PRoNTo

Pattern Recognition for Neuroimaging Toolbox

Pattern Recognition Methods: Basics

João M. Monteiro

Based on slides from
Jessica Schrouff and Janaina Mourão-Miranda

PRoNTo course
UCL, London, UK
2017

Outline

- Pattern Recognition
  - Concepts & Framework
- Examples of Linear Models/Machines
  - Classification
  - Regression
- Considerations for Neuroimaging

Pattern Recognition Concepts

Machine learning models:
Enable predictions from brain imaging

Healthy
vs.
Disease

Cognitive state #1
vs.
Cognitive state #2

Clinical Score

Pattern Recognition Concepts

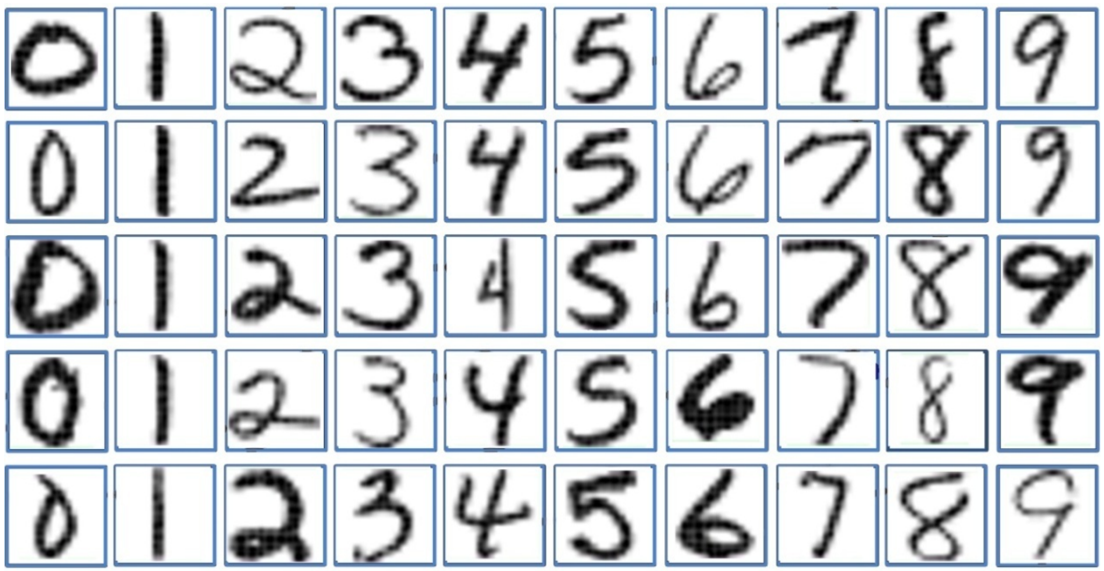
- Pattern recognition aims to assign a label to a given example (test example) based on statistical information extracted from the previous seen examples (training examples).

- The examples to be classified are usually groups of measurements or observations (e.g. brain image), defining points in an appropriate multidimensional vector space.

Currently implemented in PRoNTo

- Types of learning procedures:
  - $\boxed{\text{Supervised learning}}$
  - Unsupervised learning
  - Semi-supervised learning, reinforcement learning.

Pattern Recognition Concepts



0 1 2 3 4 5 6 7 8 9

Samples $\mathbf{X}$

Labels $\mathbf{y}$

$$f : \mathbf{X} \rightarrow \mathbf{y}$$

$$f : \mathbf{x}_* \rightarrow y_*$$

PRoNTo

Pattern Recognition Framework

Input (brain scans)

$\mathbf{x}_1$

$\mathbf{x}_2$

$\mathbf{x}_3$

No mathematical model available

Output (control/patient)

y_1

y_2

y_3

Machine Learning Methodology

Computer-based procedures that learn a function from a series of examples

Training Examples:

