

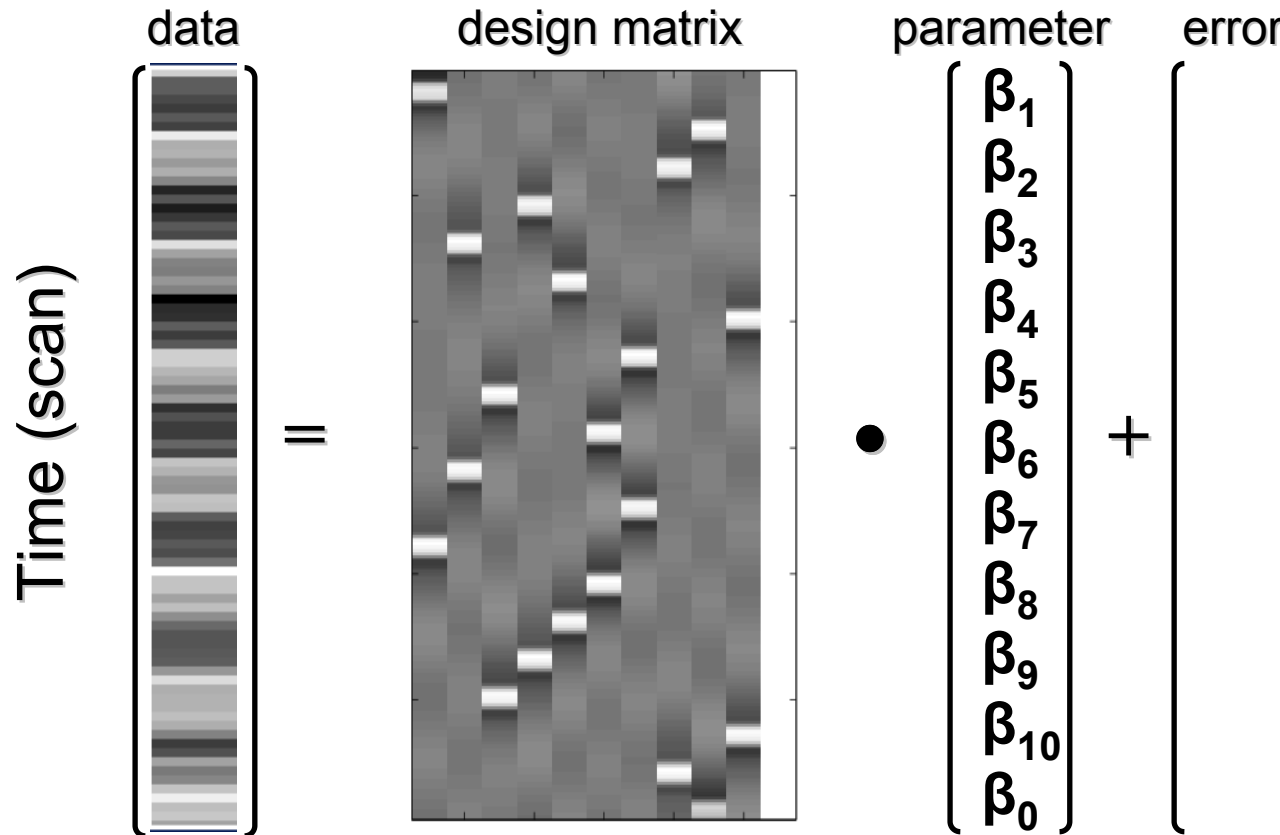
# Multivariate pattern classification



**Thomas Wolbers**  
**Space & Ageing Laboratory ([www.sal.mvm.ed.ac.uk](http://www.sal.mvm.ed.ac.uk))**  
**Centre for Cognitive and Neural Systems &**  
**Centre for Cognitive Ageing and Cognitive Epidemiology**

- **WHY PATTERN CLASSIFICATION?**
- PROCESSING STREAM
- FEATURE REDUCTION
- CLASSIFICATION
- EVALUATING RESULTS

# Why pattern class.?

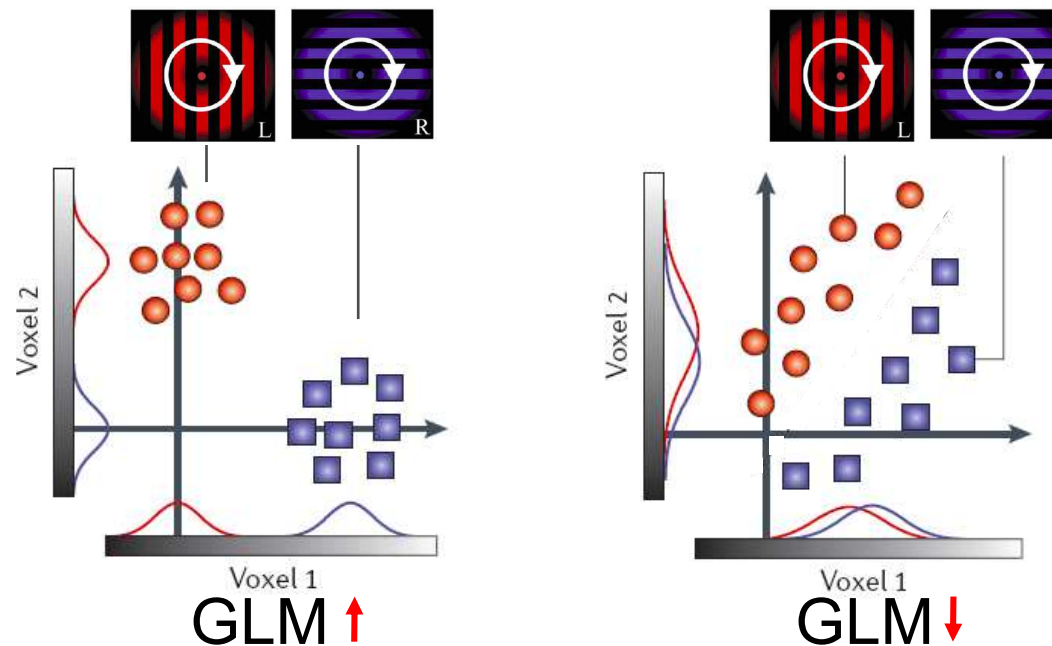


$$y = X \cdot \beta + \varepsilon$$

**GLM: separate model fitting for each voxel**  
**→ mass-univariate analysis!**

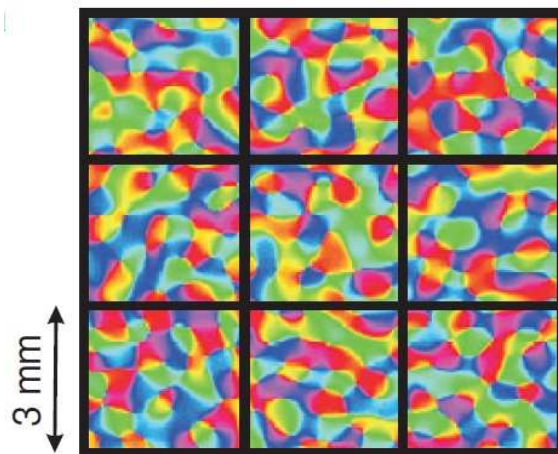
## Key idea behind pattern classification

- GLM analysis relies exclusively on the information contained in the time course of **individual** voxels
- When signal distributions for different experimental conditions show large overlap, GLM unable to detect reliable differences

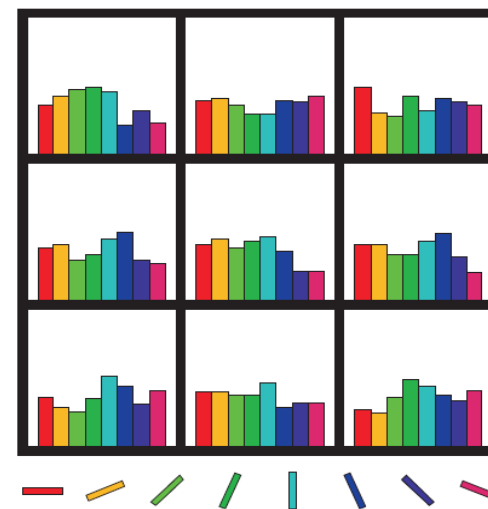


## Key idea behind pattern classification

- GLM analysis relies exclusively on the information contained in the time course of **individual** voxels
  - When signal distributions for different experimental conditions show large overlap, GLM unable to detect reliable differences
  - Multivariate analyses take advantage of the information contained in activity patterns **across multiple voxels**
- *Multivariate approaches can exploit sampling biases*

