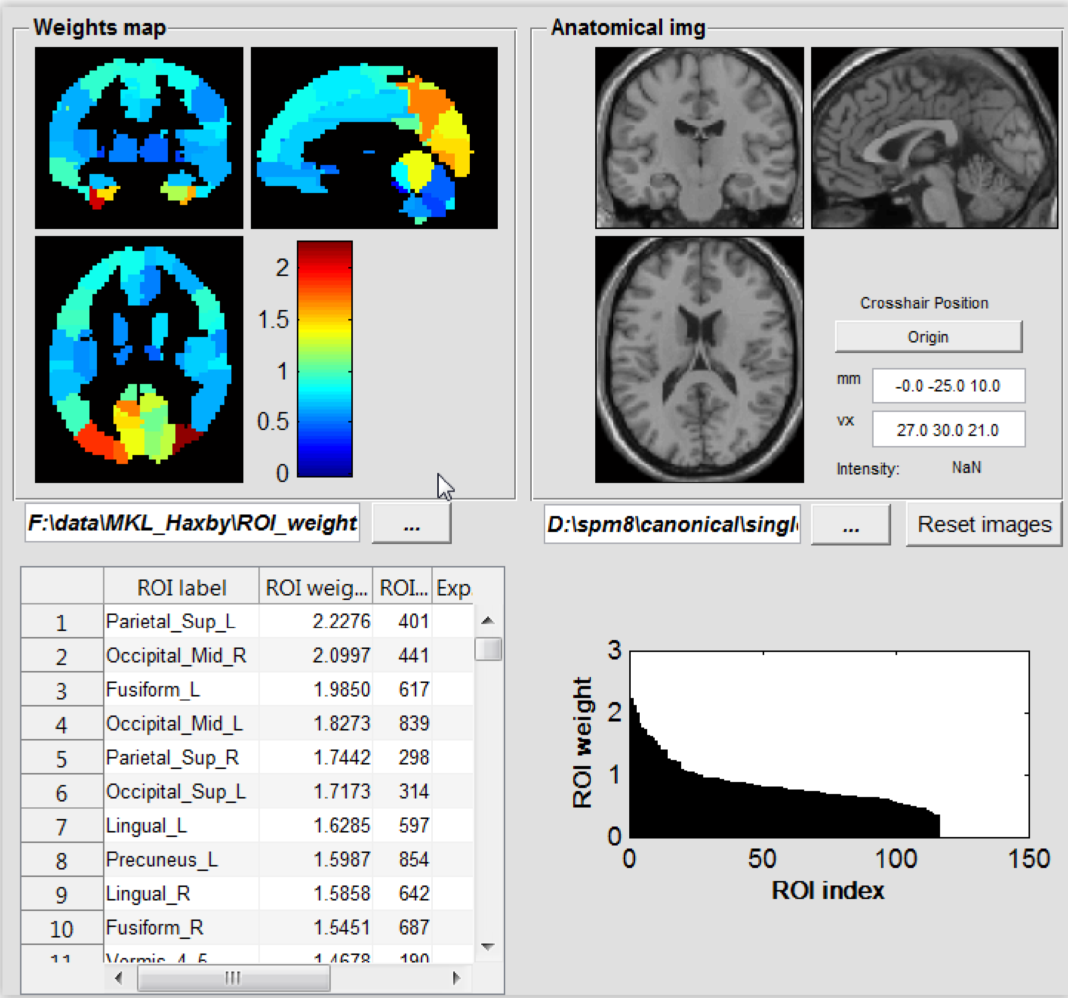
6. A posteriori knowledge: weight summarization (Schrouff et al., 2013)

Build weight maps then summarize weights according to anatomically defined regions (e.g. based on an atlas).