$\{\mathbf{x}_1, y_1\}, \cdots, \{\mathbf{x}_n, y_n\}$

Learning/Training Phase

Estimate a predictive function f such that

$$f : \mathbf{x}_i \rightarrow y_i$$

Predictive function

f

Test Example

$\mathbf{x}_*$

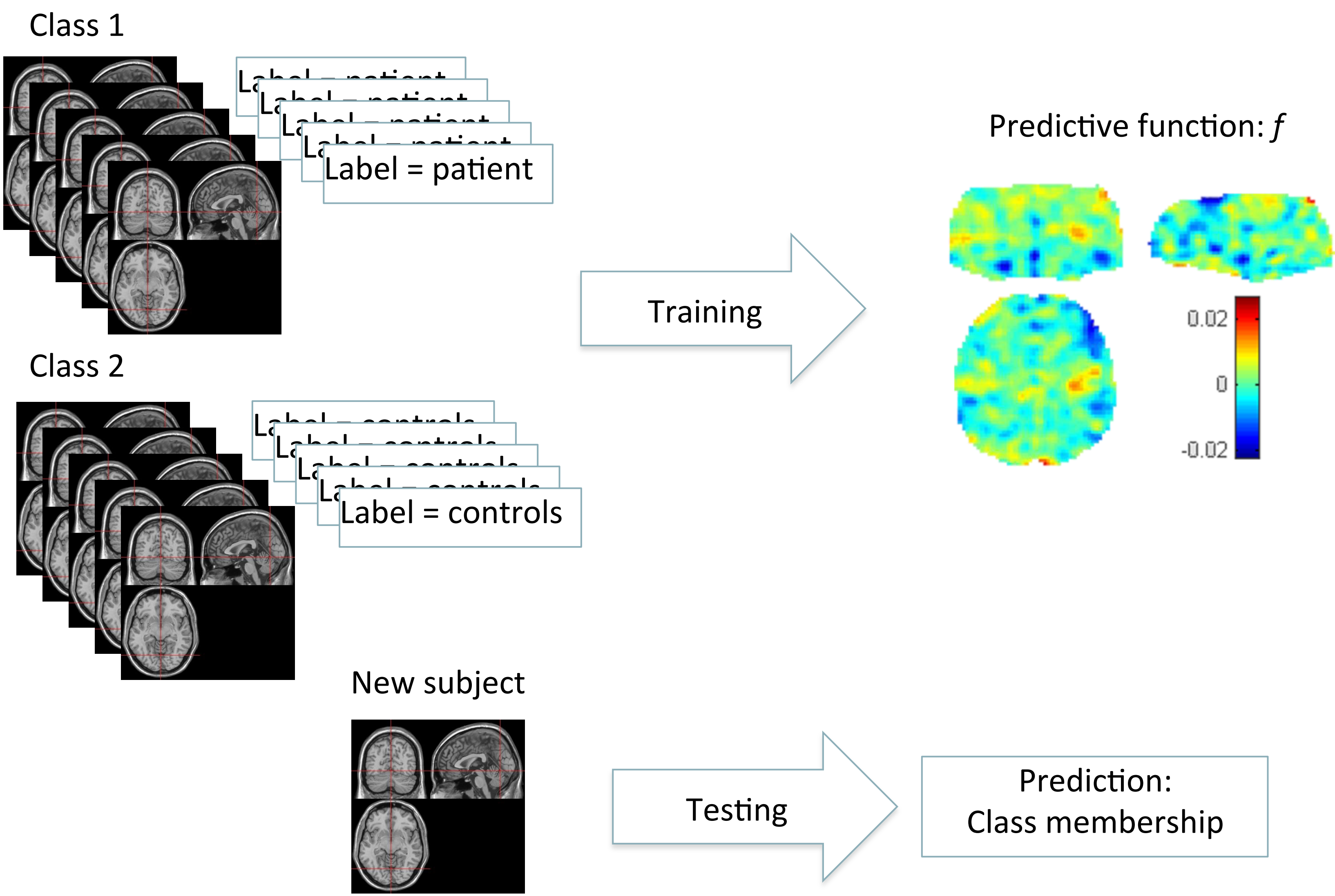
Testing Phase

Prediction

$$f(\mathbf{x}_*) = y_*$$

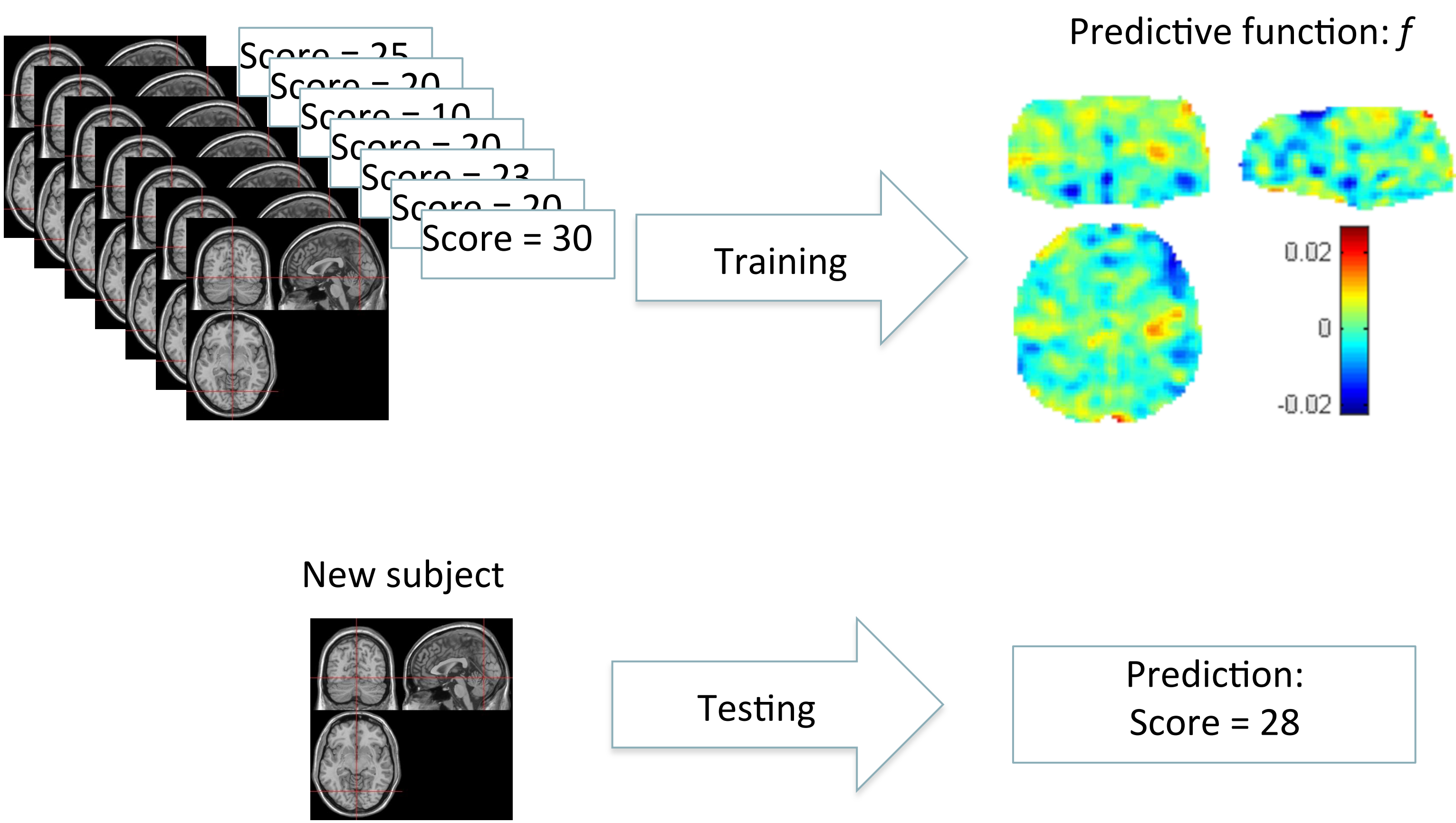
PRoNTo

Pattern recognition: classification model



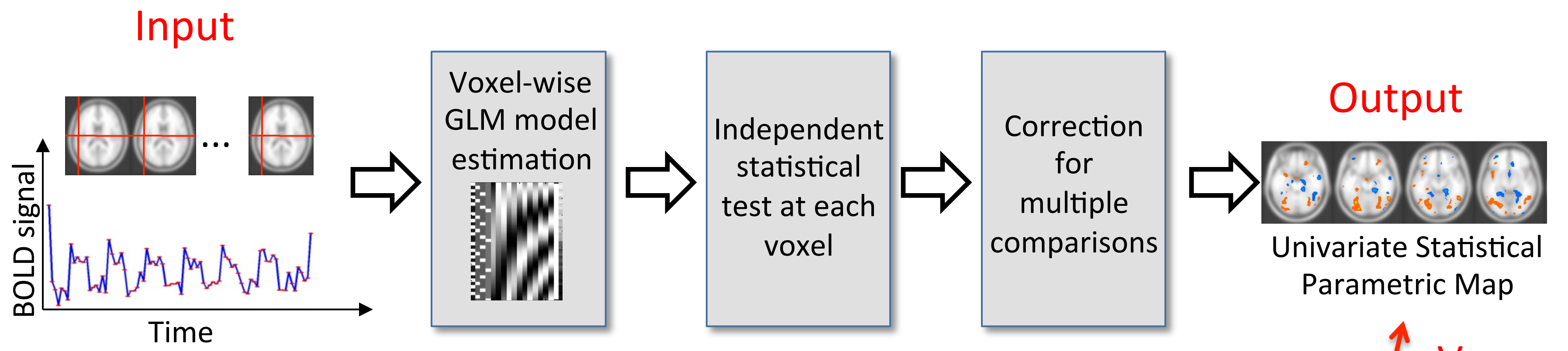
PRoNTo

Pattern recognition: regression model

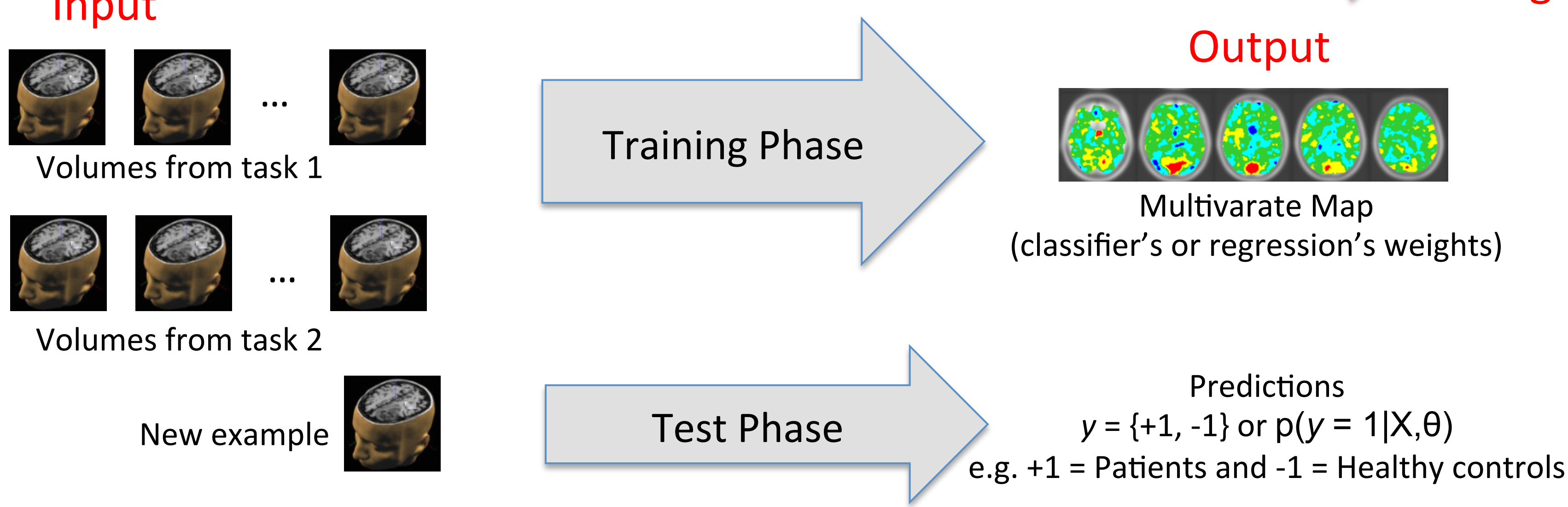


PRoNTo

Standard Statistical Analysis (mass-univatiate)



Pattern Recognition Analysis (multivariate)



Advantages of Pattern Recognition Analysis

Explore the multivariate nature of neuroimaging data

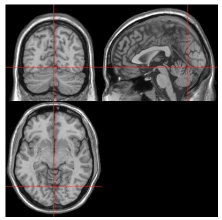
- MRI/fMRI data are multivariate since most of the brain functions are distributed processes involving a network of brain regions.

