KING'S College LONDON

UCL

Towards Machine Learning tools for diagnosis

Janaina Mourao-Miranda

Computer Science Department, UCL

Centre for Neuroimaging Sciences, Institute of Psychiatry, KCL

PASCAL2 Pattern Analysis, Statistical Modelling and Computational Learning

wellcome trust

Outline

- Introduction
- Relevant Clinical Questions
- General Analysis Framework
- Example of Clinical Applications

Introduction

- It's known that psychiatric/neurological disorders affect brain function and structure. However, to date the translation of neuroimaging research findings into diagnostic tools has been very limited due to lack of adequate analysis tools.

- In the last years there has been a substantial increase in the use of machine learning/pattern recognition approaches to analyze neuroimaging data.

- Some of the advantages of pattern recognition approaches:
  - Accounts for the spatial correlation of the data (multivariate);
  - Permit classification/prediction ('mind-reading', **clinical application**);

Relevant Clinical Questions

1. Diagnosis
   - Can we classify groups of subjects (e.g. patients vs. controls) using sMRI/fMRI scans?

2. Prognosis
   - Can we predict who will develop a disease based on a baseline scan (e.g. fMRI, sMRI)?

3. Treatment response
   - Can we predict treatment response based on brain scans?

Pattern Recognition in Functional Neuroimaging

Recently, pattern recognition approaches have become increasingly popular tools for the analysis of neuroimaging data.

Input

Training examples
$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$

$f(\mathbf{x}) \longrightarrow y$

Output

Target variables
$\{y_1, y_2, \ldots y_N\}$

$y_i \in R$: regression
$y_i \in \{0,1\}$: classification

Pattern Recognition approaches confer two advantages over conventional neuroimaging analysis techniques:

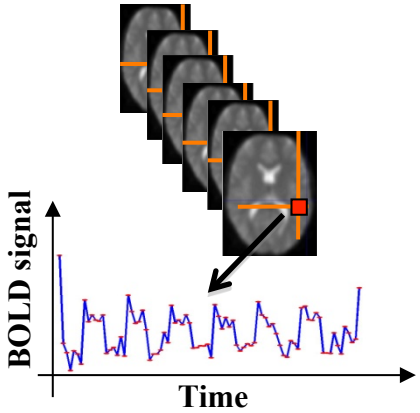
1. They have predictive capacity

2. They are multivariate and utilise spatial correlation in the data

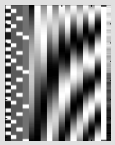
In addition to predictions, it is also desirable to understand which data features (brain regions) carry discriminating information

General Analysis Framework

Standard Statistical Analysis:



Voxel-wise GLM model estimation



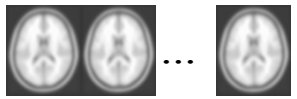
Independent statistical test at each voxel

Correction for multiple comparisons

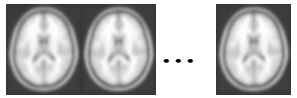
Univariate statistical Parametric map

Pattern Recognition Analysis:

Training Phase:



Class 1 (e.g. patient)



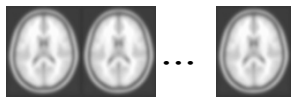
Class 2 (e.g. control)

Feature Selection (optional)

Pattern Recognition

Output 1:
Multivariate representation of the decision function (e.g. weight vector)

Testing Phase:



New Sample (unknown label)

Feature Selection (optional)

