

# Multivariate analyses & decoding

Kay Henning Brodersen

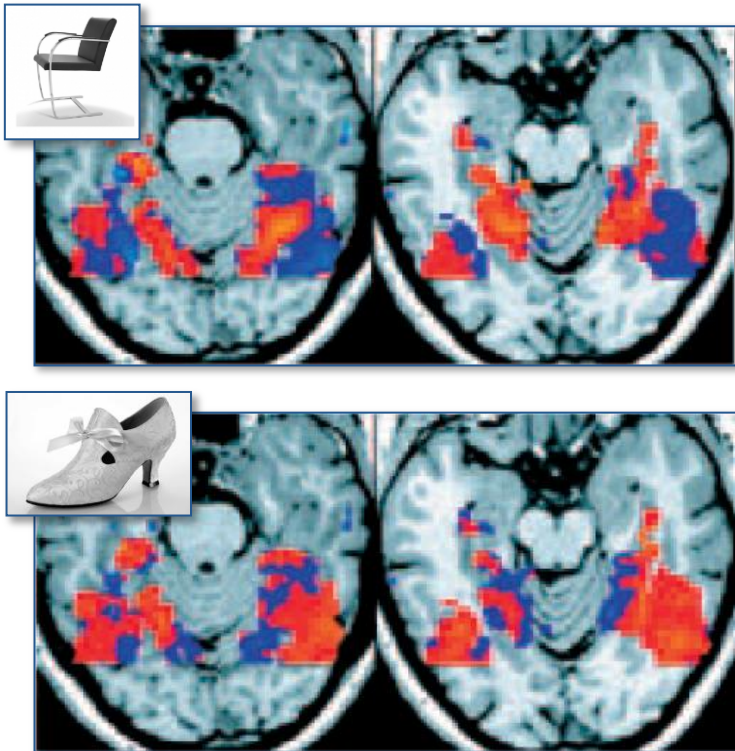
Computational Neuroeconomics Group  
Department of Economics, University of Zurich

Machine Learning and Pattern Recognition Group  
Department of Computer Science, ETH Zurich

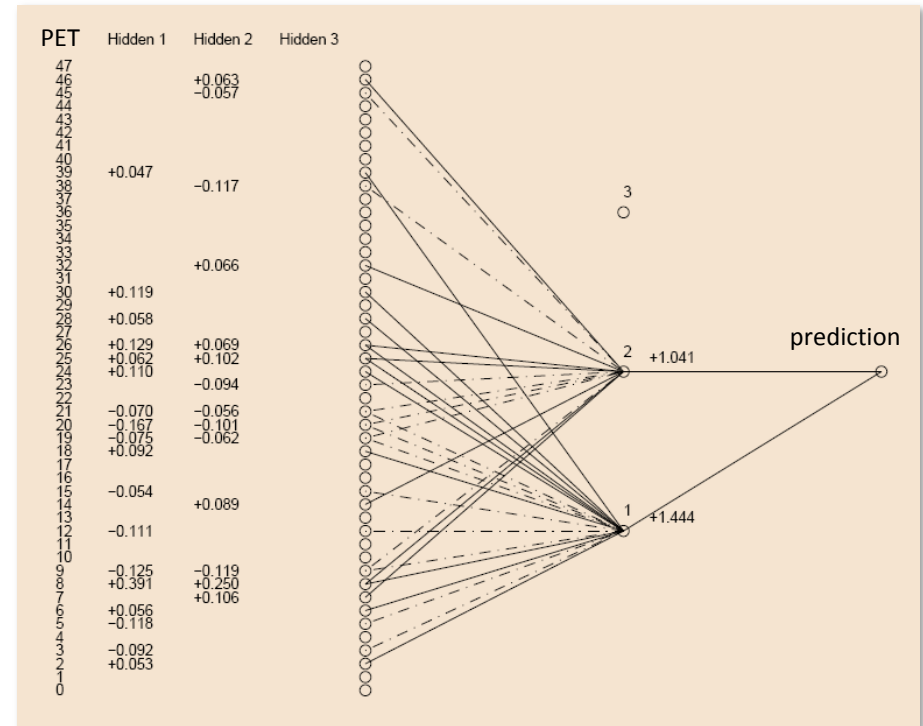
<http://people.inf.ethz.ch/bkay>

# Why multivariate?

Multivariate approaches simultaneously consider brain activity in many locations.