- Pattern recognition analysis can yield greater sensitivity than conventional analysis due to its multivariate properties.

Can be used to make predictions for new examples

- Enable clinical applications: previously acquired data can be used to make diagnostic or prognostic for new subjects.

How to extract features from MRI?

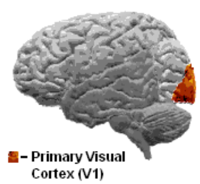


Whole brain volume

Feature Vector

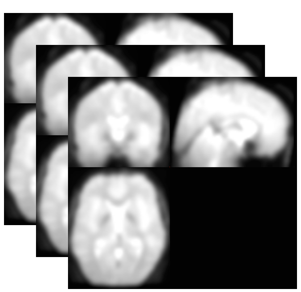
Dimensionality =
number of voxels

Region of interest (ROI)



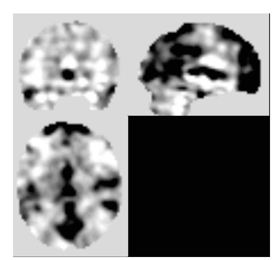
Feature Vector

Dimensionality =
number of voxels within the ROI



fMRI

General Linear Model

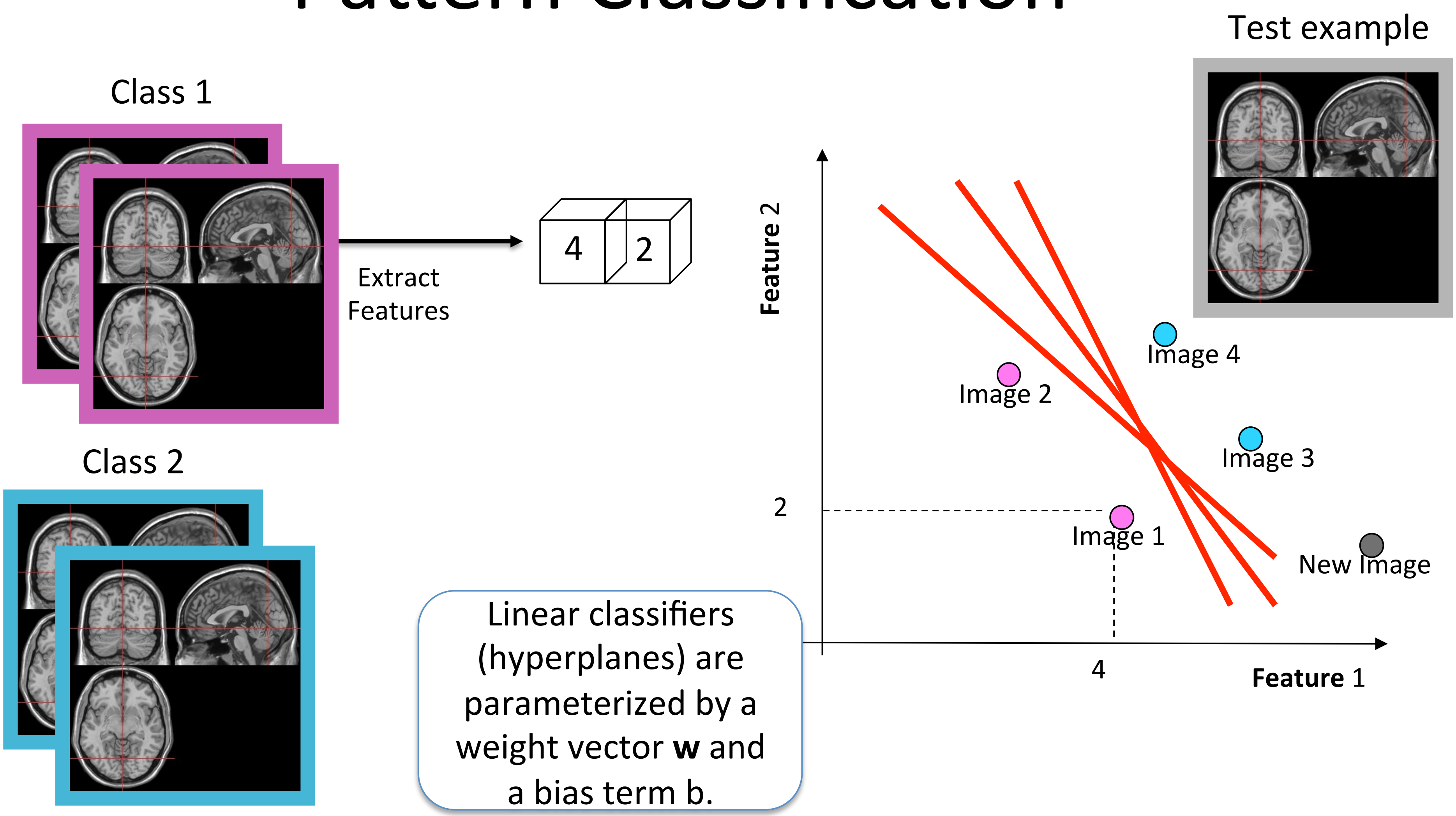


Beta/Contrast images

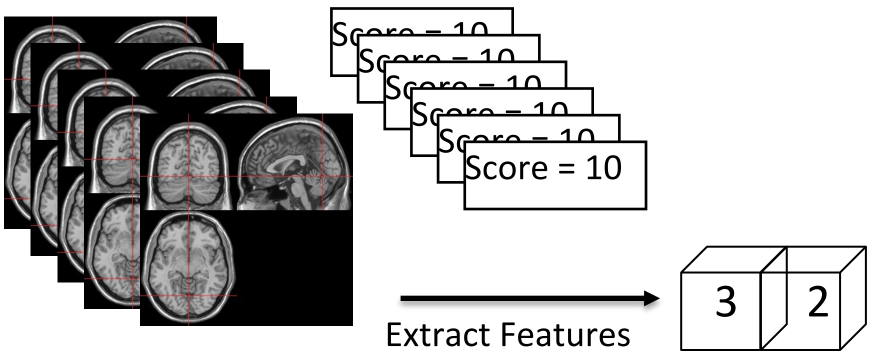
Feature Vector

Dimensionality =
number of voxels

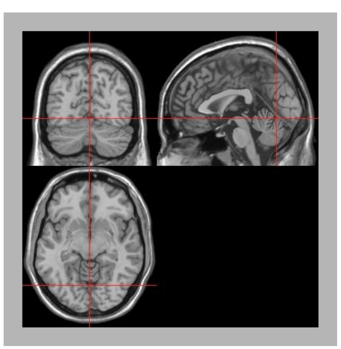
Pattern Classification



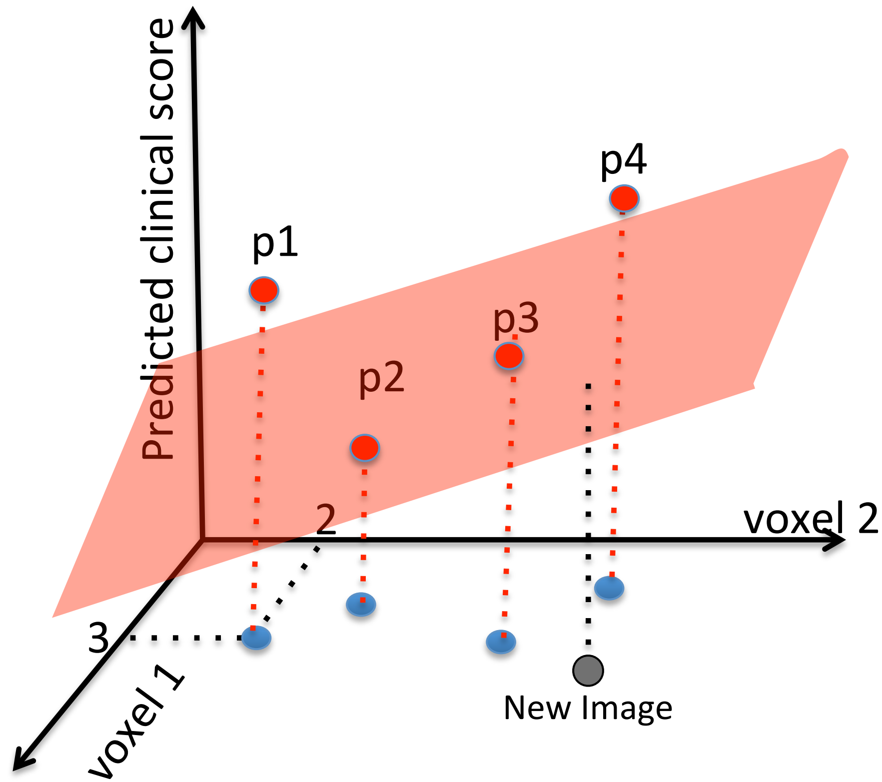
Pattern Regression



Test example



Linear regression models are also parameterized by a weight vector **w** and a bias term b.