Pattern Recognition

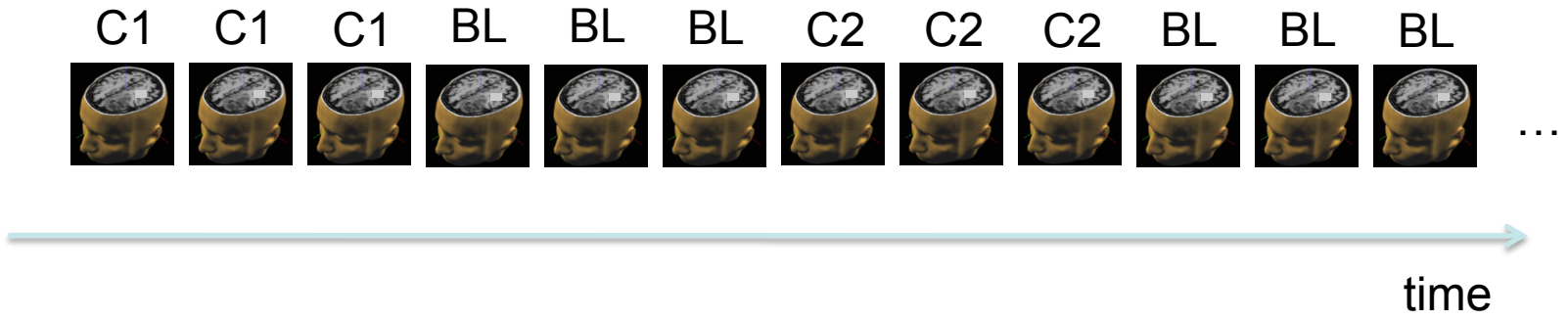
Output 2:

$y = \{+1, -1\}$
$p(y = 1|X,\theta)$

Prediction, e.g.:
-1 : Healthy
+1 : Patient

How to extract features from the fMRI?

- There are many possible ways to define fMRI based patterns as input to the pattern recognition approaches:
  - single fMRI scans.
  - mean of fMRI scans (e.g. mean of images within a block).
  - images corresponding to the GLM coefficients for each experimental condition (in general gives best results for event-related designs).



Curse of dimensionality

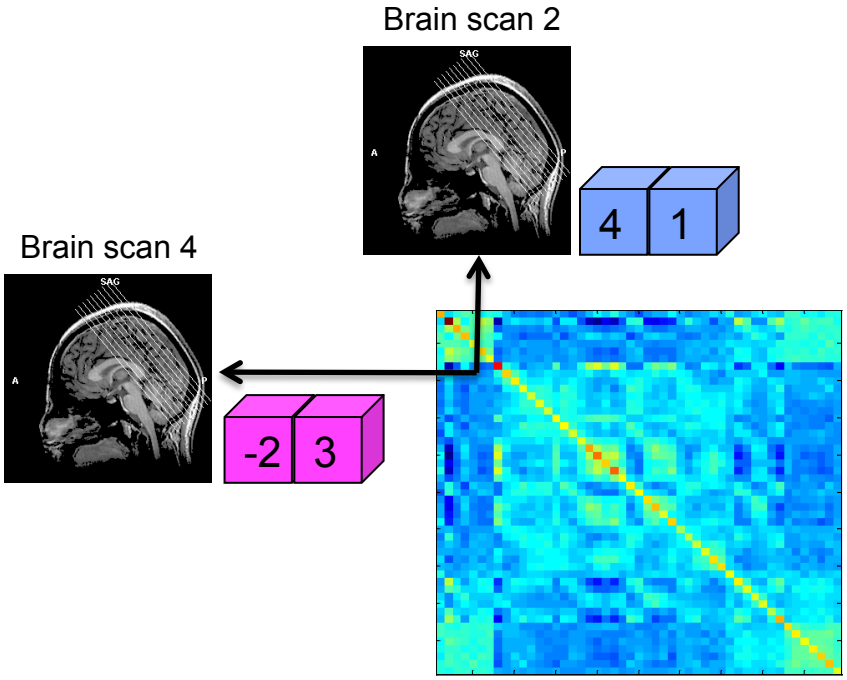
- In neuroimaging applications often the dimensionality of the data is greater than the number of examples (ill-conditioned problems).

- Possible solutions: feature selection strategies, searchlight, **Kernel Methods** (Support Vector Machine, Gaussian Processes, Kernel Ridge Regression, Kernel Fisher Discriminant).

- Kernel methods consist of two parts:
  - A mapping into the embedding or feature space (through the kernel function).
  - A learning algorithm designed to discover linear patterns in that space (e.g. non-linear relationships in the input space).

- Advantages:
  - Represent a computational shortcut which makes possible to represent linear patterns efficiently in high dimensional space.
  - Using the dual representation with proper regularization enables efficient solution of ill-conditioned problems.