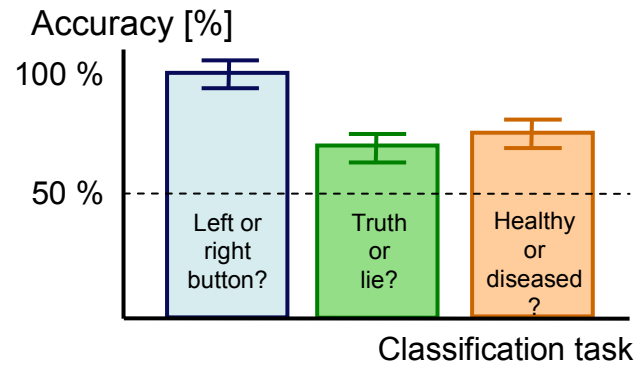


Boynton (2005). *Nat Neurosci*

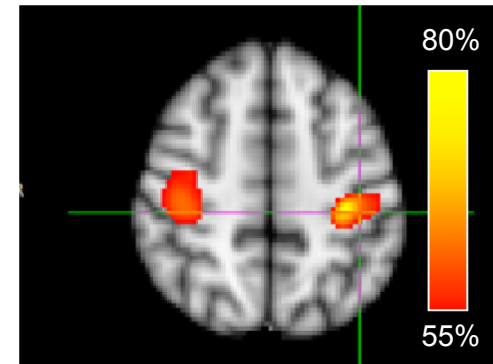


# Decoding – key questions

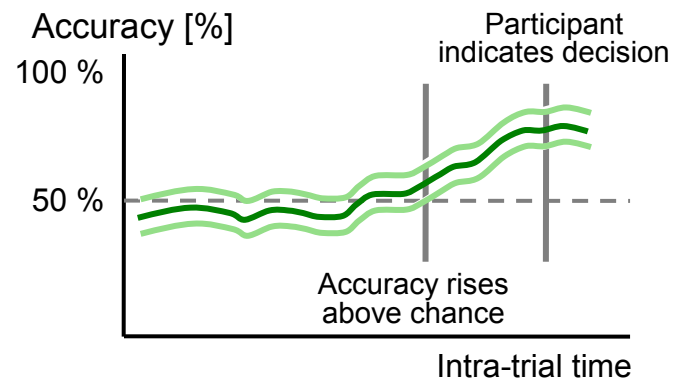
**(a) X-Y mapping overall reliability**



**(b) X-Y mapping spatial deployment**

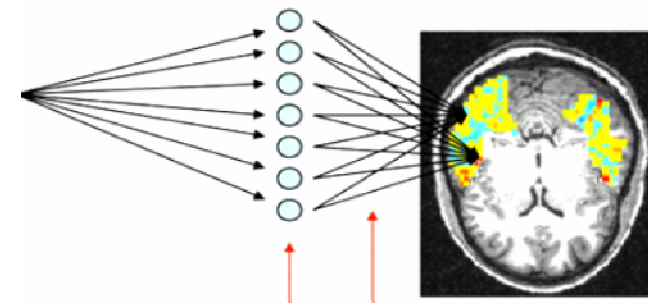


**(c) X-Y mapping temporal evolution**



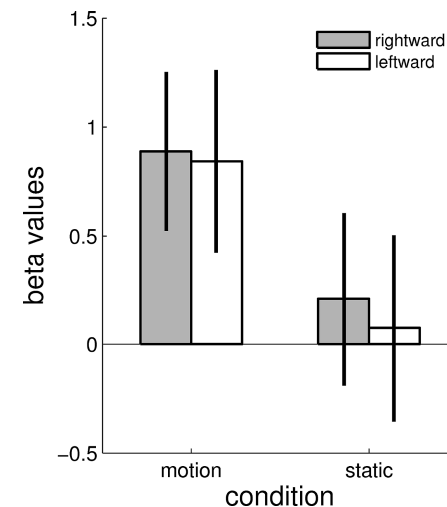
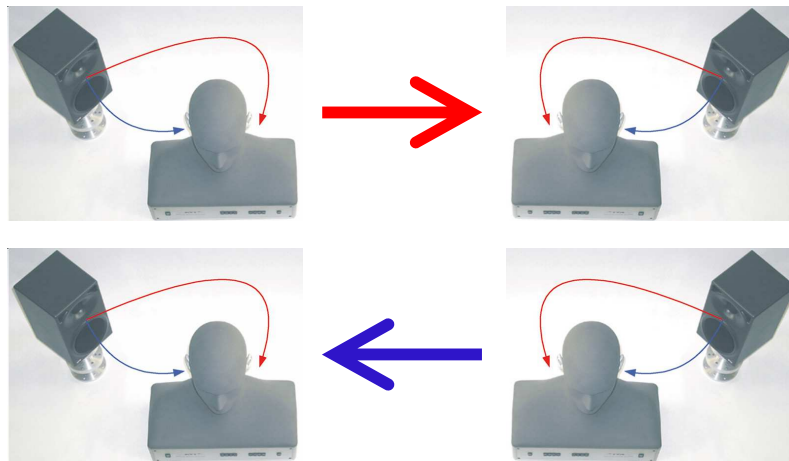
**(d) X-Y mapping: subtle issues**

- functionally selective vs segregated representations
- degenerative (many-to-one) structure-function mappings



- WHY PATTERN CLASSIFICATION?
- **PROCESSING STREAM**
- PREPROCESSING / FEATURE REDUCTION
- CLASSIFICATION
- EVALUATING RESULTS

# AUDITORY MOTION PERCEPTION IN THE BLIND



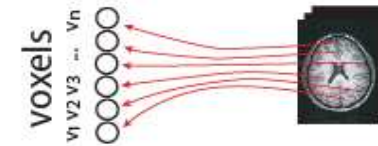
Can the direction of auditory motion be decoded from fMRI signals in the human motion complex (hMT+)?

Wolbers et al. (2011)

# Processing stream

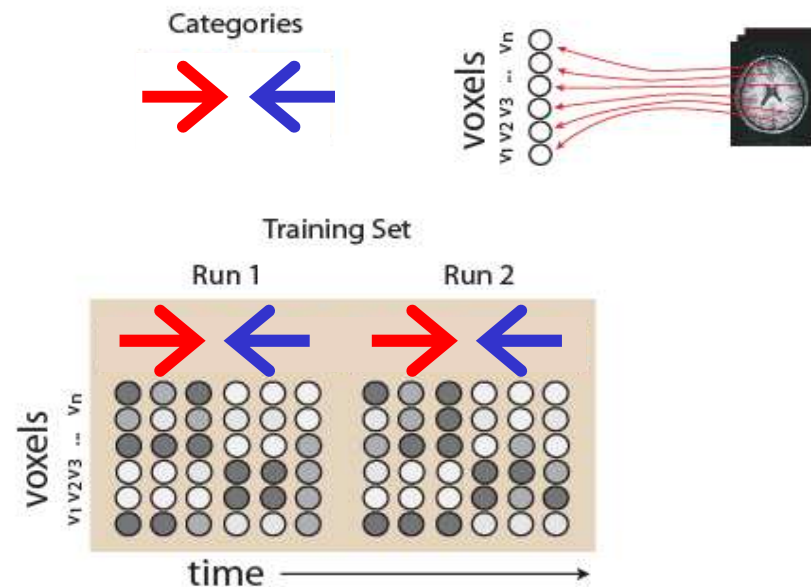
SPM Kurs HH 09/11

1. Acquire fMRI data
2. Preprocess fMRI data
3. Select relevant features  
(i.e. voxels)



# Processing stream

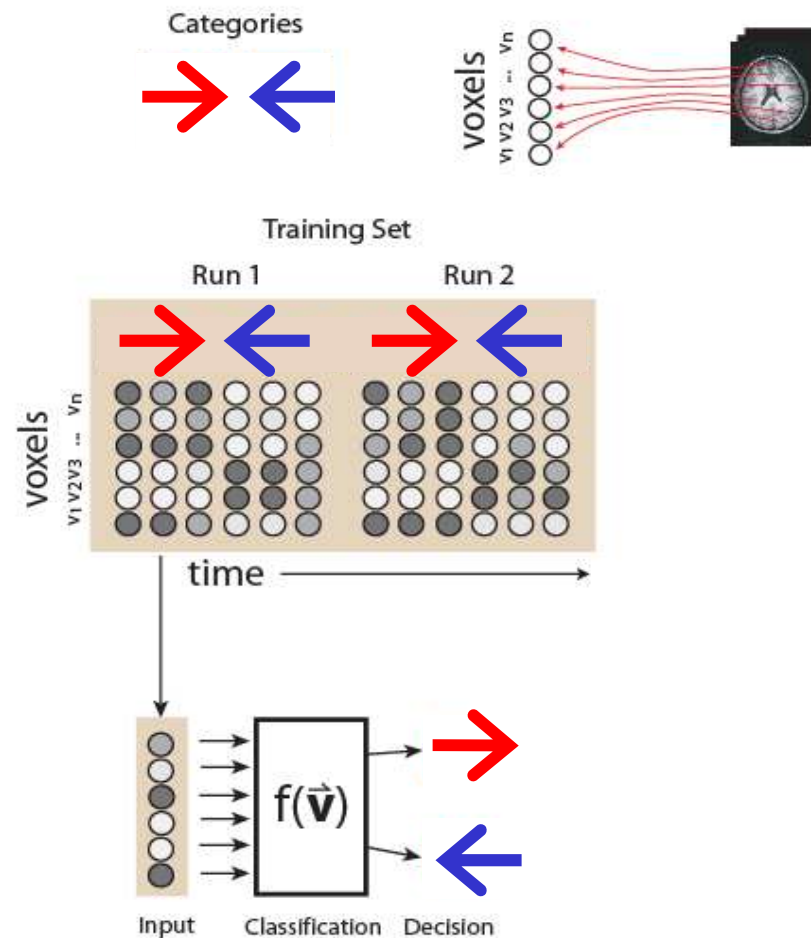
1. Acquire fMRI data
2. Preprocess fMRI data
3. Select features
4. Label fMRI patterns according to whether the subject was hearing leftward or rightward motion (adjusting for hemodynamic lag)