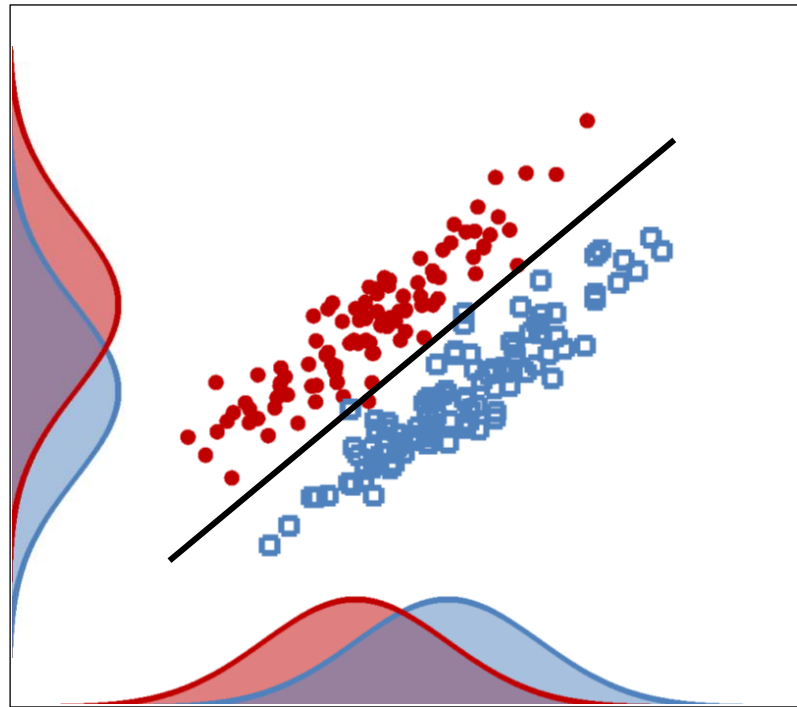
Haxby et al. (2001) *Science*



Lautrup et al. (1994) *Supercomputing in Brain Research*

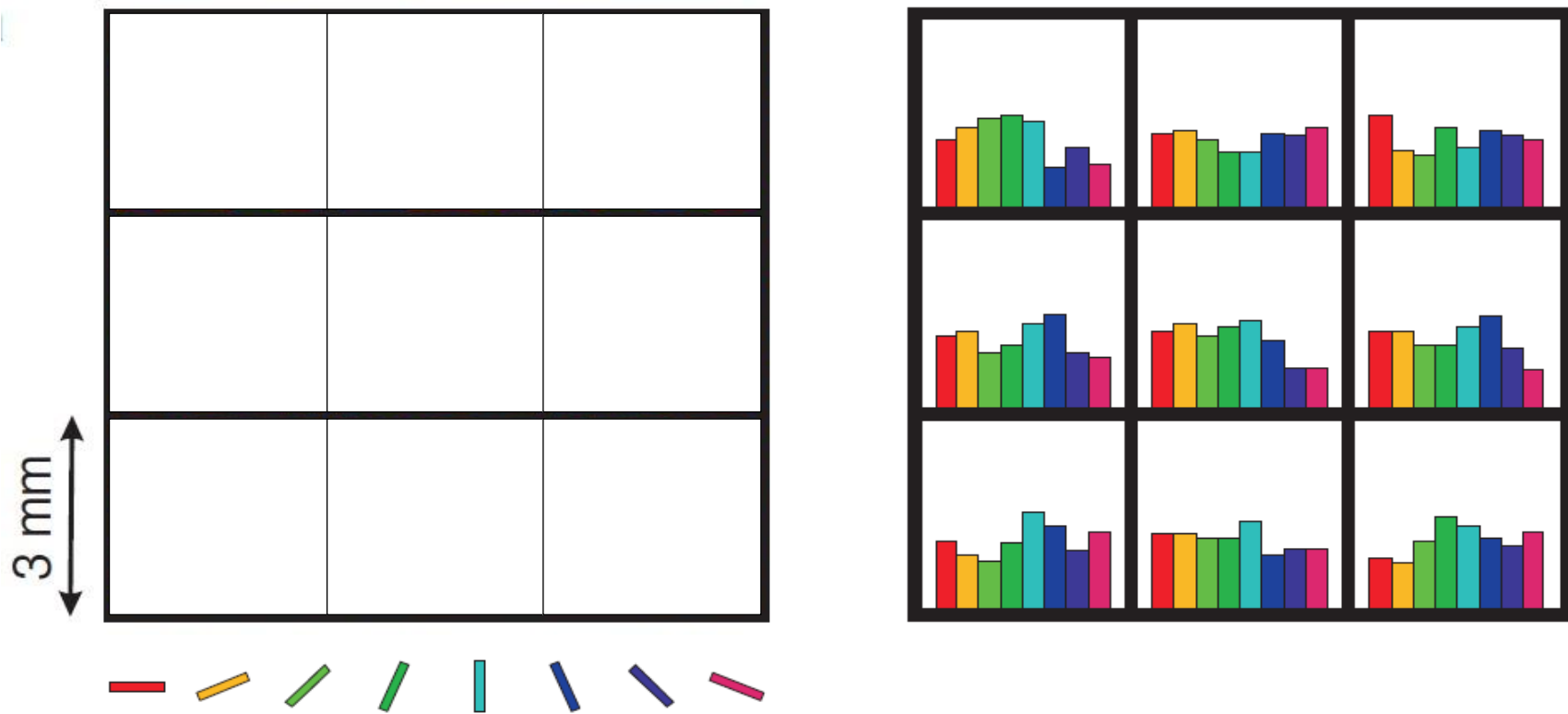
# Why multivariate?

Multivariate approaches can utilize information jointly encoded in multiple voxels. This is because multivariate distance measures can account for correlations between voxels.