Kernel Function



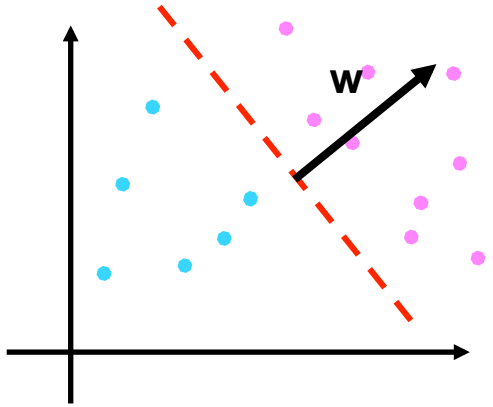
• Kernel is a function that, for given two pattern x and x*, returns a real number characterizing their similarity.

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

•Nonlinear kernels are used to map the data to a higher dimensional space as an attempt to make it linearly separable (the kernel trick enable the computation of similarities in the feature space without the computing the mapping explicitly).

Hyperplane Classifiers: binary classification can be viewed as a task of finding a hyperplane

- Given a dataset: $<\mathbf{x}_i, y_i>$, i=1,..,N
  observations: $\mathbf{x}_i \in R^2$
  labels: $y_i \in \{-1,+1\}$



- Linear classifiers (hyperplanes) are parameterized by a weight vector **w** and a bias term *b*.

- The weight vector can be expressed as a linear combination of training examples $\mathbf{x}_i$ (where $i = 1,\ldots,N$ and N is the number of training examples).

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

How to make predictions?

- The general equation for making predictions for a test example $\mathbf{x}_*$ with kernel methods is

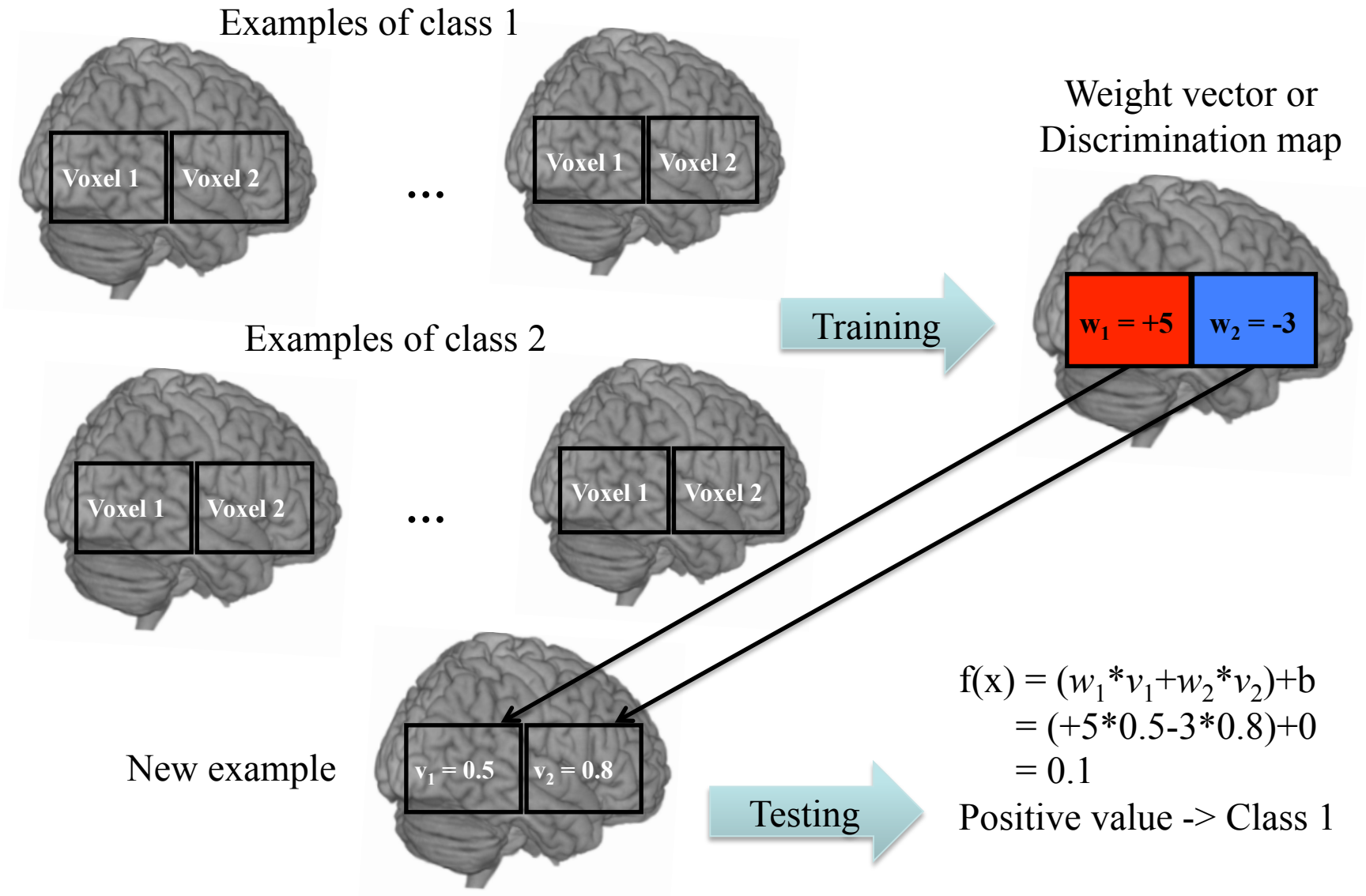
$$f(\mathbf{x}_*) = \mathbf{w} \cdot \mathbf{x}_* + b \longrightarrow \text{Primal representation}$$