# Processing stream

SPM Kurs HH 09/11

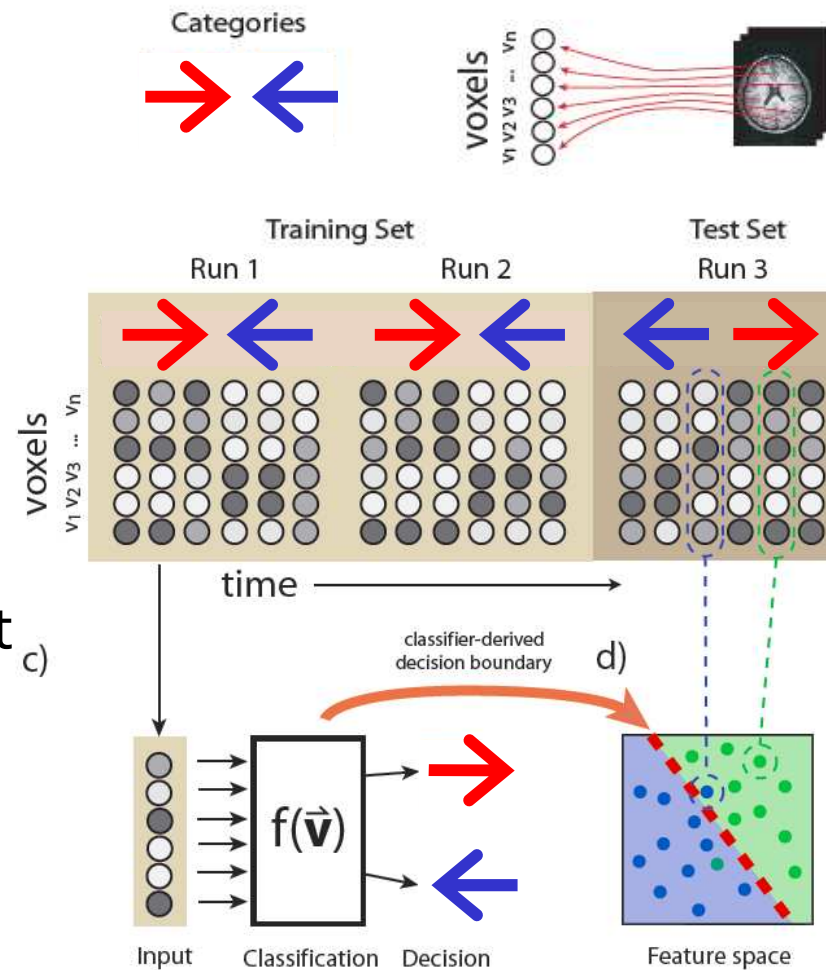
1. Acquire fMRI data
2. Preprocess fMRI data
3. Select features
4. Label fMRI patterns
5. Train a classifier to discriminate between leftward and rightward patterns



# Processing stream

SPM Kurs HH 09/11

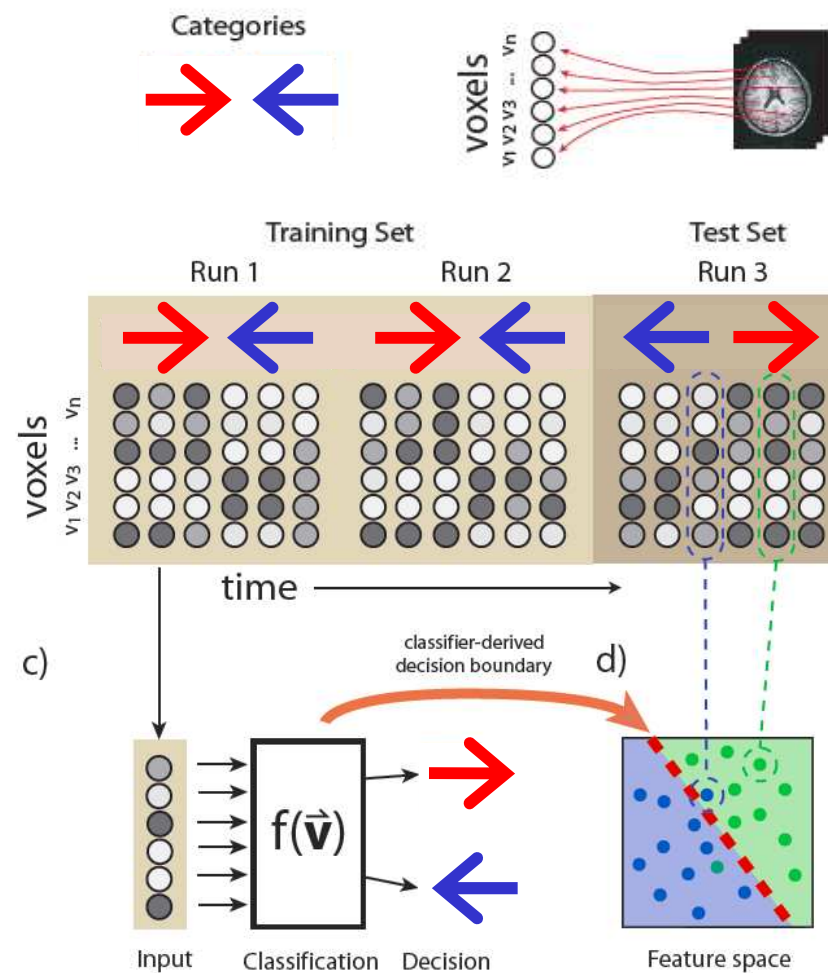
1. Acquire fMRI data
2. Preprocess fMRI data
3. Select features
4. Label fMRI patterns
5. Train the classifier
6. Apply the trained classifier to new fMRI patterns (not presented at training).



# Processing stream

SPM Kurs HH 09/11

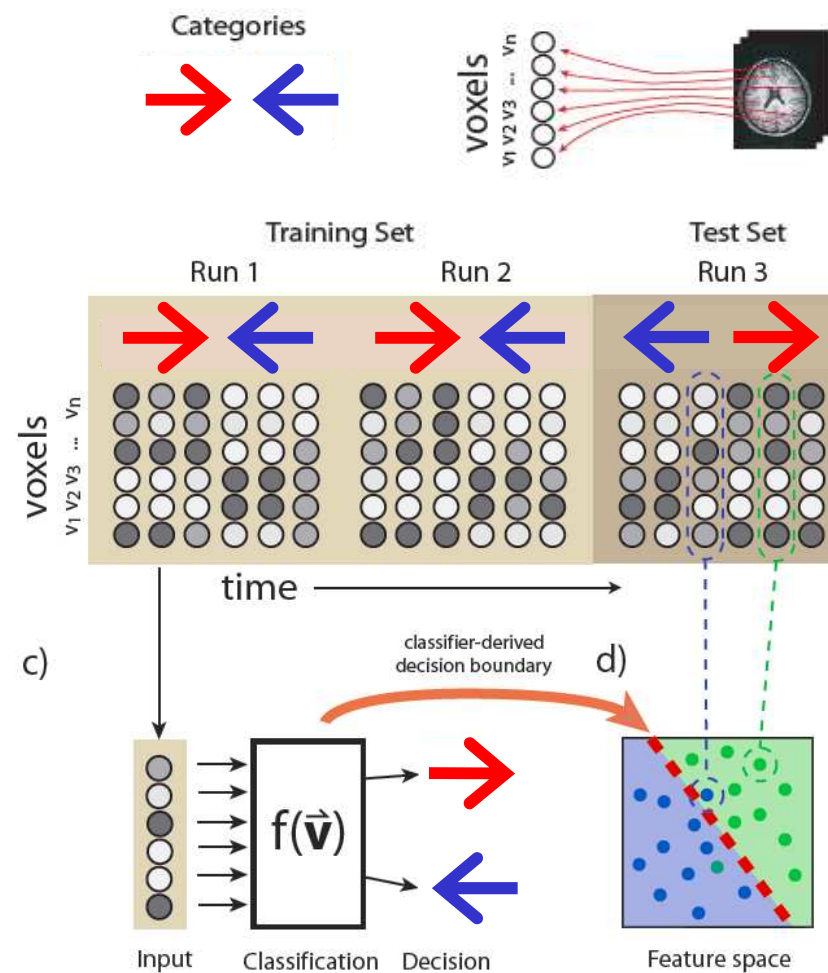
1. Acquire fMRI data
2. Preprocess fMRI data
3. Select features
4. Label fMRI patterns
5. Train the classifier
6. Apply the trained classifier to new fMRI patterns (not presented at training).
7. Crossvalidation