# Why multivariate?

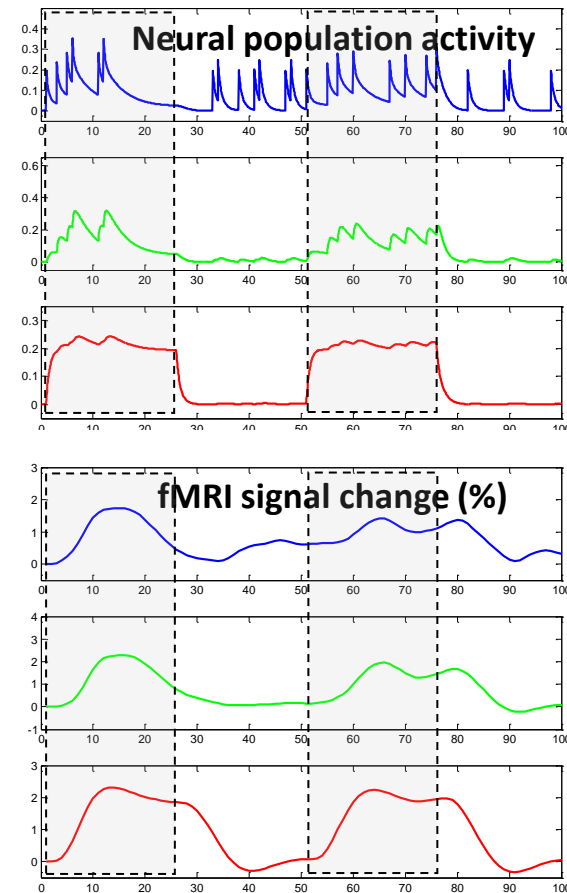
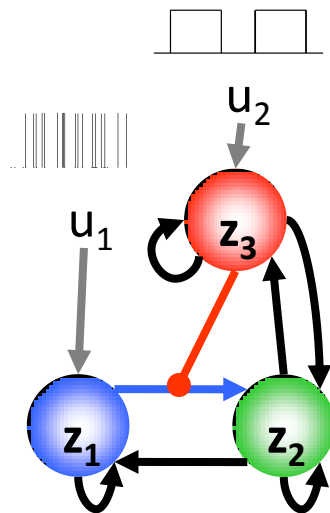
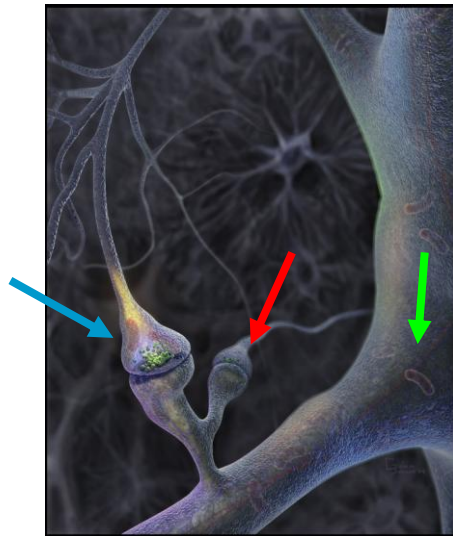
Multivariate approaches can exploit a sampling bias in voxelized images to reveal interesting activity on a subvoxel scale.



Boynton (2005) *Nature Neuroscience*

# Why multivariate?

Multivariate approaches can utilize 'hidden' quantities such as coupling strengths.



neural state equation:

$$\dot{z} = \left( A + \sum_{i=1}^m u_i B^{(i)} + \sum_{j=1}^n z_j D^{(j)} \right) z + Cu$$

Stephan et al. (2008) *NeuroImage*

# Outline

---

- 1 Foundations
- 2 Classification
- 3 Multivariate Bayes
- 4 Further model-based approaches