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

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

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

How to interpret the weight vector?



Spatial representation of the decision boundary/function, the discrimination is based on the whole pattern.

Other Issues (Clinical Applications):

- Confounds (e.g. medication load)
- Different class distributions (e.g. patient group are more heterogeneous)
- Comorbidity

Examples of Clinical Applications

Can we classify groups (e.g. patients vs. controls) using the whole brain fMRI?

BIOL PSYCHIATRY 2008

Pattern Classification of Sad Facial Processing: Toward the Development of Neurobiological Markers in Depression

Cynthia H.Y. Fu, Janaina Mourao-Miranda, Sergi G. Costafreda, Akash Khanna, Andre F. Marquand, Steve C.R. Williams, and Michael J. Brammer

We applied SVM to classify depressed patients vs. healthy controls based on their pattern of activation for emotional stimuli (sad faces).

•19 free medication depressed patients vs. 19 healthy controls

•Event-related fMRI paradigm consisted of affective processing of sad facial stimuli with modulation of the intensity of the emotional expression (low, medium, and high intensity).

•Patterns: GLM coefficients, i.e. one example per subject

•Cross-validation framework: leave one subject per group out

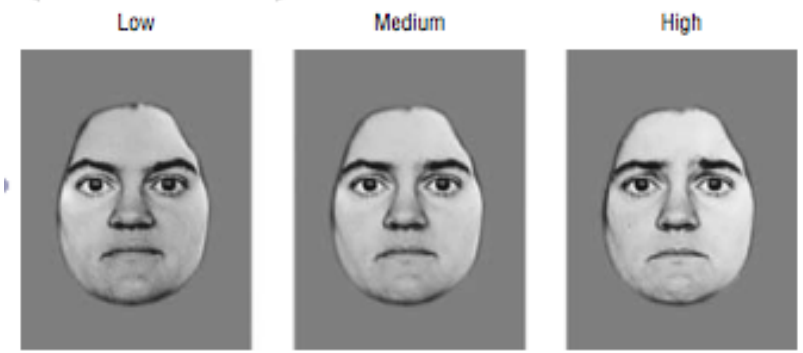
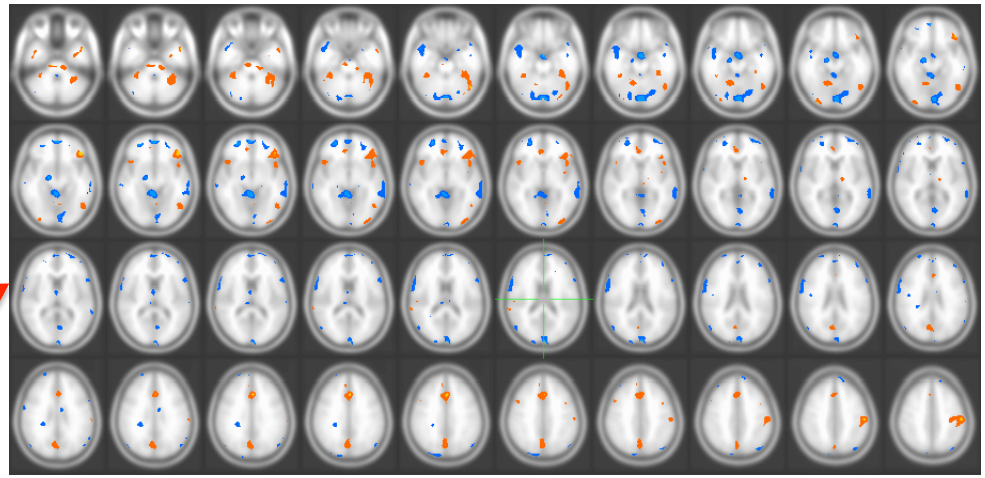


Table 2. Group Classification of Depressed and Healthy Individuals

Intensity of Sad Facial Expression	Sensitivity	Specificity	*p*-Value
Whole Brain Analysis			
Train with single events and test with mean image			
Low	58%	58%	ns
Medium	58%	58%	ns
High	74%	63%	.017
All faces	53%	58%	ns
Train and test with BOLD response convolution model			
Low	84%	89%	<.0001
Medium	68%	79%	.0030
High	68%	84%	.0008
All faces	72%	82%	.0008
Region of Interest Analysis			
Low	53%	58%	ns
Medium	58%	58%	ns
High	47%	58%	ns

BOLD, blood oxygenation level-dependent; ns, nonsignificant.

SVM weigh vector
(low intensity of sad facial expression)



The threshold value used was to 30% of the maximum (absolute) weight value.

Are patients outliers?

NeuroImage (2011)

Patient Classification as an Outlier Detection Problem: an Application of the One-Class Support Vector Machine

Janaina Mourao-Miranda[1,3], David R. Hardoon[1,3], Tim Hahn[2], Andre F. Marquand[3], Steve C R Williams[3], John Shawe-Taylor[1], Michael Brammer[3]

- The OC-SVM is a special case of the SVM algorithm for novelty or outlier detection (Scholkopf et al., 2001):
  - ✓ Finds in a general kernel defined feature space the smallest hypersphere containing most of the data;
  - ✓ Detects an outlier when a new data lies outside the hypersphere;

- Uses a training set to learn the support of the distribution of the 'normal examples'.

- After the training the resulting pattern function identify in the test set any abnormal example that appears not to have been generated from the same distribution.