# Processing stream

SPM Kurs HH 09/11

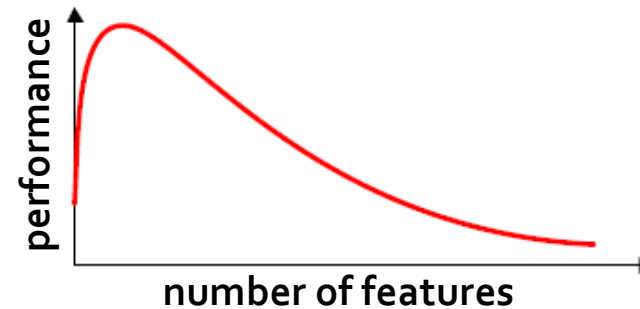
1. Acquire fMRI data
2. Preprocess fMRI data
3. Select features
4. Label fMRI patterns
5. Train the classifier
6. Apply the trained classifier to new fMRI patterns (not presented at training).
7. Crossvalidation
8. Statistical inference



- WHY PATTERN CLASSIFICATION?
- PROCESSING STREAM
- **FEATURE REDUCTION**
- CLASSIFICATION
- EVALUATING RESULTS

## The problem

- fMRI data are typically sparse, high-dimensional and noisy
- Classification is sensitive to information content in all voxels  
→ many uninformative voxels = poor classification (i.e. due to overfitting)



## Solution 1: Feature selection



- select subset with the most informative features
- original features remain unchanged

## 'External' Solutions

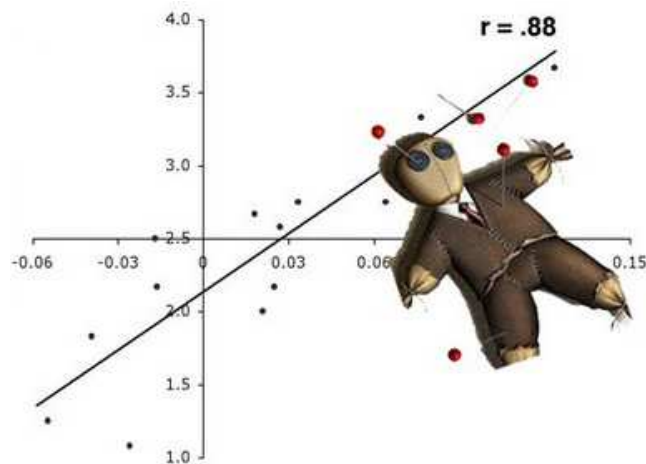
- Anatomical regions of interest
- Independent functional localizer (i.e. retinotopic mapping to identify early visual areas)

## Solutions using the data to be classified

- Searchlight classification: define region of interest (i.e. sphere) and move it across the search volume
- univariate tests, i.e. activation vs. baseline (t-Test), mean difference between conditions (ANOVA)
- recursive feature elimination (De Martino et al, 2008): eliminate voxels based on discriminative weights

## Peeking #1

- testing a trained classifier needs to be performed on *independent* test datasets
- if entire dataset is used for feature selection, ...



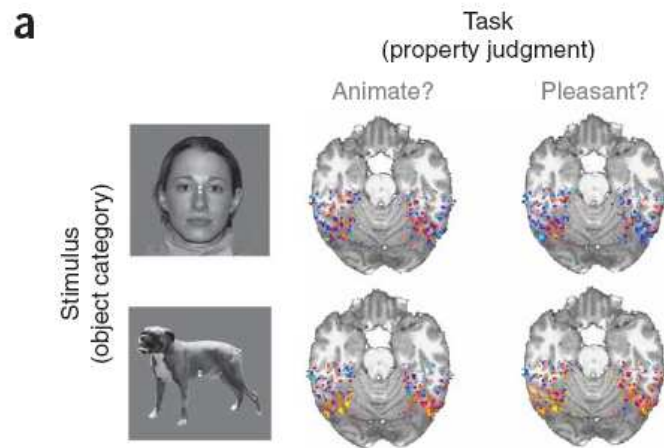
Puzzlingly High Correlations  
in fMRI Studies of Emotion,  
Personality, and Social  
Cognition<sup>1</sup>

Edward Vul,<sup>1</sup> Christine Harris,<sup>2</sup> Piotr Winkielman,<sup>2</sup> & Harold Pashler<sup>2</sup>

Circular analysis in systems neuroscience:  
the dangers of double dipping

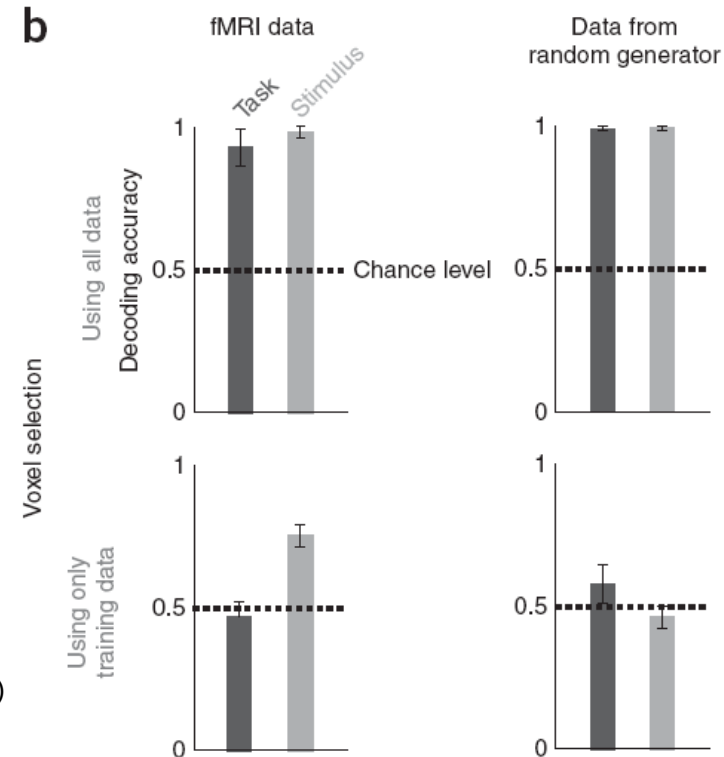
Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan & Chris I Baker

# Feature selection



ROI definition in inferior temporal cortex based on two sided t-tests comparing conditions

Kriegeskorte et al. (2009)  
*Nat Neurosci*



→ if entire dataset is used for feature selection, training and test data are no longer independent => overfitting, optimisation of decision function (partially) based on noise, classification accuracy become overly optimistic

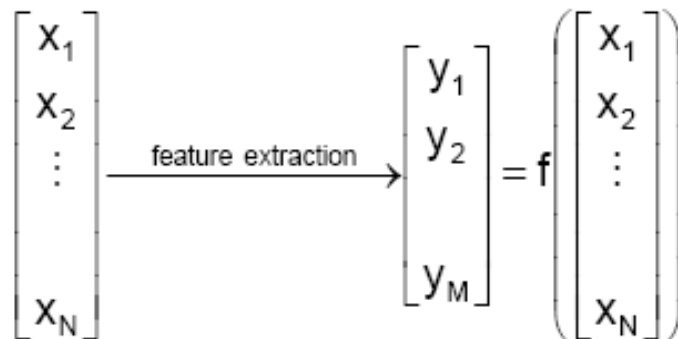
→ **nested crossvalidation**

## Solution 1: Feature selection



- select subset from all available features
- original features remain unchanged

## Solution 2: Feature extraction



- create new features as a function of existing features
- Linear functions (PCA, ICA, ...)
- Nonlinear functions during classification (i.e. hidden units in a neural network)

- WHY PATTERN CLASSIFICATION?
- PROCESSING STREAM
- FEATURE REDUCTION
- **CLASSIFICATION**
- EVALUATING RESULTS

**Discriminative classifiers**, e.g. Logistic Regression, SVM:

- $X$  – data;  $C$  - classes
- Assume some functional form for  $P(C|X)$ , e.g.:

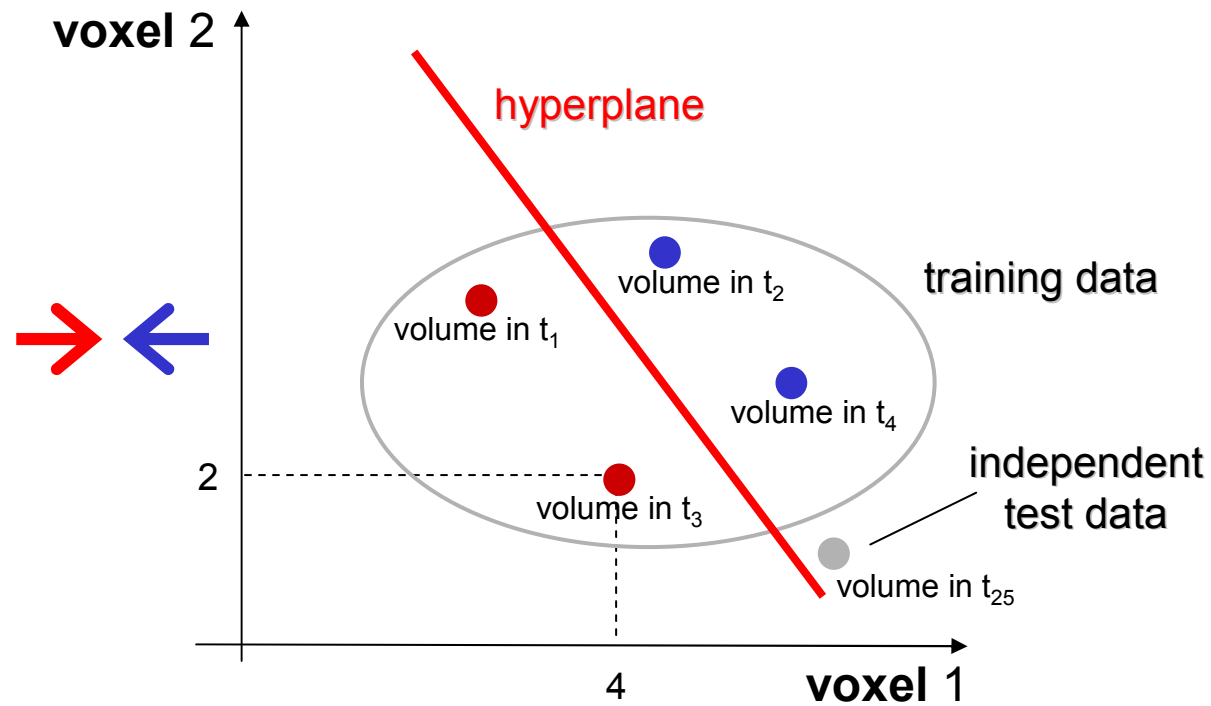
$$P(C / X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

- Estimate parameters of  $P(C|X)$  directly from training data

**Generative classifiers**, e.g. Gaussian Naïve Bayes, LDA:

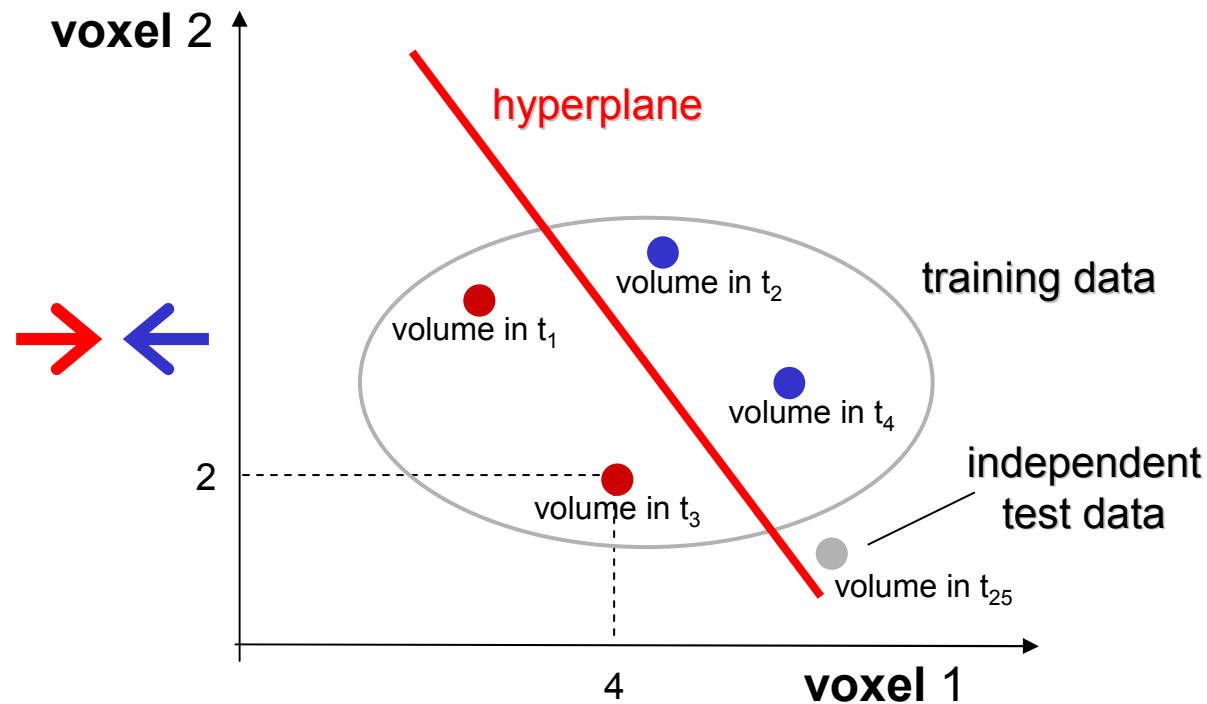
- Assume some functional form for  $P(X|C)$ ,  $P(C)$
- Estimate parameters of class conditional probabilities  $P(X|C)$  (and priors  $P(C)$ ) directly from training data
- Use Bayes rule to calculate  $P(C|X)$
- **Indirect** computation of  $P(C|X)$  through Bayes rule

## The Decoding problem



- our task: find a *hyperplane* that separates both conditions

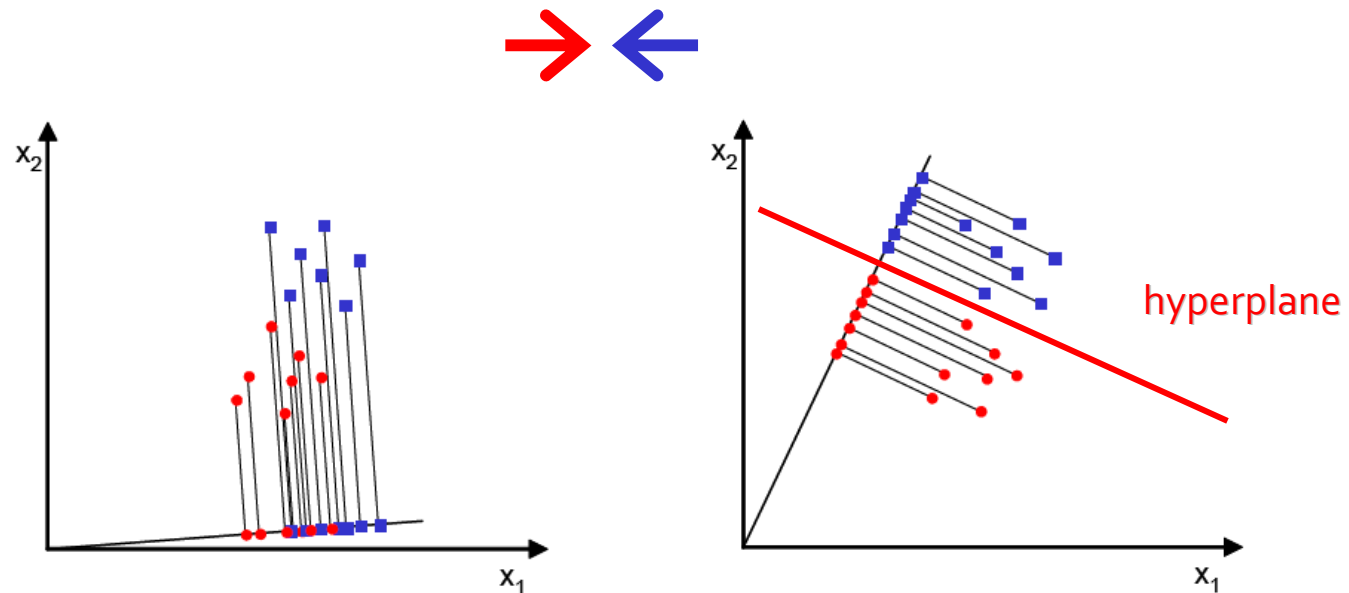
## The Decoding problem



a linear decision function:  $y = f(x) = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$