# Outline

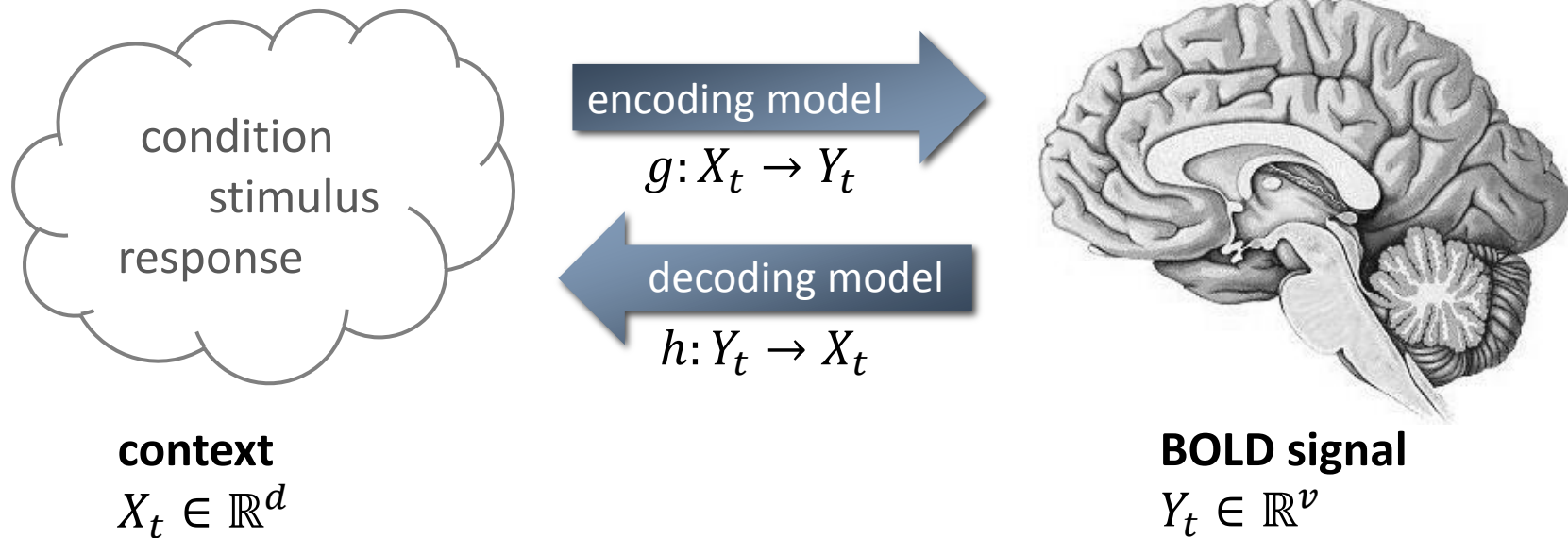
---

- 1 Foundations
- 2 Classification
- 3 Multivariate Bayes
- 4 Further model-based approaches

# Modelling terminology

## 1 Encoding vs. decoding

An **encoding** model (or generative model) relates context to brain activity.  
A **decoding** model (or recognition model) relates brain activity to context.





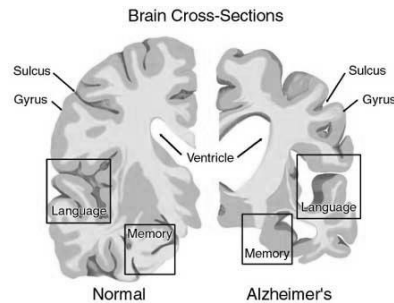
# Modelling terminology

## 2 Prediction vs. inference

The goal of **prediction** is to find a highly accurate encoding or decoding function.

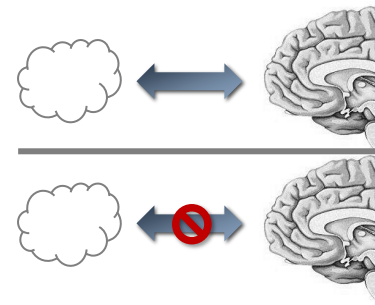


predicting a cognitive state using a brain-machine interface

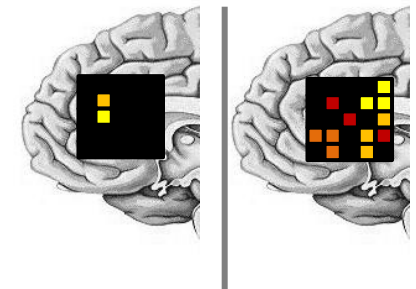


predicting a subject-specific diagnostic status

The goal of **inference** is to decide between competing hypotheses about structure-function mappings in the brain.



comparing a model that links distributed neuronal activity to a cognitive state with a model that does not



weighing the evidence for sparse coding vs. dense coding

predictive density

$$p(X_{new}|Y_{new}, X, Y) = \int p(X_{new}|Y_{new}, \theta)p(\theta|X, Y)d\theta$$

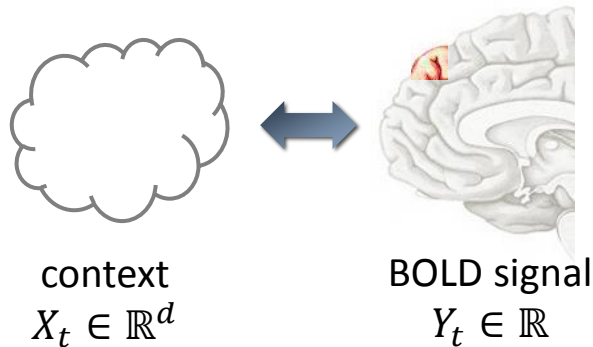
marginal likelihood

$$p(X|Y) = \int p(X|Y, \theta)p(\theta)d\theta$$

# Modelling terminology

## 3 Univariate vs. multivariate

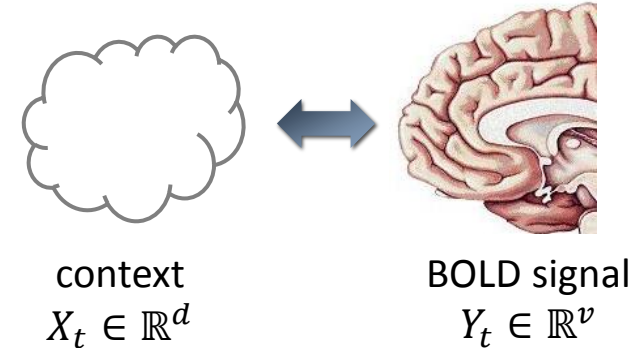
**A univariate model** considers a single voxel at a time.