Pattern Recognition Models

- Linear predictive models (classifier or regression) are parameterized by a weight vector **w** and a bias term *b*.

- The weight vector **w** can be expressed as a linear combination of training examples $\mathbf{x}_i$ (N = number of training examples).

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$$

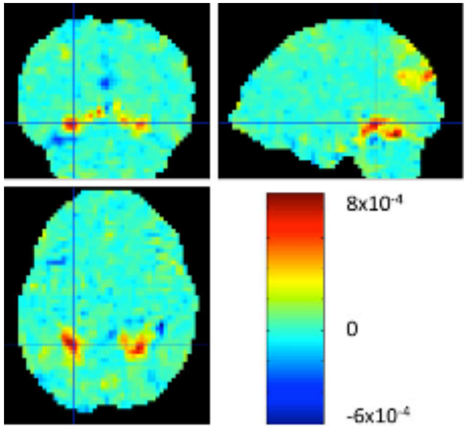
- The general equation for making predictions for a test example x_* is:

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

PRoNTo

Weight map or predictive pattern

- The weight vector **w** has the same dimensionality of the input data/image and can be plotted as an image.



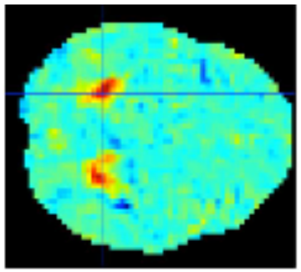
- The weight vector might provide potential insights into brain function or structure that drives the prediction, but the interpretation should be done with care!

Using the weights for prediction

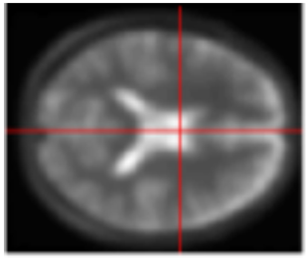
Predictive function

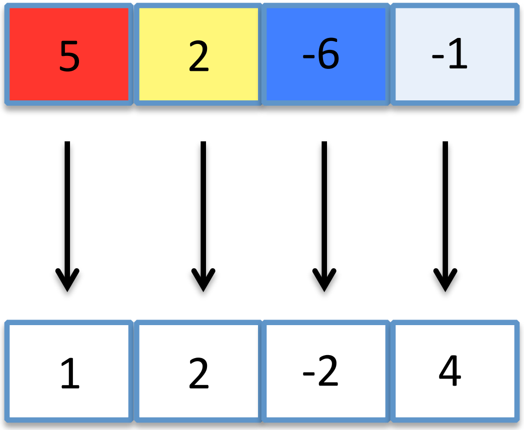
$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b$$

Weight map (**w**)



New example (**x***)



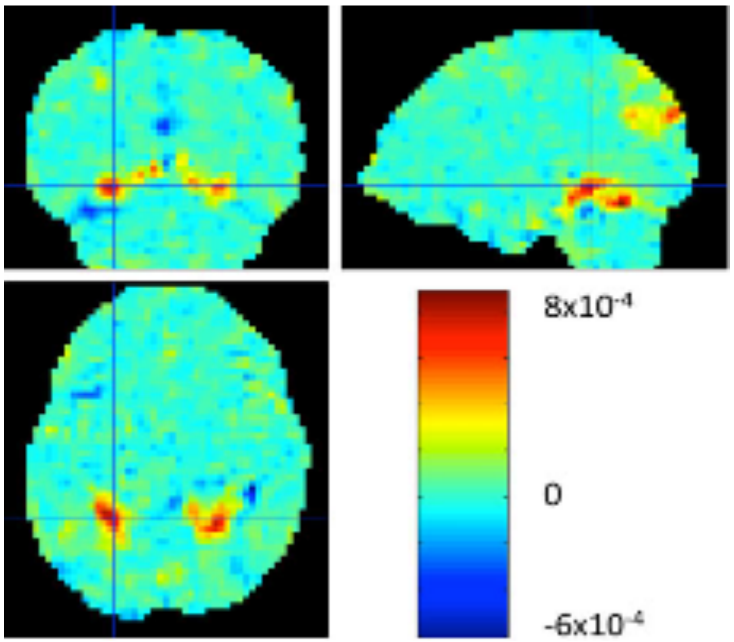


$$f(\mathbf{x}_*) = (5\times 1) + (2\times 2) + (-6\times -2) + (-1\times 4) + 0$$

$$f(\mathbf{x}_*) = 5 + 4 + 12 - 4 = 17$$

$f(\mathbf{x}_*)$ is the predicted score for regression or the distance to the decision boundary for classification models.

How to interpret the weight vector w?



- ✓ It is a spatial representation of the predictive function.
- ✓ Shows the contribution of each feature/ voxel for the prediction.
- ✓ Multivariate pattern -> All voxels with weights different from zero contribute to the final prediction (no arbitrary threshold should be applied).
- ✓ **No local inferences can be made!**

Challenges in Neuroimaging

- In neuroimaging applications often the dimensionality of the data is greater than the number of examples (ill-conditioned problems).

- Possible solutions:
  - Region of interest (ROI)
  - Feature selection strategies
  - Searchlight
  - **Regularization + Kernel Methods**

PRoNTo

Regularization

- **Regularization** is a technique used in an attempt to **solve ill-posed problems** and to **prevent overfitting** in statistical/machine learning models.

- Regularized methods find **w** minimizing an objective function consisting of a data fit term ***E*** and a penalty/regularization term ***J***

Regularization hyper-parameter

$$\min_{\mathbf{w}\in R^p} \{E(\mathbf{w}) + \lambda J(\mathbf{w})\}$$

The data fit term is a **error function *E***

The **regularisation term *J***

- Many machine learning algorithms are particular choices of ***E*** and ***J*** (e.g. Kernel Ridge Regression (KRR), Support Vector Machine (SVM)) .