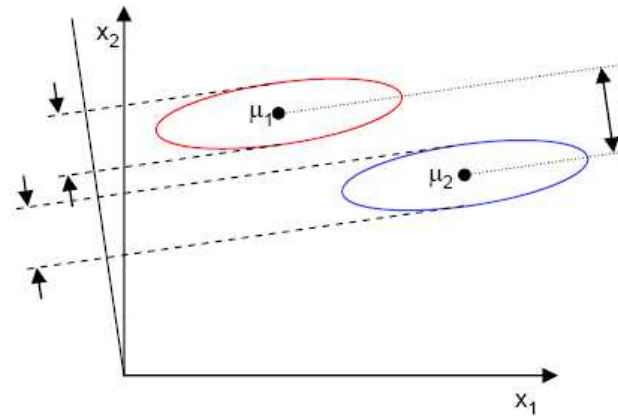
- if  $y < 0$ , predict **rightward** // if  $y > 0$ , predict **leftward**
- prediction = linear function of features

## Fisher Linear Discriminant (FLD)



- Project data on a new axis that maximizes the class separability
- Hyperplane is orthogonal to the best projection axis

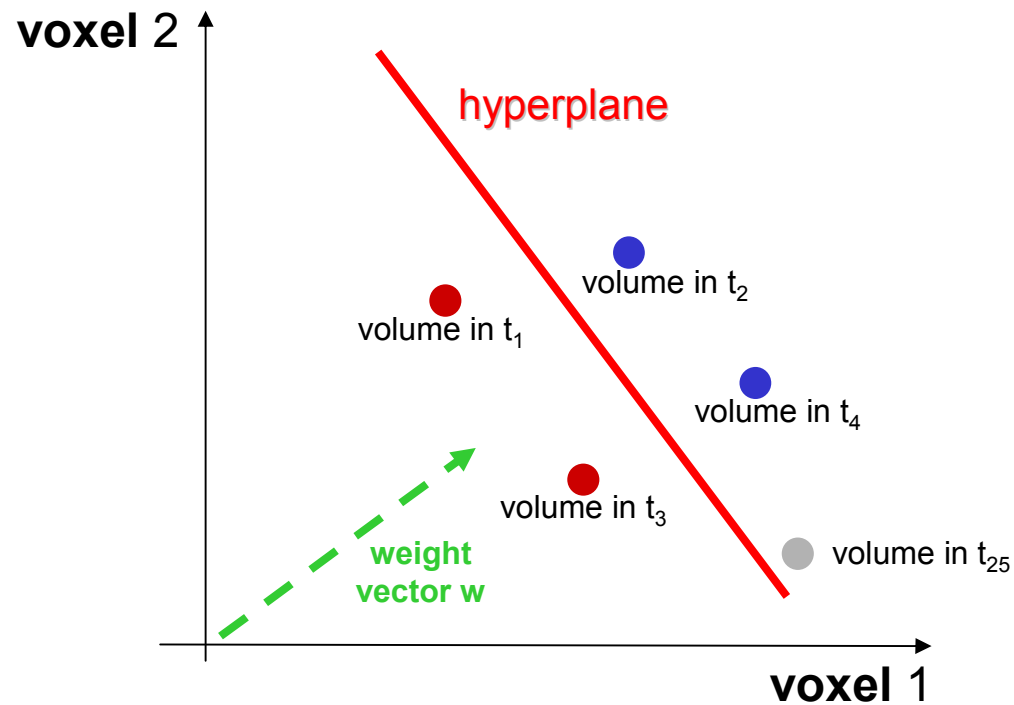
## Fisher Linear Discriminant (FLD)



- FLD classifies by projecting the training set on the axis that is defined by the difference between the center of mass for both classes, corrected by the within class scatter
- separation is maximised for:

$$w = \frac{m_1 - m_2}{\frac{1}{N_1} \sum (x_n - m_1)(x_n - m_1)^T + \frac{1}{N_2} \sum (x_n - m_2)(x_n - m_2)^T}$$

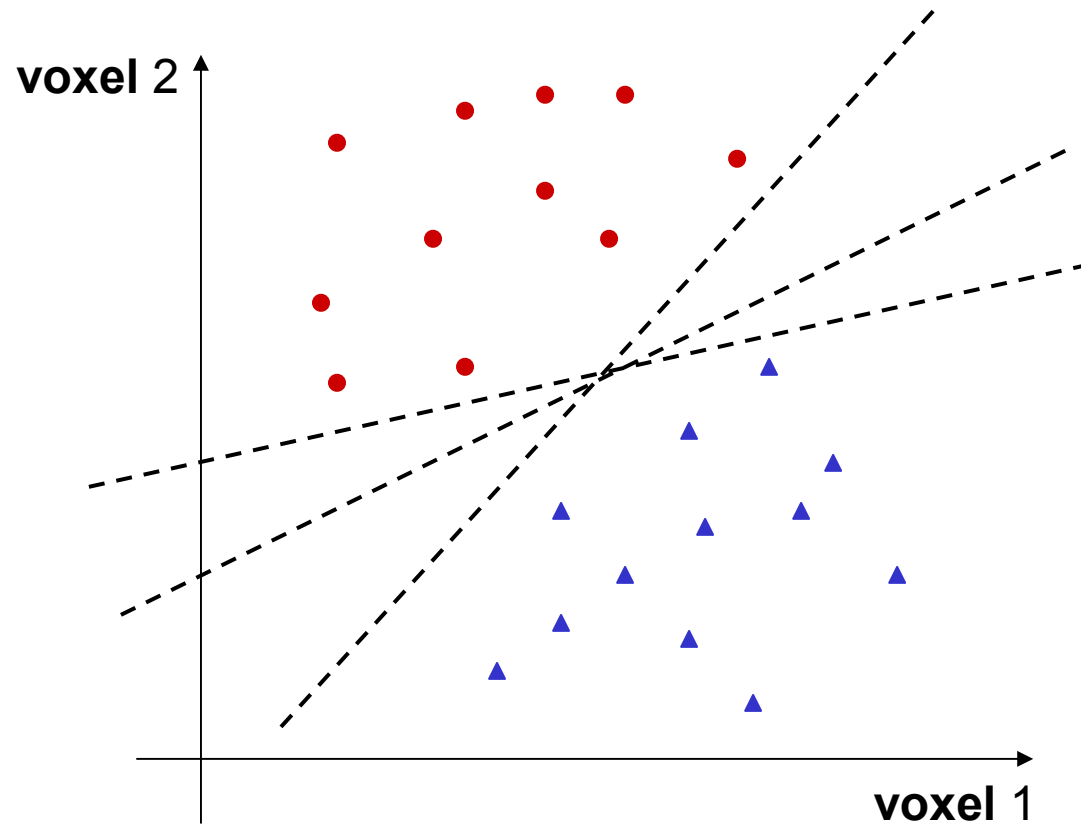
## Fisher Linear Discriminant (FLD)



$$y = \mathbf{w}\mathbf{x} + b$$

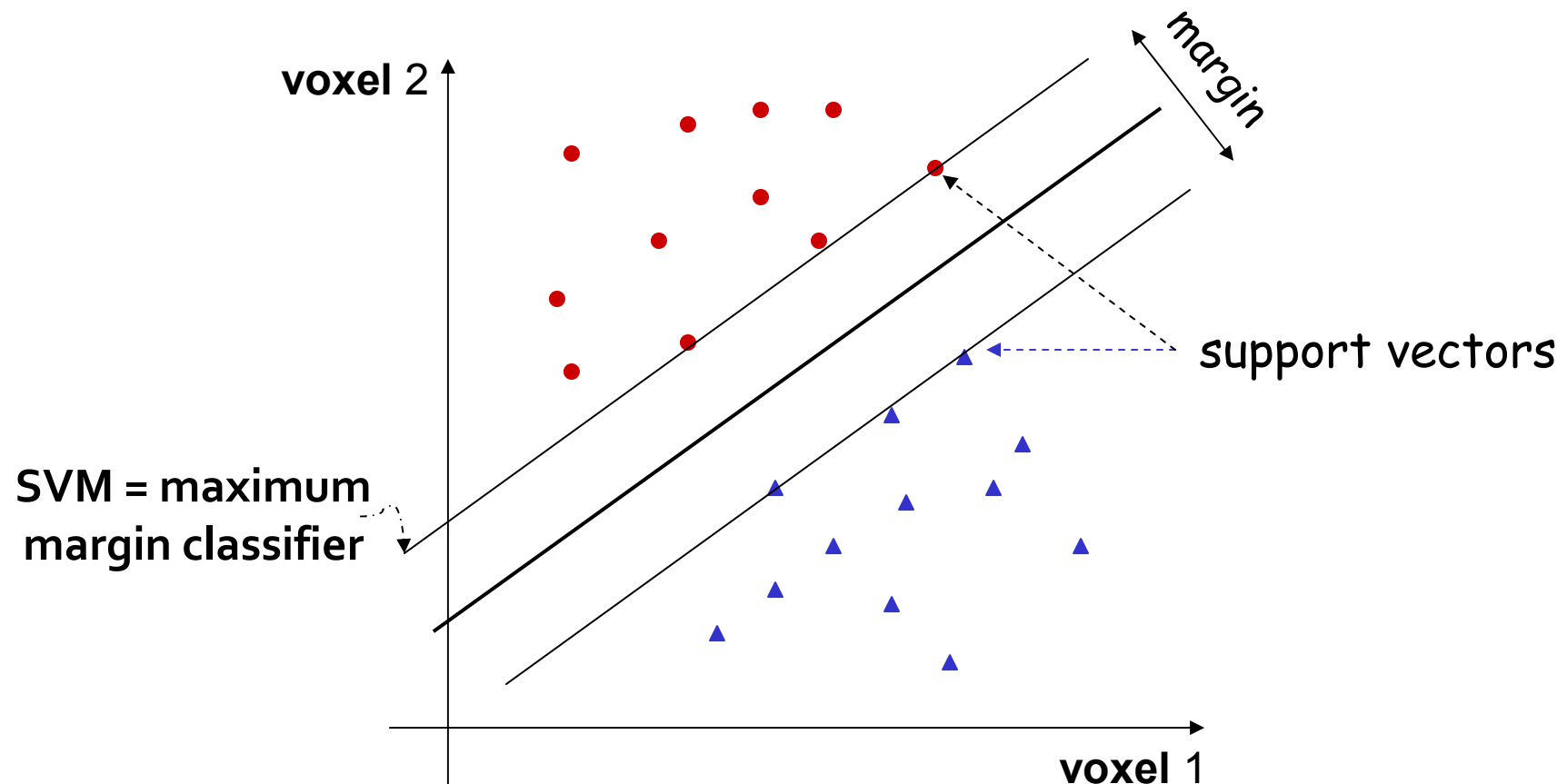
- orientation of the hyperplane defined by weight vector  $w$  (normal vector with length 1)
- $b$ : offset of the hyperplane from the origin along the weight vector