The implicit likelihood of the data factorizes over voxels,  $p(Y_t|X_t) = \prod_{i=1}^v p(Y_{t,i}|X_t)$ .

Spatial dependencies between voxels are introduced afterwards, through random field theory. This enables multivariate inferences over voxels (i.e., cluster-level or set-level inference).

**A multivariate model** considers many voxels at once.



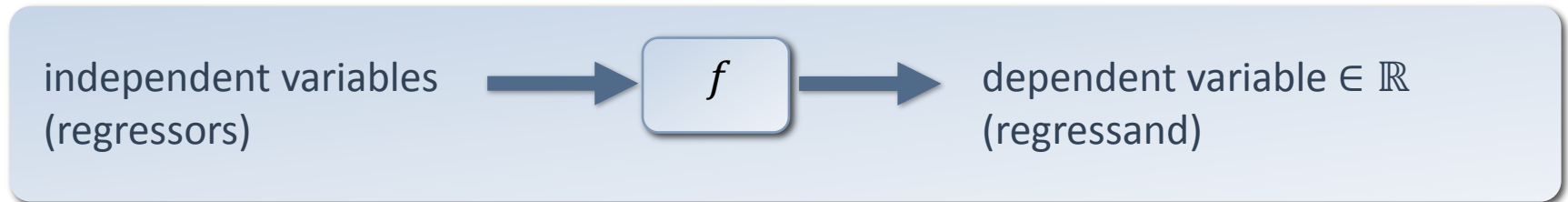
Multivariate models relax the assumption about independence of voxels.

They enable inference about distributed responses without requiring focal activations or certain topological response features. They can therefore be more powerful than univariate analyses.

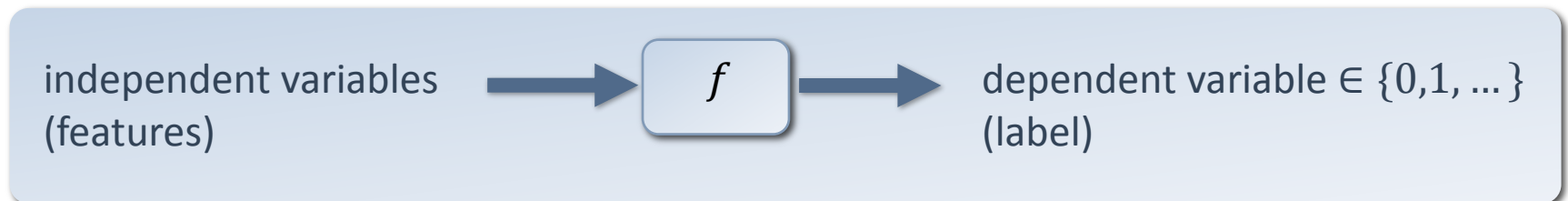
# Modelling terminology

## 4 Regression vs. classification

In a **regression** model, the dependent variable is continuous.



In a **classification** model, the dependent variable is categorical (e.g., binary).

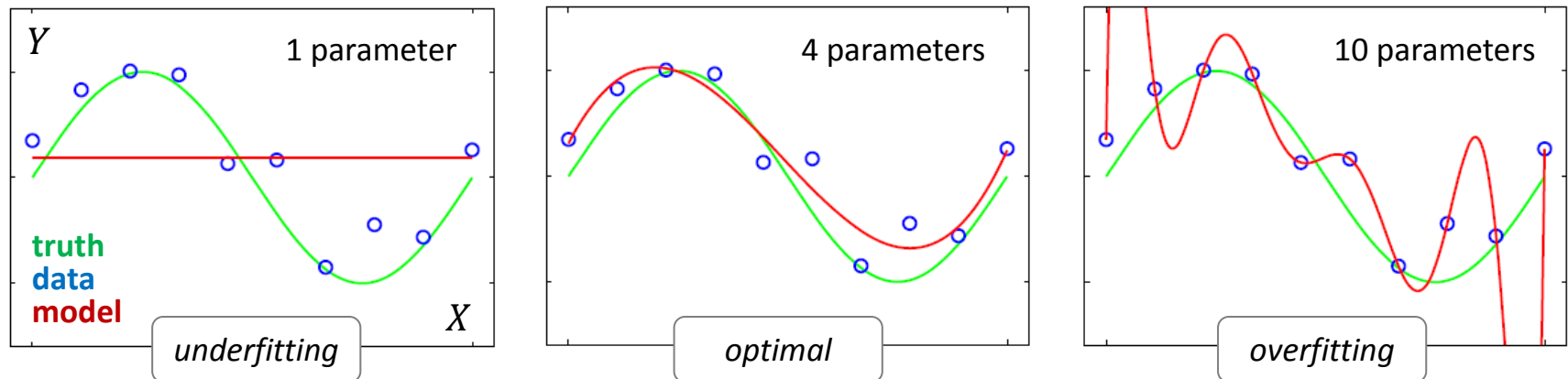


# Modelling terminology

## 5 Goodness of fit vs. complexity

A measure of **goodness of fit** assesses how well the model fits available data.