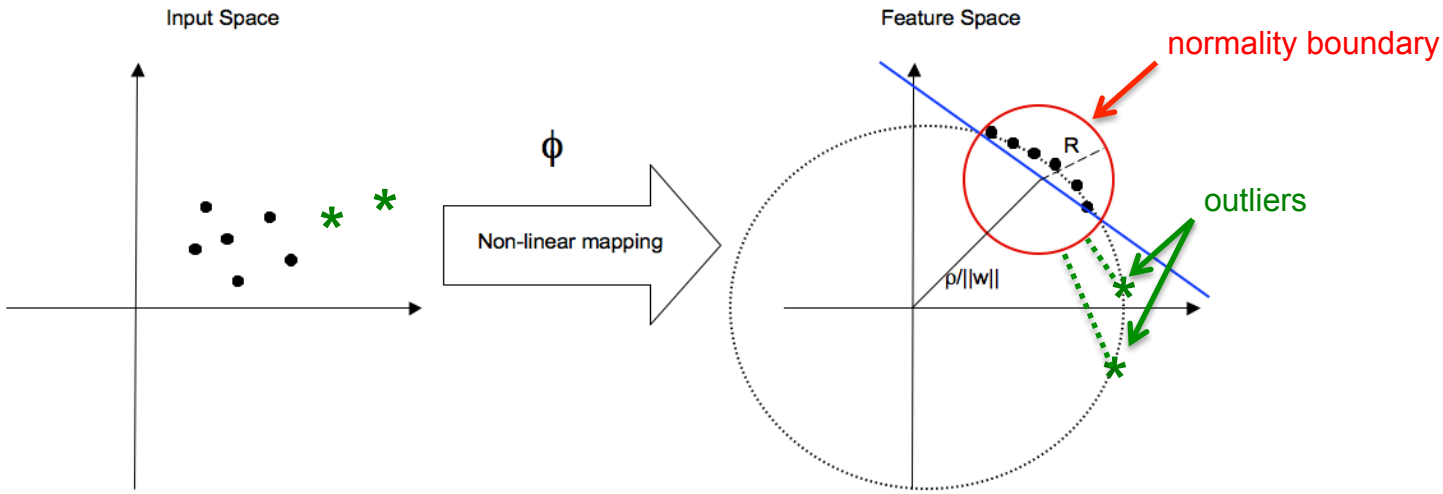
- The purpose of the OC-SVM algorithm is to estimate a decision function or boundary (hypershere) $f(\mathbf{x})$ that takes the value +1 in a small region capturing most of the training examples, and -1 elsewhere.

NeuroImage (2011)

Hyperspheres and Hyperplanes

•If the data is normalized it can be viewed as lying on the surface of a hypersphere in the feature space.