## Support Vector Machine (SVM)



Which of the linear separators is the optimal one?

## Support Vector Machine (SVM)

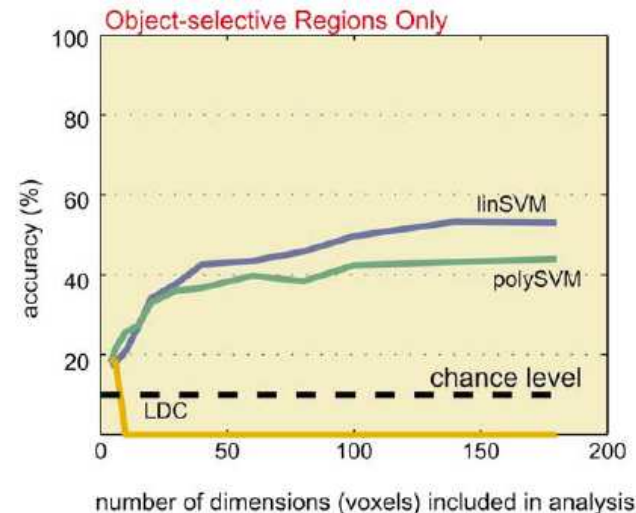
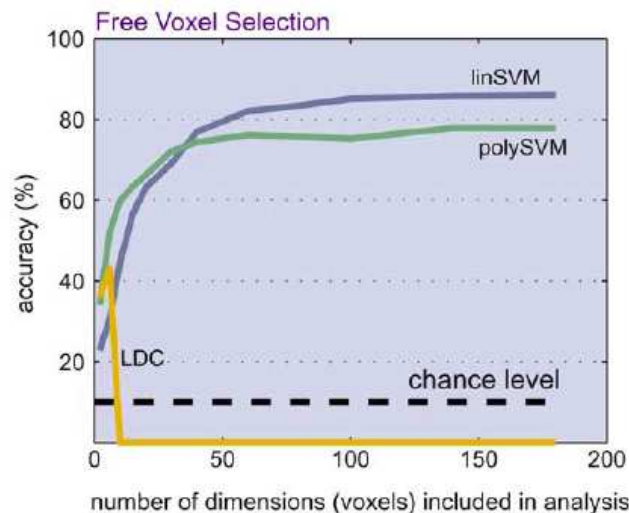


Usually, classes have overlapping distributions: SVM's account for misclassification errors by introducing additional slack variables

## How to choose the right classifier?

**Situation 1: scans ↓, features ↑ (i.e. whole brain data)**

- FLD unsuitable: depends on reliable estimation of covariance matrix
- GNB inferior to SVM and LR → the latter come with regularisation that help weigh down the effects of noisy and highly correlated features

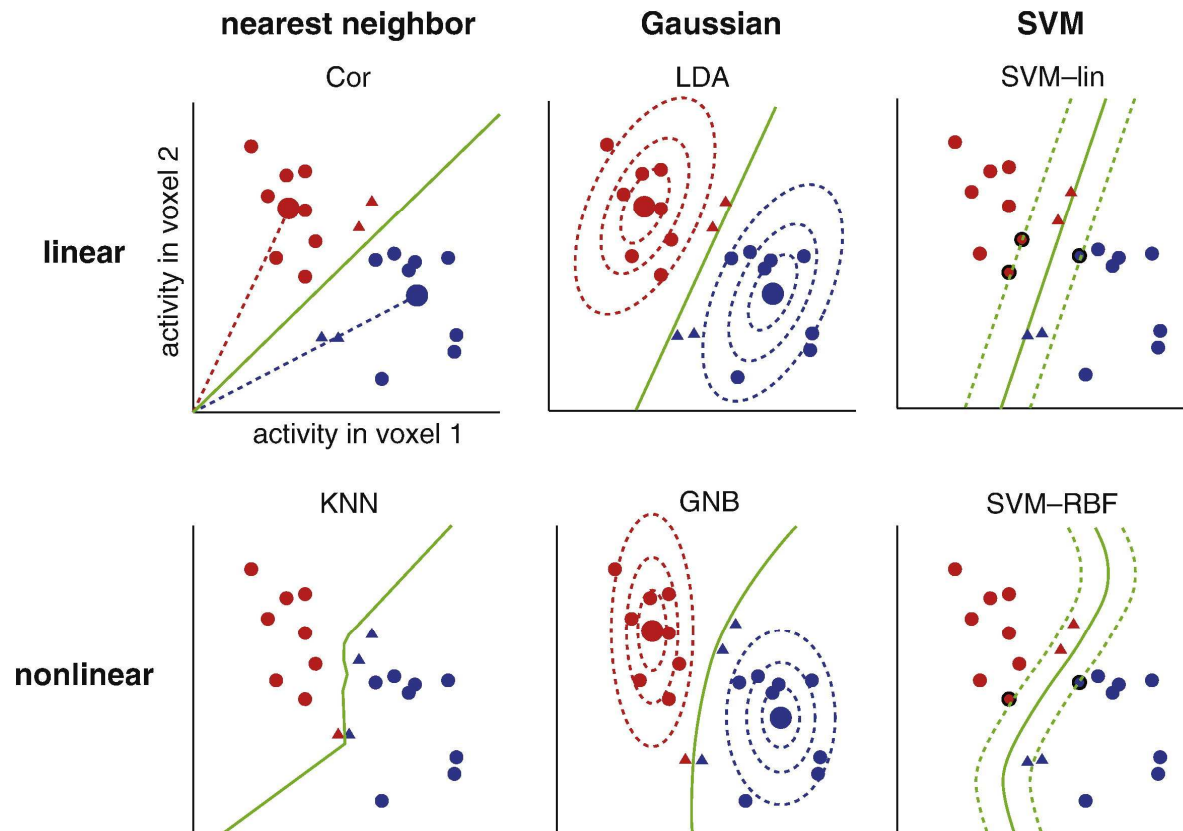


Cox & Savoy (2003). NeuroImage

***Situation 2: scans ↓, features ↓ (i.e. feature selection or feature extraction)***

- GNB, SVM and LR: often similar performance
- SVM originally designed for two-class problems only
- SVM for multiclass problems: multiple binary comparisons, voting scheme to identify classes
  
- accuracy of SVM increases faster than GNB when the number of scans increase
- see Mitchell et al. (2005) and Misaki et al. (2010) for further comparisons between different classifiers

# Classification



Misaki et al. (2010)  
*NeuroImage*

- non-linear decision boundaries can adapt to the idiosyncrasies of the noise in the training data → overfitting, poor generalisation!
- particularly problematic when number of data points is low and number of features is high