A measure of **complexity** assesses the amount of structure and the degree of flexibility of a model, including, among other things, its number of parameters.



We typically wish to find the model that **generalizes best to new data**. This is the model that optimally trades off goodness of fit and complexity.

# Summary of modelling terminology

## General Linear Model (GLM)

mass-univariate encoding model for regressing context onto brain activity and inferring on topological response features

## Dynamic Causal Modelling (DCM)

multivariate encoding model for comparing alternative connectivity hypotheses

## Classification

based on multivariate decoding models for predicting a categorical context label from brain activity

## Multivariate Bayes (MVB)

multivariate encoding model for comparing alternative coding hypotheses

• • •

• • •

# Outline

---

- 1 Foundations
- 2 Classification**
- 3 Multivariate Bayes
- 4 Further model-based approaches

# Constructing a classifier

In classification, we aim to predict a target variable  $X$  from data  $Y$ ,

$$h: Y_t \rightarrow X_t \in \{1, \dots, K\}$$

Most classifiers are designed to estimate the unknown probabilities of an example belonging to a particular class:

$$h(Y_t) = \arg \max_k p(X_t = k | Y_t, X, Y)$$

## Generative classifiers

use Bayes' rule to estimate  
 $p(X_t | Y_t) \propto p(Y_t | X_t) p(X_t)$

*e.g., Gaussian Naïve Bayes  
Linear Discriminant Analysis*

## Discriminative classifiers

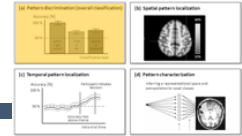
estimate  $p(X_t | Y_t)$  directly  
without Bayes' theorem

*e.g., Logistic regression  
Relevance Vector Machine*

## Discriminant classifiers

estimate  $h(Y_t)$  directly

*e.g., Support Vector Machine  
Fisher's Linear Discriminant*



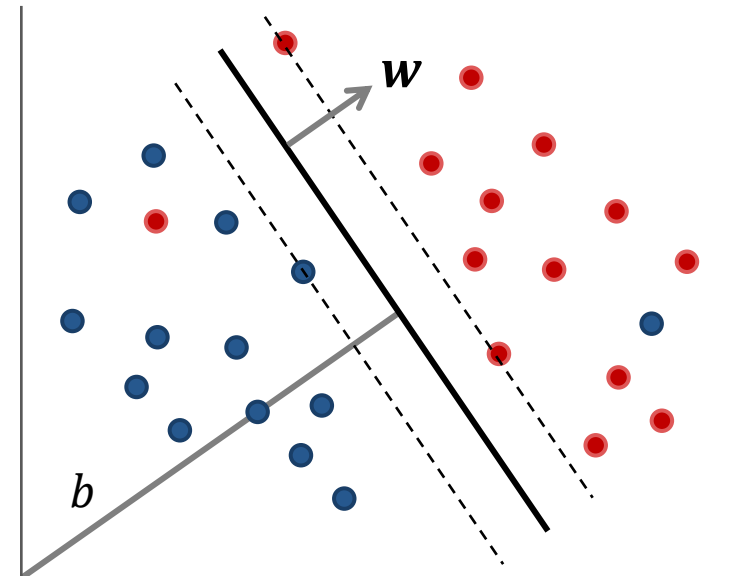
**The support vector machine (SVM) is a discriminant classifier.**

- Training – Find a hyperplane with a maximal margin to the nearest examples on either side.
- Test – Assign a new example to the class corresponding to its side of the plane.

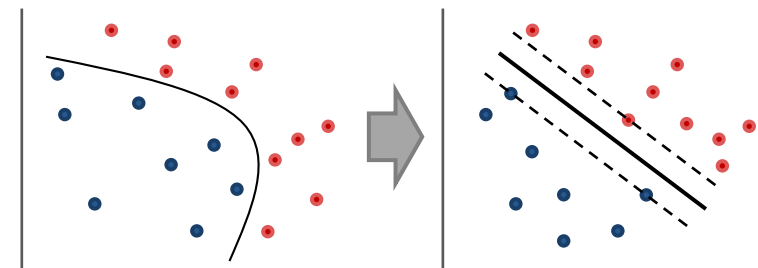
**SVMs are used in many domains of application.**

- Efficiency – SVMs are fast and easy to use.
- Performance – SVMs usually perform well compared to other classifiers.
- Flexibility – The need for vectorial representations of examples is replaced by a similarity measure, defined via a kernel function  $k(Y_i, Y_j)$ .

**Linear SVM**



**Nonlinear SVM**





# Generalizability of a classifier

Typically, we have many more voxels than observations. This means that there are infinitely many models that enable perfect classification of the available data. But these models might have overfit the data.

Overfitting is usually not an issue in GLM analyses, where the number of regressors is much smaller than the number of observations.