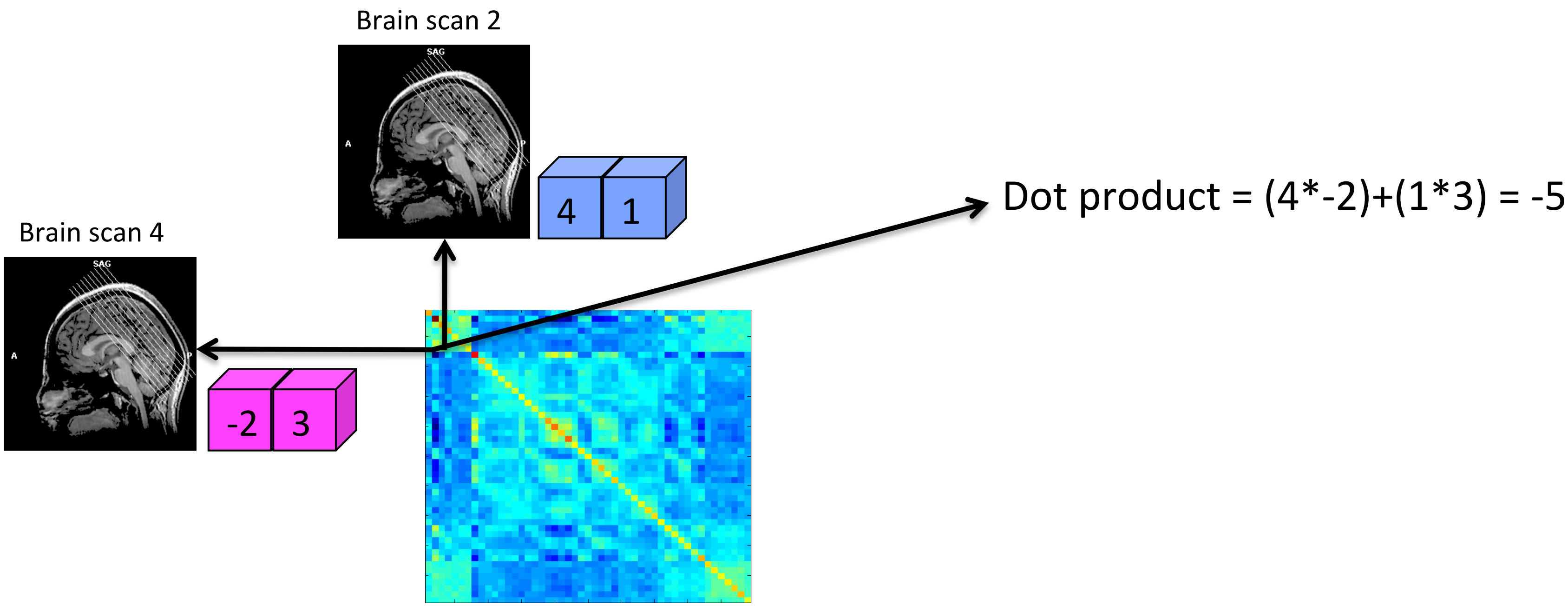
Kernel Methods

- **Kernel methods** provide a powerful and unified framework for investigating general types of relationships in the data (e.g. classification and regression)

- **Consists of two parts:**
  - Computation of the kernel matrix (mapping into the feature space)
  - A learning algorithm based on the kernel matrix

- **Main advantage:**
  - Computational efficiency

Kernel Function ("similarity" measure)



- Kernel is a function that, given $\mathbf{x}$ and $\mathbf{x}_*$, returns a real number characterizing their similarity;

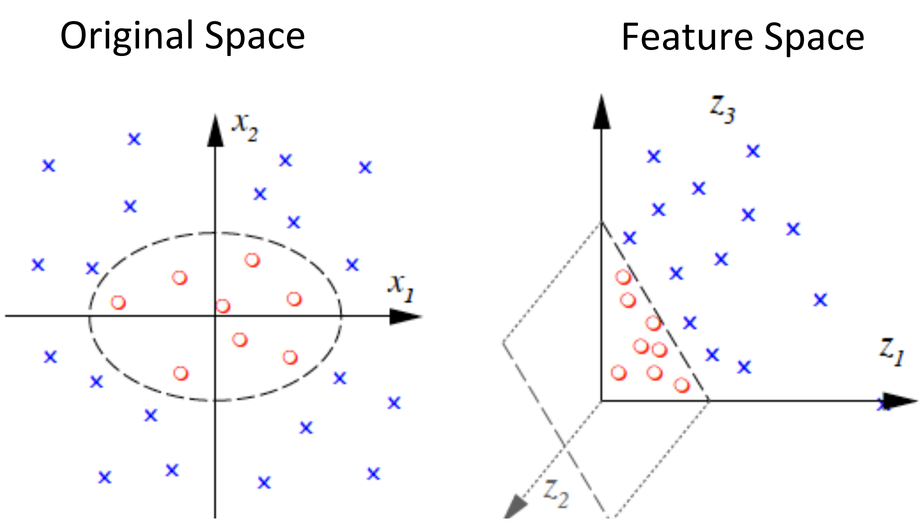
- A simple type of similarity measure between two vectors is a dot product (linear kernel).

PRoNTo

Nonlinear Kernels

- There are more general "similarity measures", i.e. nonlinear kernels: Gaussian kernel, Polynomial kernel, etc.

- Nonlinear kernels are used to map the data to a higher dimensional space as an attempt to make it linearly separable.



$$\Phi : R^2 \to R^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$

Advantage of linear models

- Neuroimaging data are extremely high-dimensional and the sample sizes are very small, therefore non-linear kernels often don't bring any benefit.

- Linear models reduce the risk of overfitting the data and allow direct extraction of the weight vector as an image (i.e. predictive map).

Learning with kernels

- Making predictions with kernel methods

$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b \longrightarrow$$

Primal representation

$$f(\mathbf{x}_*) = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \cdot \mathbf{x}_* + b$$

$$f(\mathbf{x}_*) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}_*) + b \longrightarrow$$

Dual representation

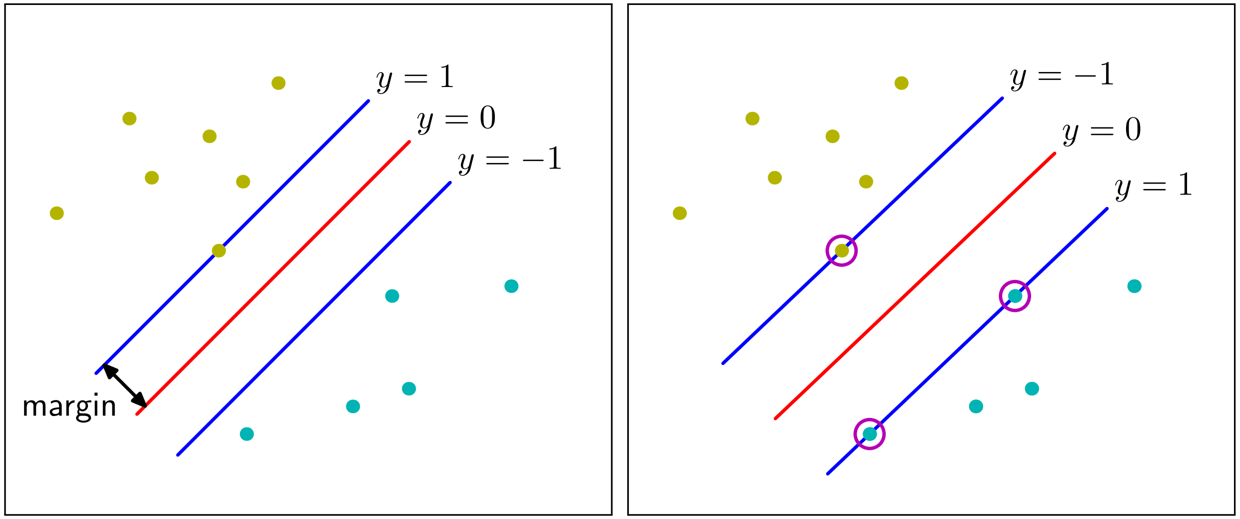
Algorithms available in PRoNTo

Kernel methods

- **Classification:**
  - ✓ Support Vector Machine (SVM)
  - ✓ Multiple Kernel Learning (MKL) Classifier
  - ✓ Binary Gaussian Process Classifier (GPC) -> probabilistic
  - ✓ Multiclass Gaussian Process Classifier (GPC) -> probabilistic

- **Regression:**
  - ✓ Kernel Ridge Regression (KRR)
  - ✓ Multiple Kernel Learning (MKL) Regression
  - ✓ Relevance Vector Regression (RVR) -> probabilistic
  - ✓ Gaussian Process Regression (GPR) -> probabilistic