•In this case there is a correspondence between hyperspheres in the feature space and hyperplanes: since the decision boundary is determined by the intersection of the two hypersphere, it can equally be described by the intersection of a hyperplane with the unit hypersphere.



NeuroImage (2011)

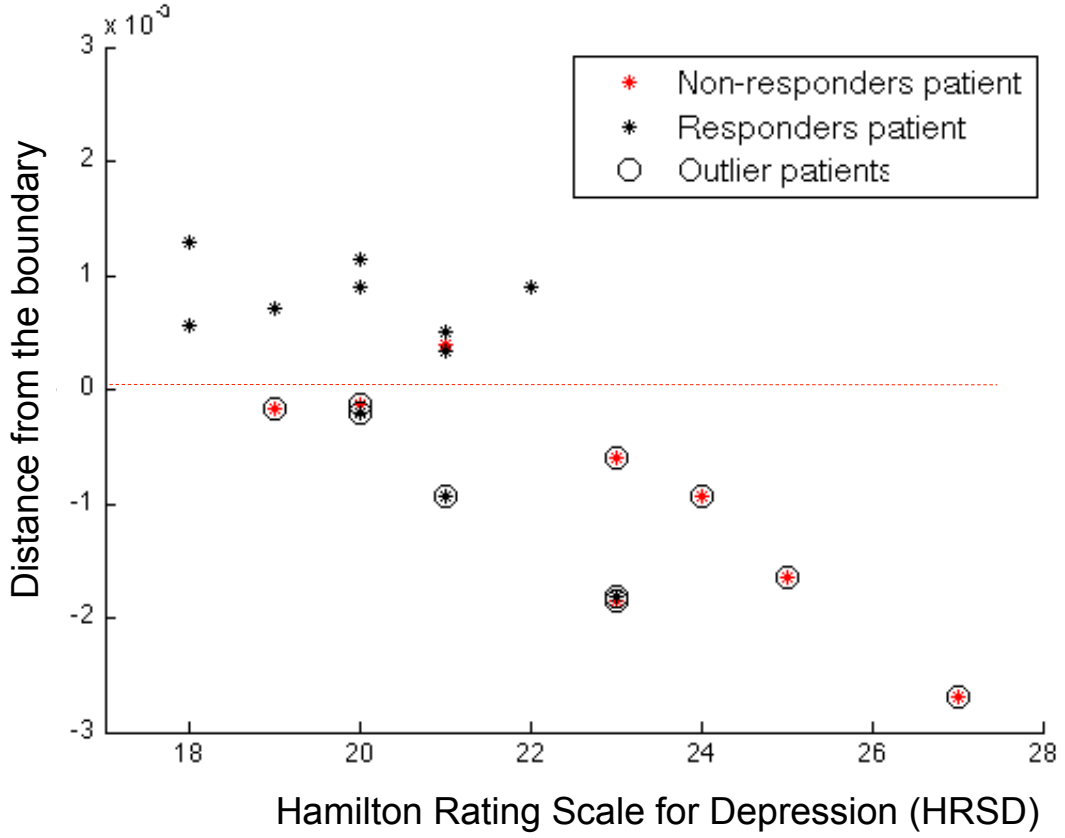
We applied the OC-SVM with Radial Basis Function (RBF) Kernel to investigate three hypotheses:

• The pattern of fMRI response to sad faces in healthy subjects is homogeneous enough to enable the definition of a **"normality boundary".**

• This pattern is altered in depressed patients.

• The amount of departure from the "normality boundary" as measured by the OC-SVM is related to the severity of the depression.

NeuroImage (2011)



Main Results

•Correlation between OC-SVM predictions and Hamilton Rating Scale for Depression (HRSD) = -0.81 ($p<0.001$).

•79% of controls were detected as non-outlier.
•52% of patients were detected as outlier.

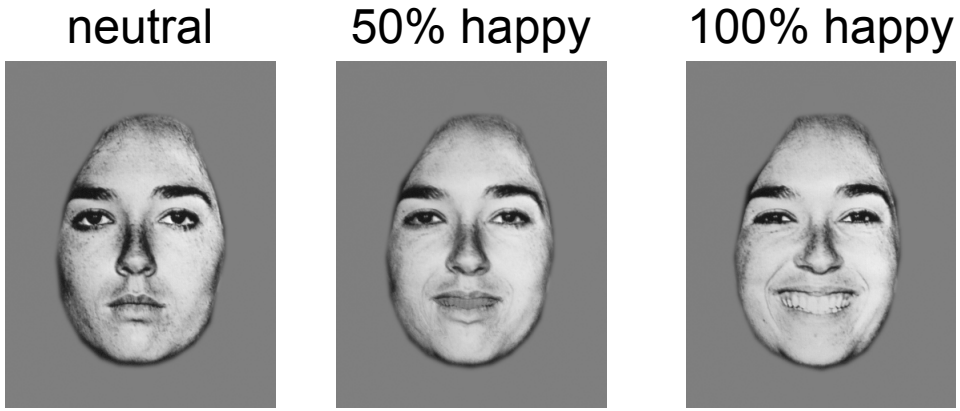
•89% patients classified as non-outliers respond to treatment
•70% of patients classified as outliers did not respond to treatment.

•The OC-SVM split the patient groups into two subgroups whose membership were associated with future response to treatment.