We want to find a classification model  $h: Y \rightarrow X$  that generalizes well to new data. Given some training data, we might consider the probability

$$P\left(h(Y^{(\text{test})}) = X^{(\text{test})} \mid Y^{(\text{train})}, X^{(\text{train})}\right).$$

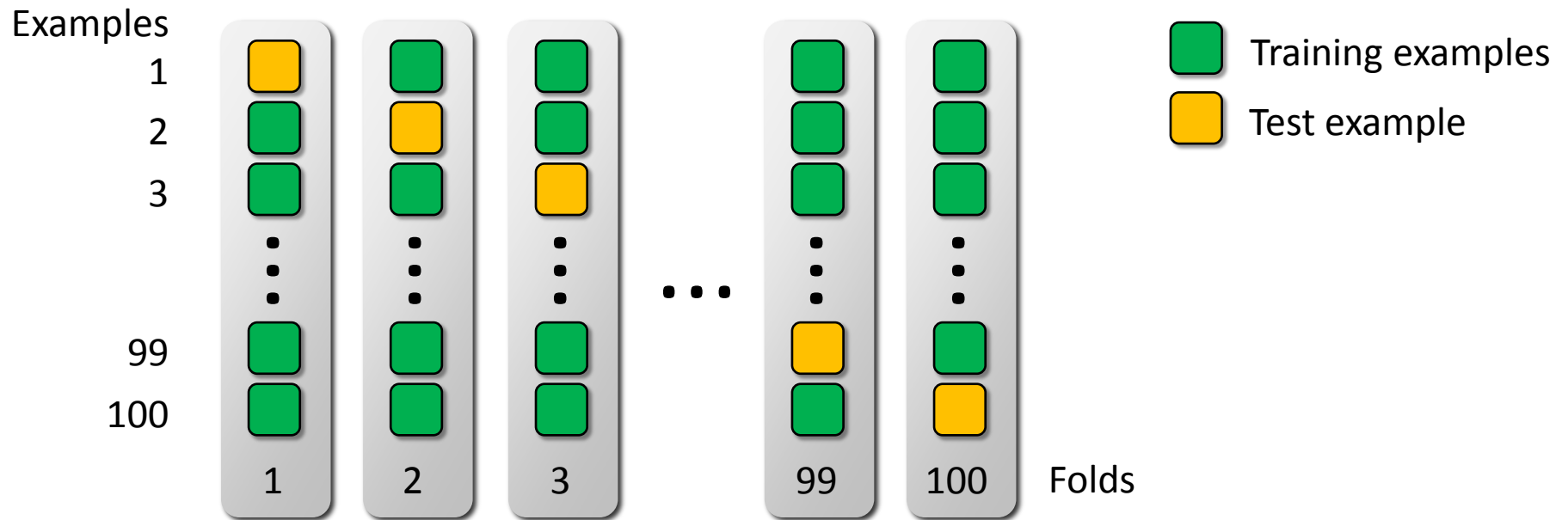
However, this quantity is dependent on the training data. So instead we should consider the generalizability

$$E_{\text{training}} \left[ P\left(h(Y^{(\text{test})}) = X^{(\text{test})} \mid Y^{(\text{train})}, X^{(\text{train})}\right) \right],$$

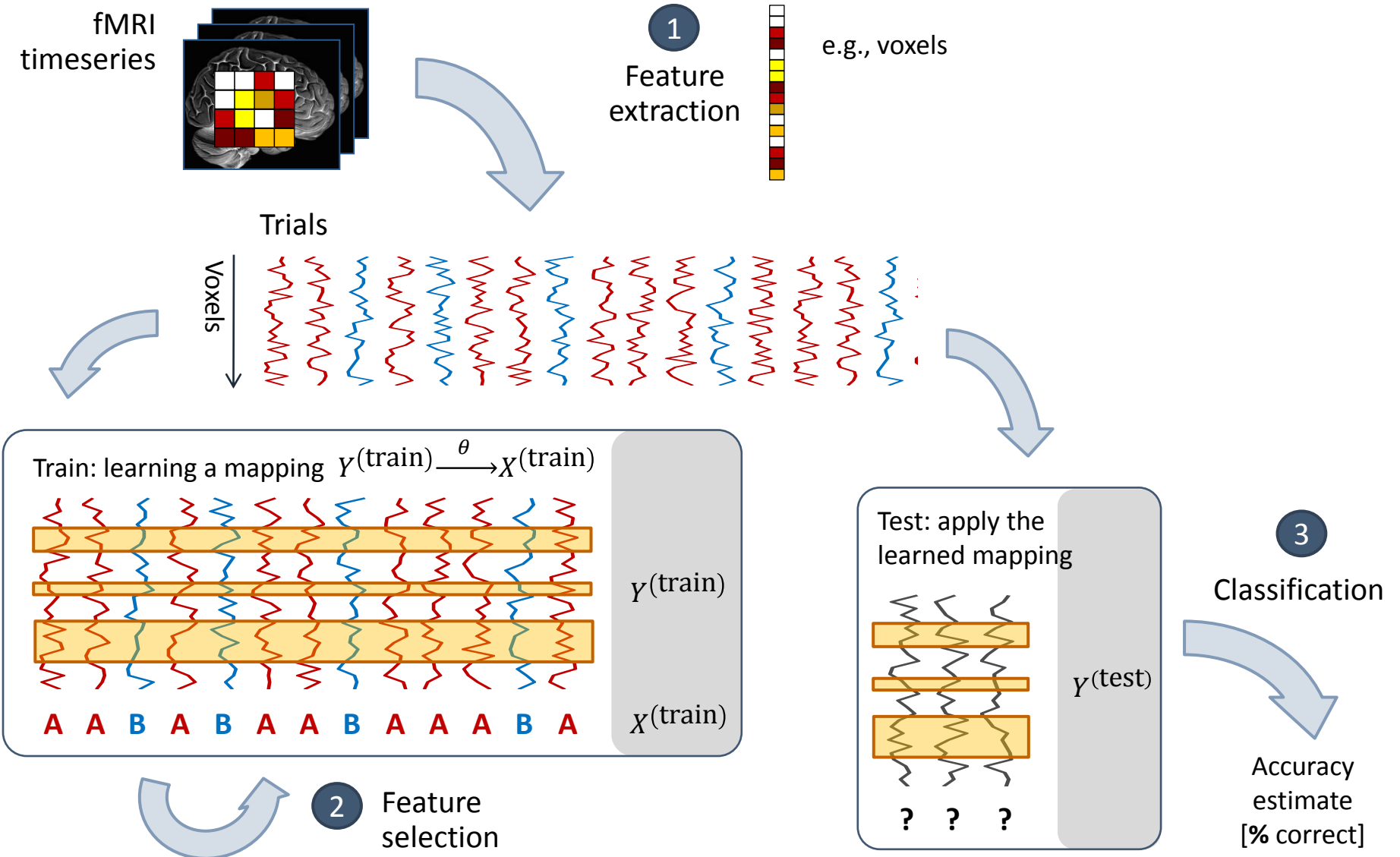
which we can approximate using cross-validation.