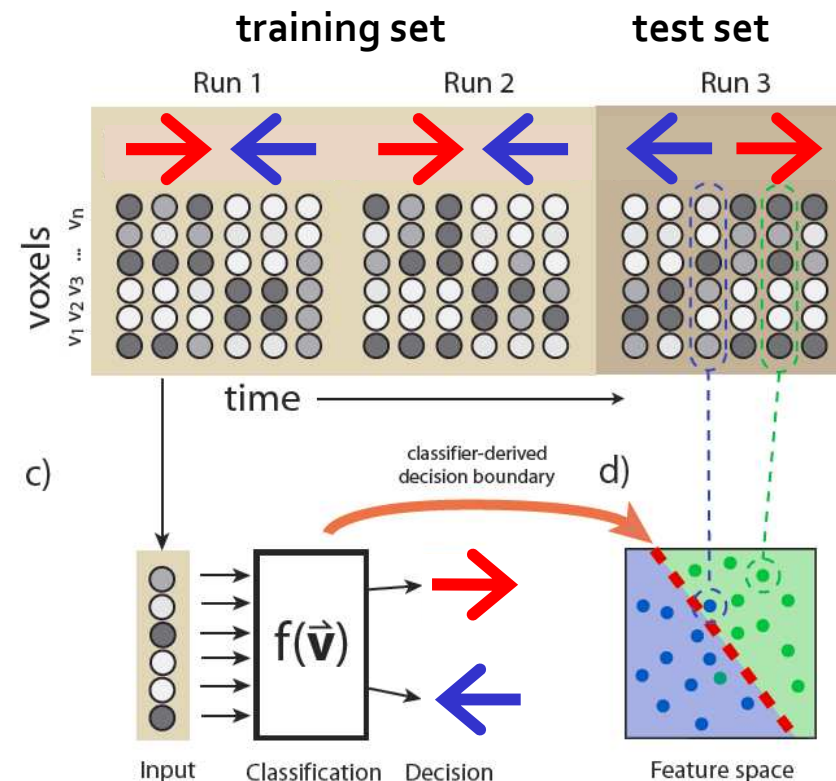
## Peeking #2

- classifier performance = unbiased estimate of classification accuracy
- ➔ how well would the classifier label a new example randomly drawn from the same distribution?
- testing a trained classifier needs to be performed on a dataset the classifier has never seen before
- ➔ if entire dataset is used for training a classifier, classification estimates become overly optimistic

**Solution: leave-one-out crossvalidation**

## Crossvalidation

- standard approach: leave-one-out crossvalidation
- split dataset into  $n$  folds (i.e. runs)
- train classifier on  $1:n-1$  folds
- test the trained classifier on fold  $n$
- rerun training/testing while withholding a different fold
- repeat procedure until each fold has been withheld once
- Classification accuracy usually computed as mean accuracy



- WHY PATTERN CLASSIFICATION?
- PROCESSING STREAM
- FEATURE REDUCTION
- CLASSIFICATION
- **EVALUATING RESULTS**

## Can I publish my data with 57% classification accuracy in Science or Nature?

### Independent test data

- Classification accuracy = unbiased estimate of the true accuracy of the classifier
- Question: what is the probability of obtaining 57% accuracy under the null hypothesis (no information about the variable of interest in my data)?
- Binary classification: p-value can be calculated under a binomial distribution with  $n$  (# of trials),  $p$  (probability of success) and  $k$  (number of correctly labeled examples)

$$F(x, n, k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}$$

## Nonparametric approaches

Permutation tests (i.e. Polyn et al, 2005):

- create a null distribution of performance values by repeatedly generating scrambled versions of the classifier output
- MVPA: wavelet based scrambling technique (Bullmore et al., 2004)  
→ can accommodate non-independent data

## Bootstrapping

- estimate the variance and distribution of a statistic (i.e. classification accuracy) by data resampling with replacement

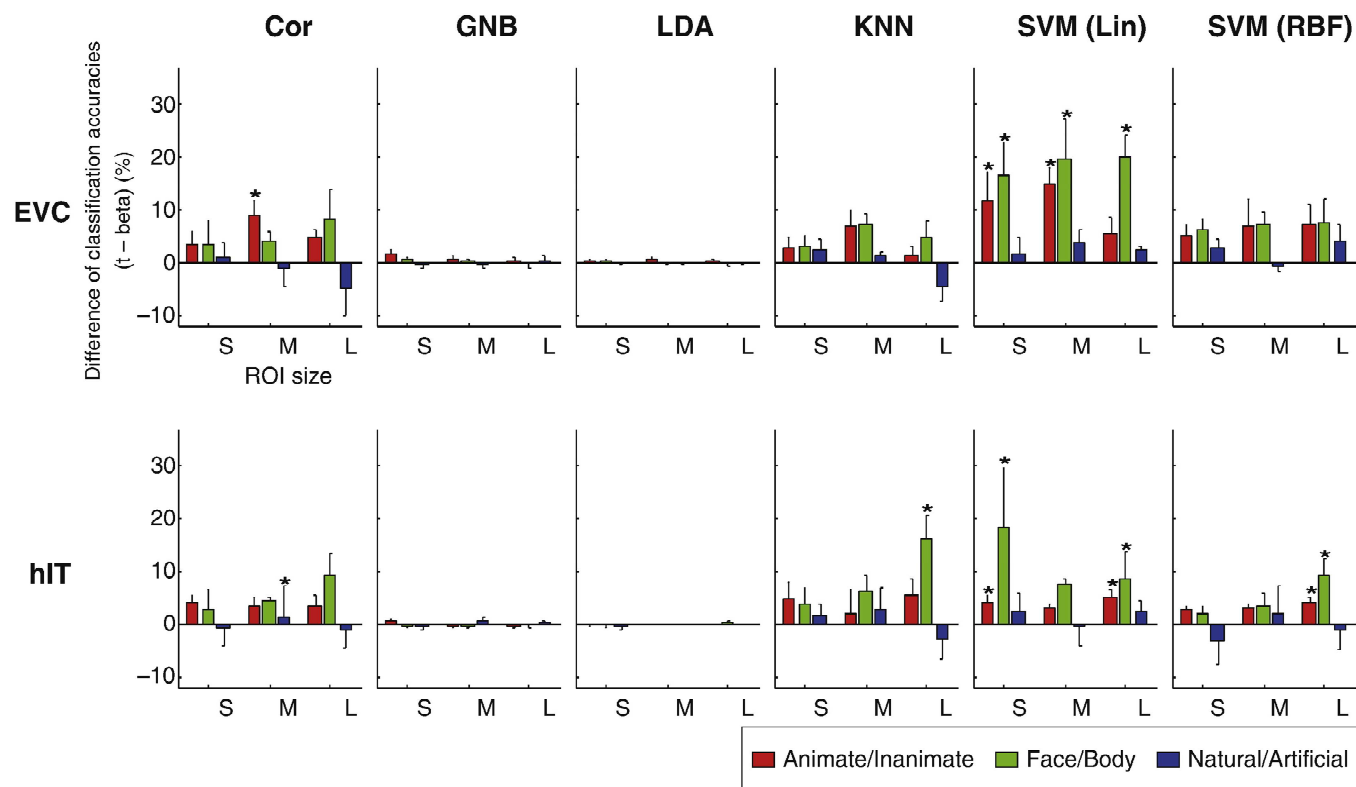
## Group inference

- (non-)parametric tests on classification accuracies, voxel weights etc.

## Design considerations

- acquire as many training examples as possible → classifier needs to be able to „see through the noise“
- averaging consecutive TR's can help to reduce the impact of noise (but may also eliminate natural, informative variation)
- avoid using consecutive scans for training a classifier → lots of highly similar datapoints do not give new information
- acquire as many test examples as possible → increases the power of significance test
- balance conditions → if not, classifier may tend to focus on predominant condition
- alternative to averaging: use beta weights or t-images from a GLM analysis (i.e. based on FIR or HRF)

## Classification on t- vs. beta images



Misaki et al. (2010)  
*NeuroImage*

- normalisation by standard error can down-weight noisy voxels
- Linear SVMs sensitive to scaling of the data → decision hyperplane is not simply scaled along with the data points but can reorient as it finds a new maximum-margin configuration
- LDA and GNB take each voxel's response variance into account automatically via the diagonal entries of the covariance matrix

## What you also might be interested in...

- how to use decoding to predict continuous variables (i.e. age, disease severity, ratings etc.) – relevance/support vector regression, relevance voxel machine (see Chu et al., 2011; Formisano et al., 2008; Wang et al., 2010)
- fMRI adaptation vs. pattern classification => two approaches assumed to provide subvoxel resolution (see Sapountzis et al., 2010, for a comparison in early visual areas)
- inferential limitations of decoding approach (vs. encoding models, see Naseri et al., 2011; Friston et al., 2008)

## Further reading

- Formisano E, De Martino F, Valente G (2008) Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magn Reson Imaging* 26(7):921-34.
- Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103:3863-3868.
- Misaki M. et al. (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* 53, 103-118.
- Mitchell TM, et al. (2004) Learning to Decode Cognitive States from Brain Images. *Machine Learning* 57:145-175.
- O'Toole et al. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cogn Neurosci*.19(11):1735-52
- Pereira F, Mitchell TM, Botvinick M (2009) Machine Learning Classifiers and fMRI: a tutorial overview. *Neuroimage* 45(1 Suppl):S199-209.