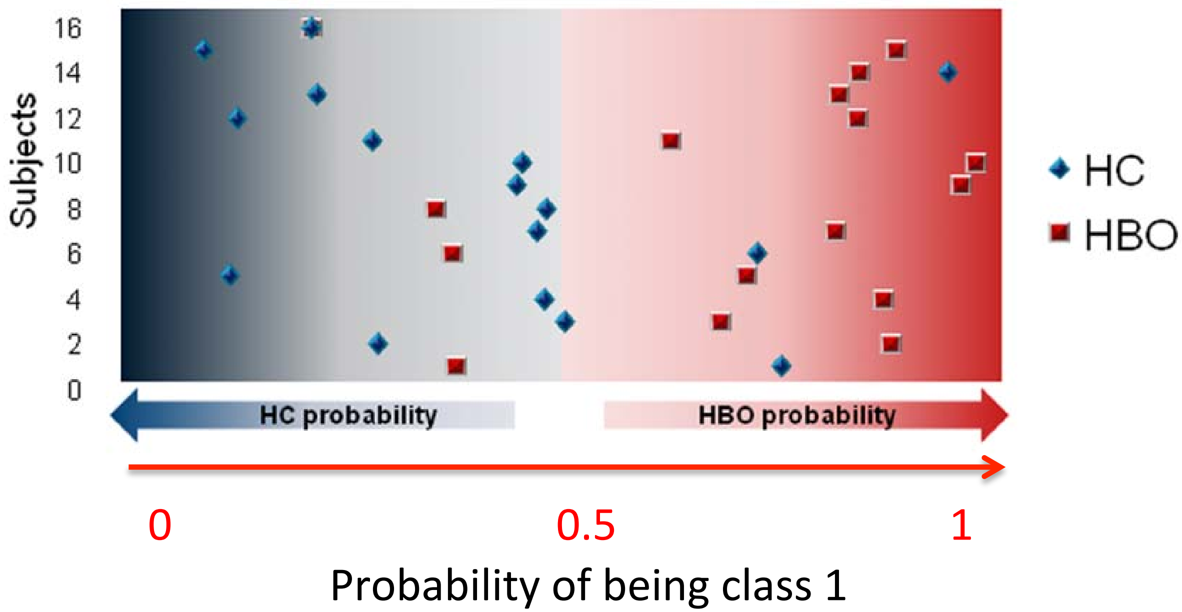
Support Vector Machine (SVM)



- Sparse solution in terms of examples (support vectors)
- Computational efficient
- Gives good results for most problems

PRoNTo

Gaussian Process Classifier – Binary/Multiclass



- Explicit probabilistic framework
- Natural extension to direct multi-class classification
- Provide mechanisms for automatic parameter optimization

PRoNTo

Kernel Ridge Regression

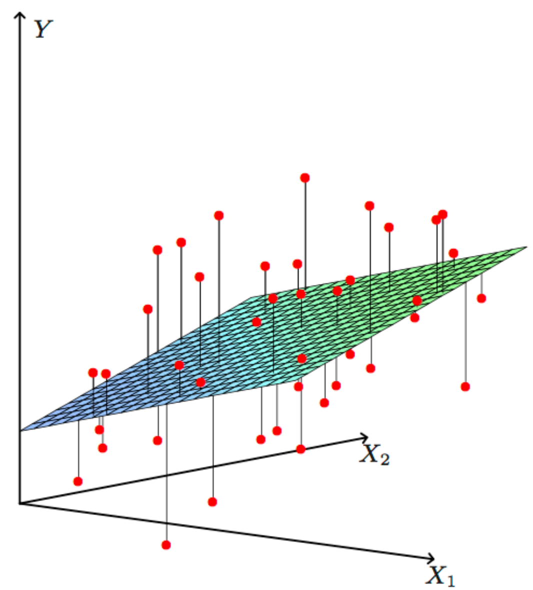


Illustration of a linear least squares fitting with $X \in \mathrm{IR}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.

Hastie, Tibshirani & Friedman, 2009

- Dual representation of ridge regression, also known as the linear least square regression with Tikhonov regularization (Chu et al. 2011).

PRoNTo

Relevance vector machine

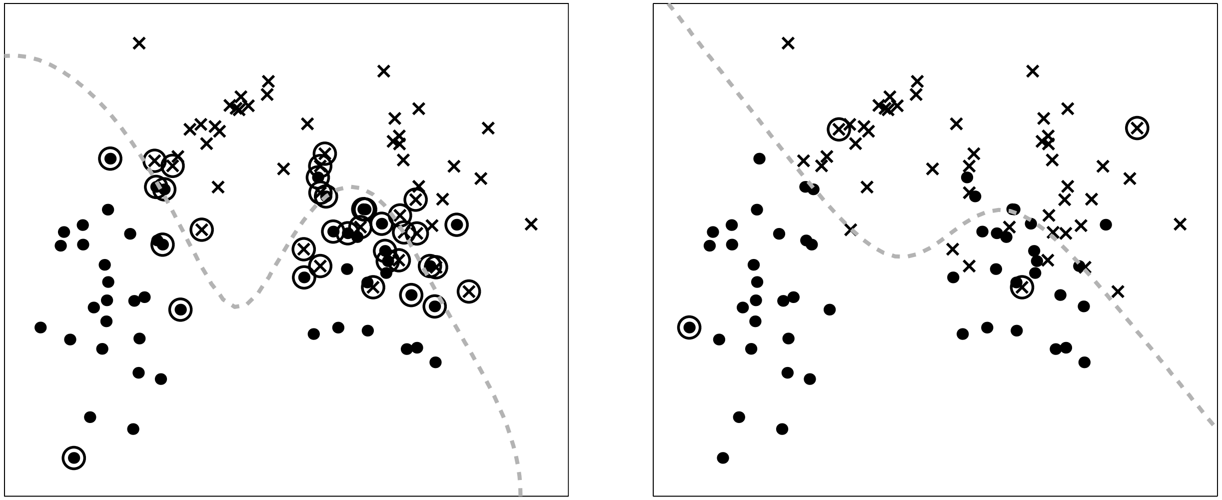
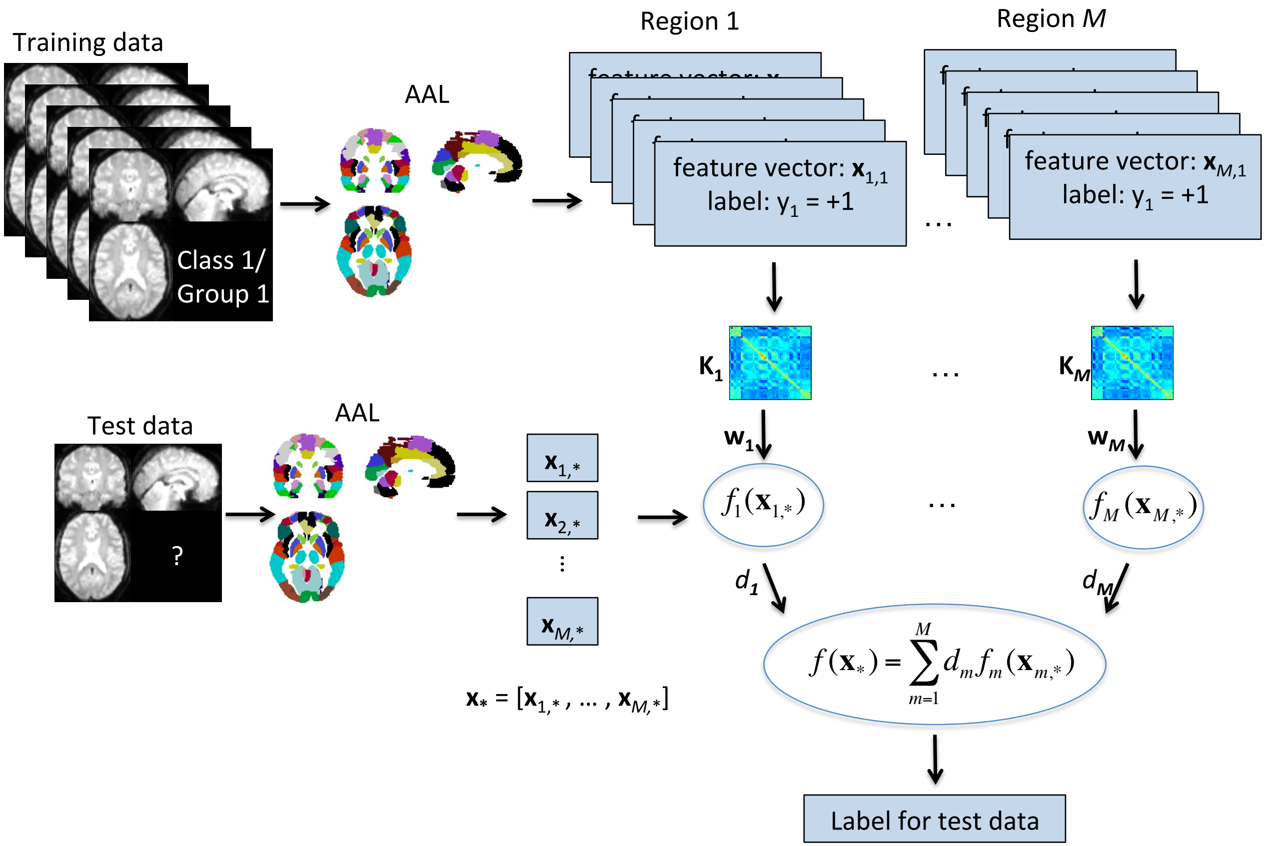


Figure 3: SVM (left) and RVM (right) classifiers on 100 examples from Ripley's Gaussian-mixture data set. The decision boundary is shown dashed, and relevance/support vectors are shown circled to emphasise the dramatic reduction in complexity of the RVM model.