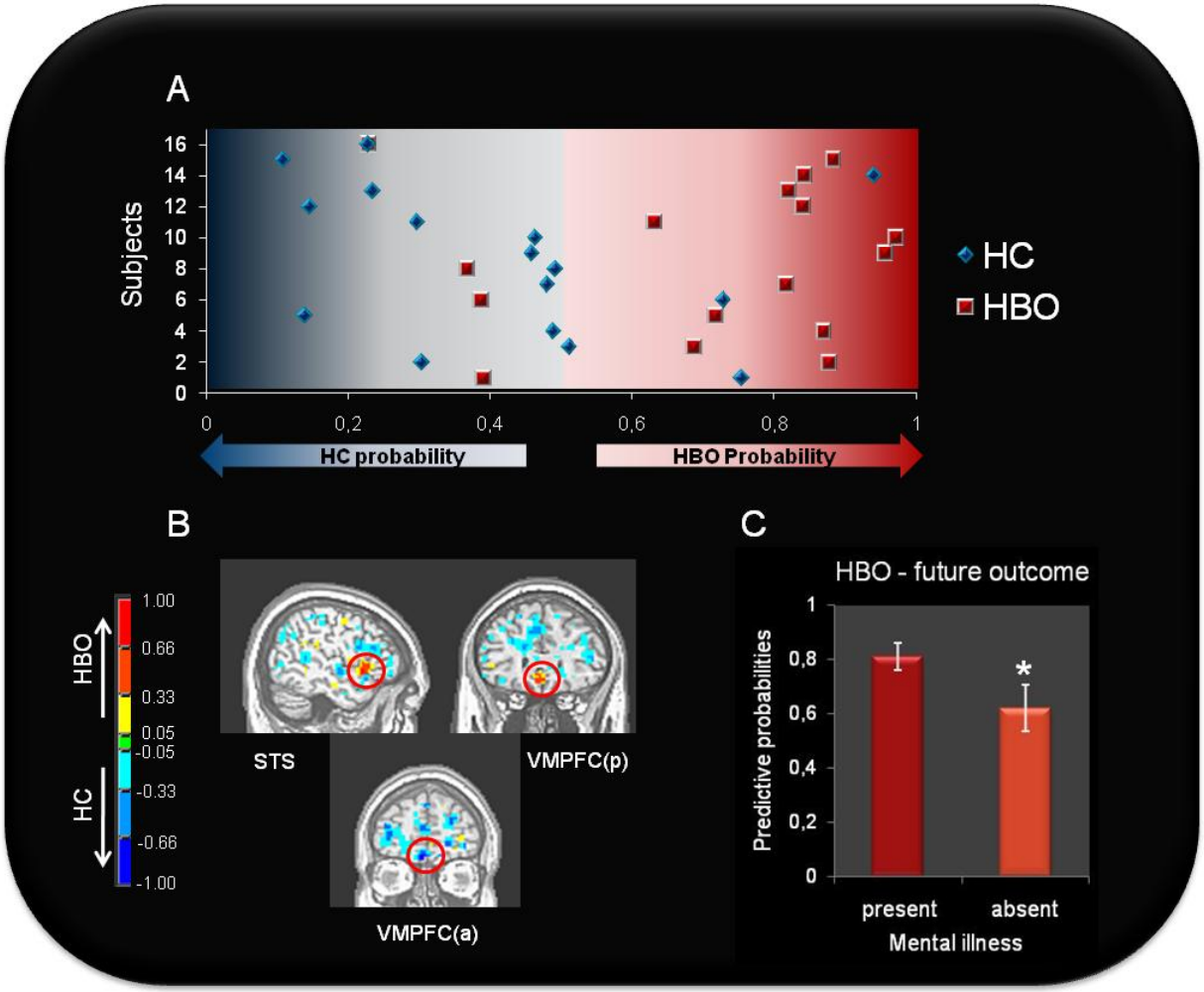
Can we predict future mental illness in at-risk adolescents ?

Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents.

- 16 healthy bipolar offspring (HBO) and 16 healthy, age- and gender ratio-matched healthy controls (HC).
- Event-related experiment consisting of affective processing of **happy facial** stimuli with modulation of the intensity of the emotional expression (neutral, 100% happy, and 50% happy).
- Gender label task.
- Patterns: GLM coefficients, i.e. one example per subject.
- Classifier: Gaussian Process & Recursive Feature Elimination



Collaboration with Prof. Mary Phillips and Dr. Leticia Oliveira



•Using GPC wholebrain activity to neutral faces accurately and significantly differentiated HBO from HC with 75% of accuracy (sensitivity =75%, specificity =75%, permutation test p=0.008).

•The predictive probabilities of HBO who developed depression or anxiety disorders, were significantly higher for these 6 HBO than for HBO remaining healthy at follow-up (up to 4 years').

Summary

- Pattern classification applied to whole brain fMRI data can be used for diagnosis and prognosis.

- Pattern recognition analysis can be used to detect patients as outliers in relation to a “healthy control” pattern.

- This demonstrates the potential clinical relevance of pattern classification approaches in neuroimaging.

Acknowledgements

- Prof. John Shawe-Taylor (Computer Science Department, UCL)
- Prof. Steve Williams (Department of Neuroimaging, KCL)
- Prof. Mary Phillips (Department of Psychiatry, Pittsburgh University)
- Leticia Oliveira, PhD
- Andre Marquand, PhD

- Wellcome Trust
- Pascal 2