Mathematically:

$$NW_{ROI} = \frac{\sum_{v \in ROI} |W_v|}{\# v \in ROI}$$

# 6. A posteriori knowledge: weight summarization



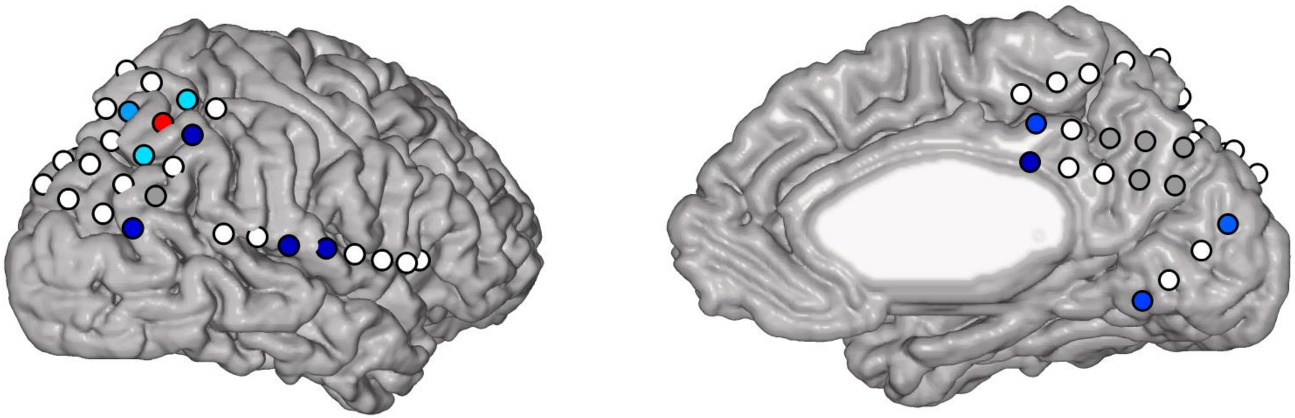

AAL atlas (Tzourio-Mazoyer, 2002).

## 6. A posteriori knowledge: weight summarization

| + | - |
|---|---|
| Anatomically/ functionally relevant | No thresholding of the sorted list of regions |
| Used the full pattern to build the model | Depends on atlas |
| Sorted list of regions according to their proportion of $NW_{ROI}$ | |

How about electrophysiological recordings?

- All strategies explained previously apply
- For MKL, can be applied on each dimension (i.e. channels, time point, frequency band) or combination (e.g. channel + frequency band)



## Conclusions:

- Interpreting machine learning based models can be complex.
- Different strategies exist, but always a trade-off between different parameters:
  - Thresholded list of voxels/regions contributing
  - Anatomical/functional relevance
  - Multivariate power
  - Stability of the pattern
  - Computational time

⟹ Your (informed) choice!

**Acknowledgments:**

- F.R.S. – F.N.R.S. Belgian National Research Funds
- Belgian American Educational Foundation (BAEF)
- Medical Foundation of the Liège Rotary Club
- Stanford University
- University College London
- Wellcome Trust

Figures created using PRoNTo (www.mlnl.cs.ucl.ac.uk/pronto/, development version to be released soon).

# References:


- Baldassarre, L., J. Mourao-Miranda, et M. Pontil. «Structured Sparsity Models for Brain Decoding from fMRI data.» *Proceedings of the 2nd conference on Pattern Recognition in NeuroImaging.* 2012.
- Filippone, M., A. Marquand, C. Blain, C. Williams, J. Mourao-Miranda, et M. Girolami. «Probabilistic prediction of neurological disorders with a statistical assessement of neuroimaging data modalities.» *Annals of Applied Statistics* 6 (2012): 1883-1905.
- Kriegeskorte, Nikolaus, Rainer Goebel, et Peter Bandettini. «Information-based functional brain mapping.» *PNAS* 103 (2006): 3863-3868.
- Rakotomamonjy, Alain, Francis R. Bach, Stéphane Canu, et Yves Grandvalet. «SimpleMKL.» *Journal of Machine Learning* 9 (2008): 2491-2521.
- Schrouff, Jessica, Julien Cremers, Gaëtan Garraux, Luca Baldassarre, Janaina Mourão-Miranda, et Christophe Phillips. «Localizing and comparing weight maps generated from linear kernel machine learning models.» *Proceedings of the 3rd workshop on Pattern Recognition in NeuroImaging.* 2013.
- Tibshirani, R. «Regression shrinkage and selection via the lasso.» *Journal of the Royal Statistical Society.* 58 (1996): 267-288.
- Tzourio-Mazoyer, N., et al. «Automated Anatomical Labeling of activations in SPM using a Macroscopic Anatomical Parcellation of the MNI MRI single-subject brain.» *NeuroImage* 15 (2002): 273-289.
- Zou, H., et T. Hastie. «Regularization and variable selection via the elastic net.» *J. R. Statist. Soc. B* 67 (2005): 301-320.