- Probabilistic: apply a Bayesian treatment to SVM
- Similarly to SVM finds a sparse solution (relevance vectors)
- Risk of local minima during optimization

PRoNTo

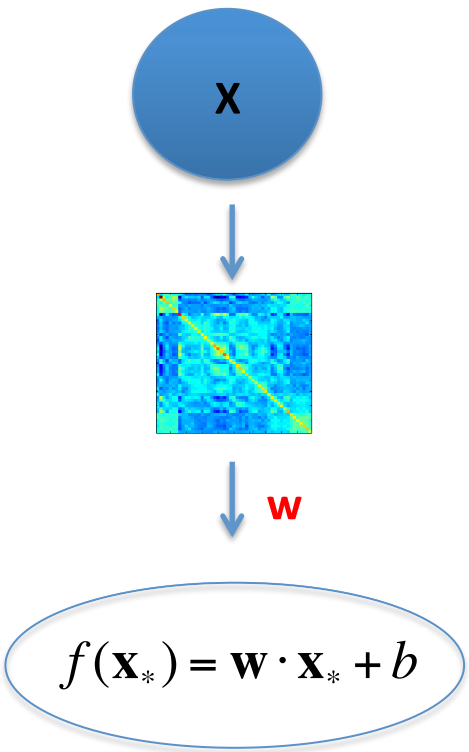
Multiple Kernel Learning (MKL)



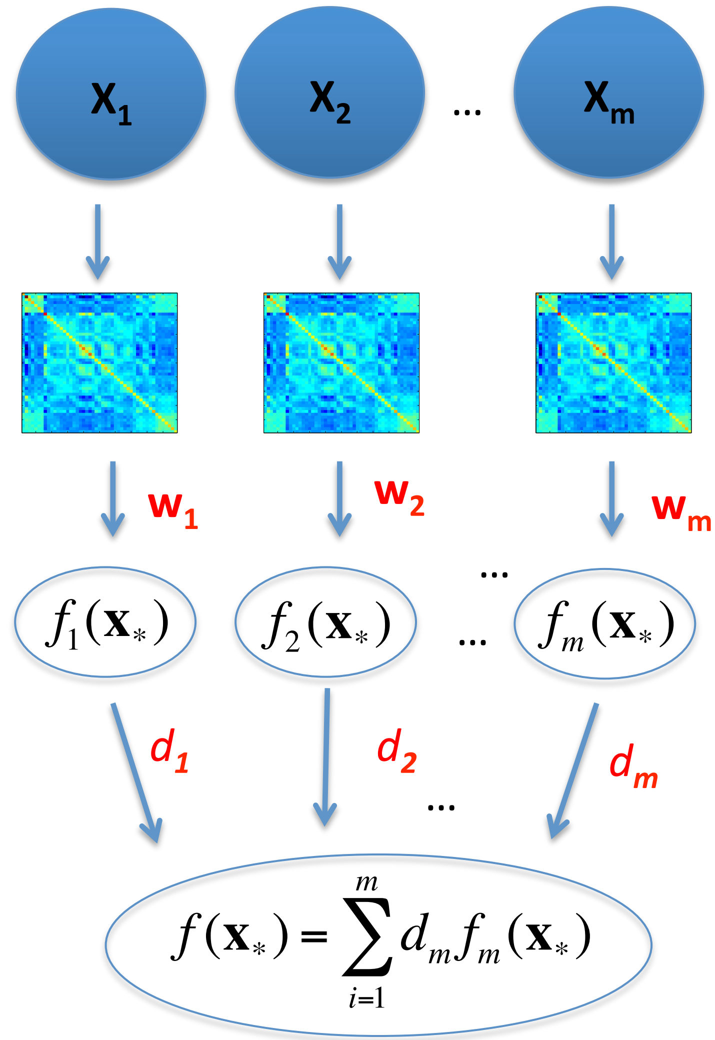
- MKL has been proposed as an approach to learn a decision function based on a predefined set of kernels.

PRoNTo

Single kernel SVM



Multiple kernel SVM

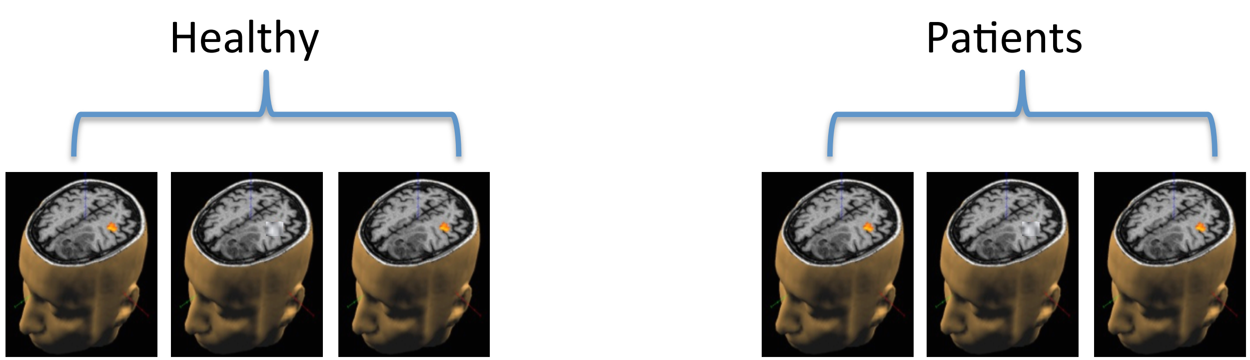


PRoNTo

Considerations for Neuroimaging

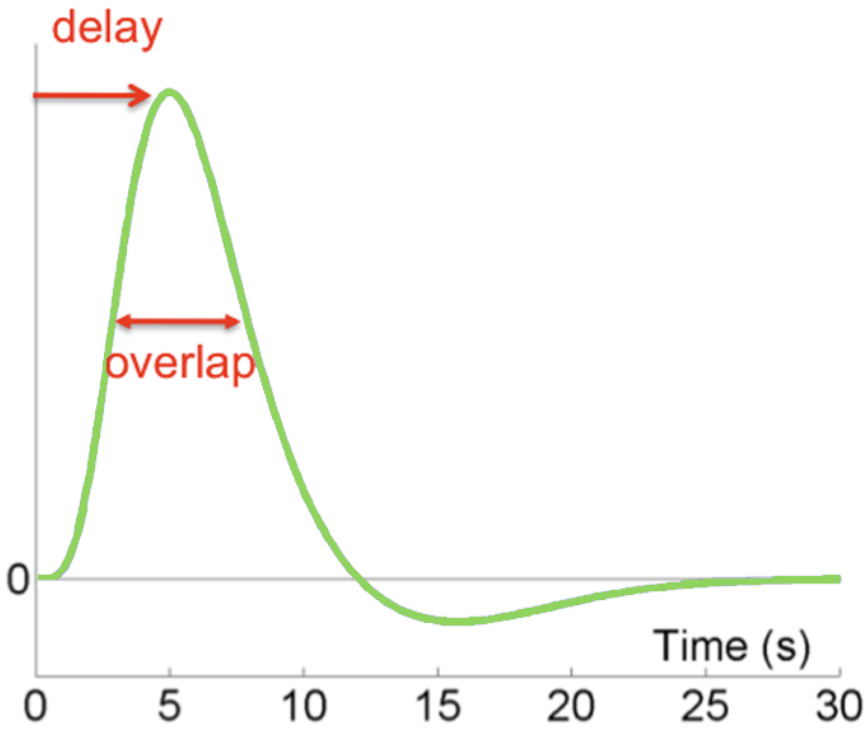
Define your question:

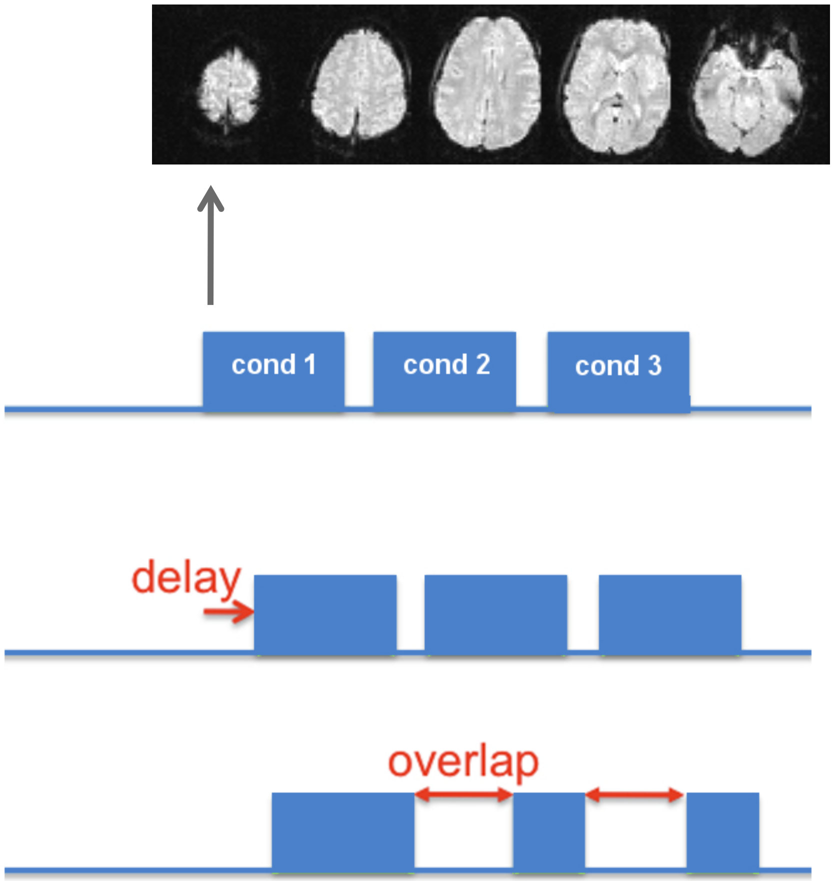
- Classification or regression?
- Specify the subjects and/or conditions



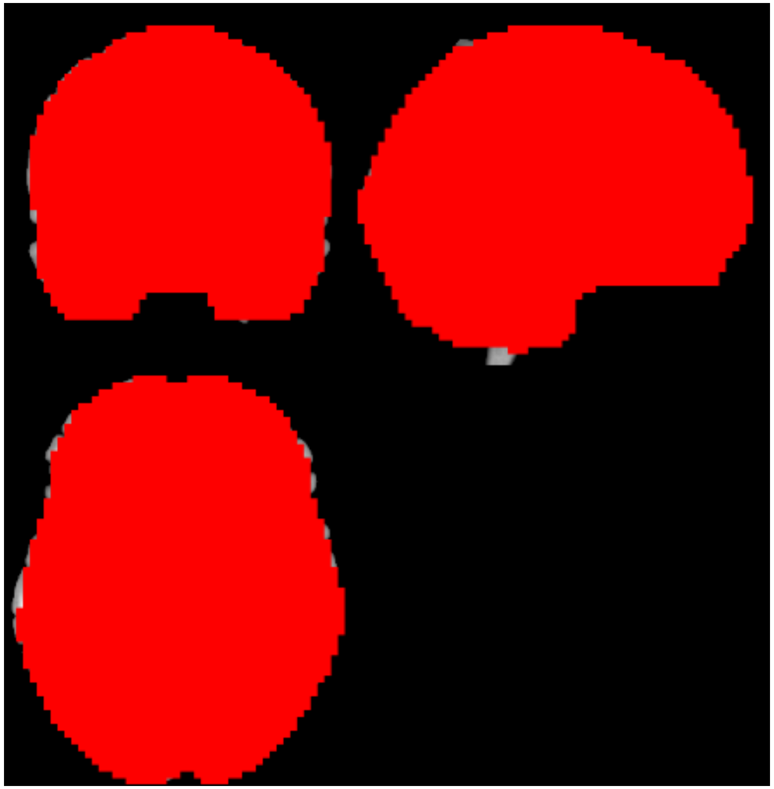
PRoNTo

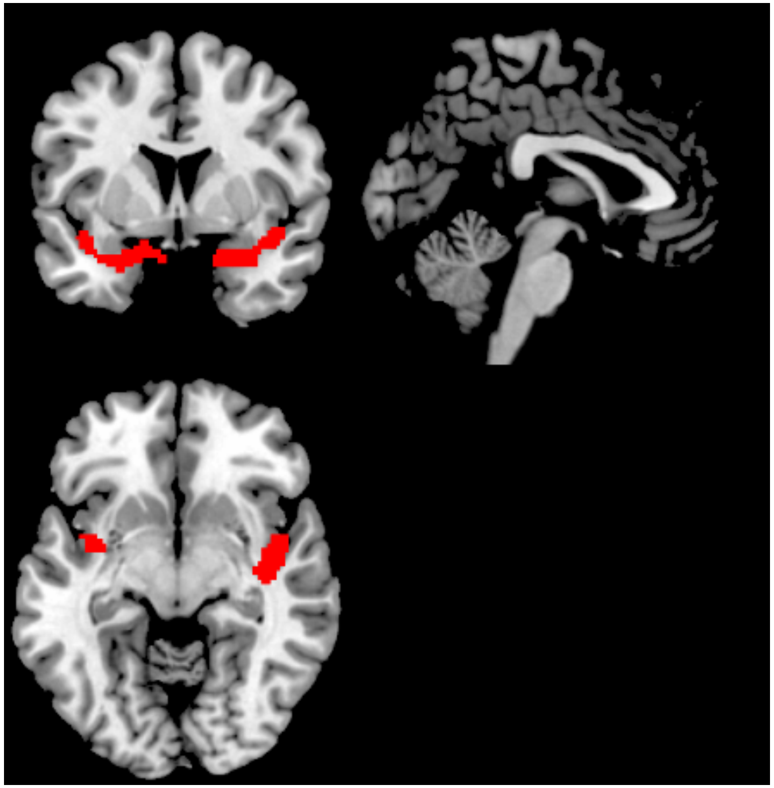
Considerations for fMRI (BOLD)





How to extract features from MRI?





References

PRoNTo paper:

- Schrouff J*, Rosa MJ*, Rondina J, Marquand A, Chu C, Ashburner J, Phillips C, Richiardi J, Mourao-Miranda J. PRoNTo: Pattern Recognition for Neuroimaging Toolbox, Neuroinformatics, February 2013. *co-first authors.

Reviews:

- Haynes and Rees (2006) Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.,* 7, 523-534
- Pereira, Mitchell, Botnivik (2009) Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage,*45, S199-S209

Books:

- Hastie , Tibishirani, Friedman (2003) Elements of Statistical Learning. *Springer*
- Bishop, Jordan, Kleinberg, Schölkopf (2006) Pattern Recognition and Machine learning. *Springer*

Machines:

- Burges (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.
- Rasmussen, Williams (2006) Gaussian Processes for Machine Learning. *The MIT Press*.
- Tipping (2001) Sparse Bayesian Learning and the Relevance Vector Machine *Journal of Machine Learning Research,* 1, 211-244
- Breiman (1996) Bagging Predictors Machine Learning, 24, 123-140
- Dietterich, Bakiri (1995) Solving multiclass learning problem via error-correcting output codes. Journal of Artificial Intelligence Research, 2: 263-286.
- Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. Journal of Machine Learning Research, 9, 2491-2521.