# Cross-validation

Cross-validation is a resampling procedure that can be used to estimate the generalizability of a classifier.

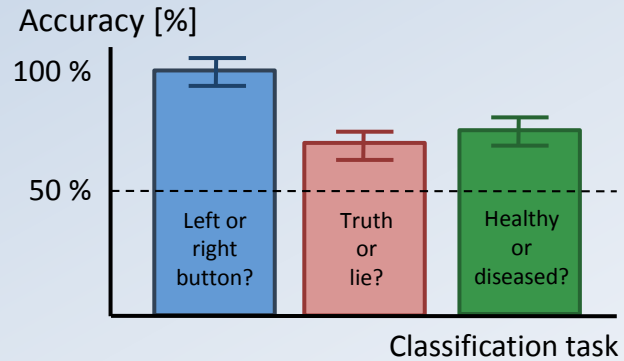


# Trial-by-trial classification of fMRI data

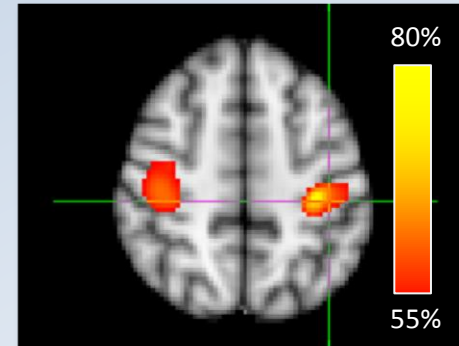


# Target questions in classification studies

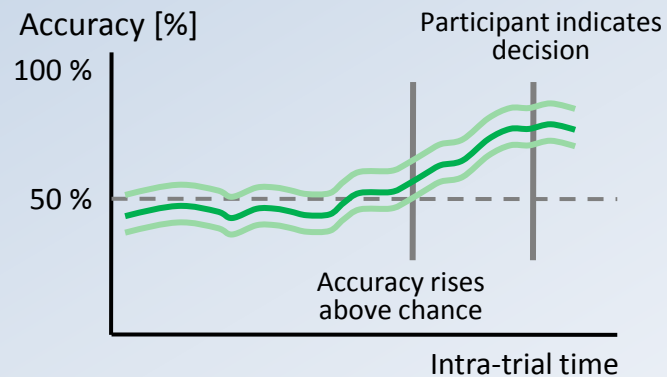
## A Overall classification accuracy



## B Spatial deployment of informative regions

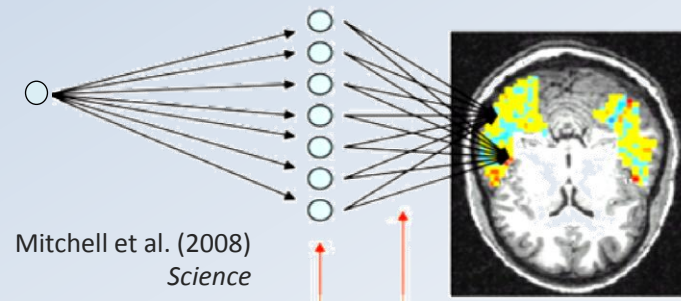


## C Temporal evolution of informativeness

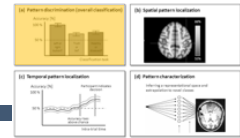


## D Characterization of distributed activity

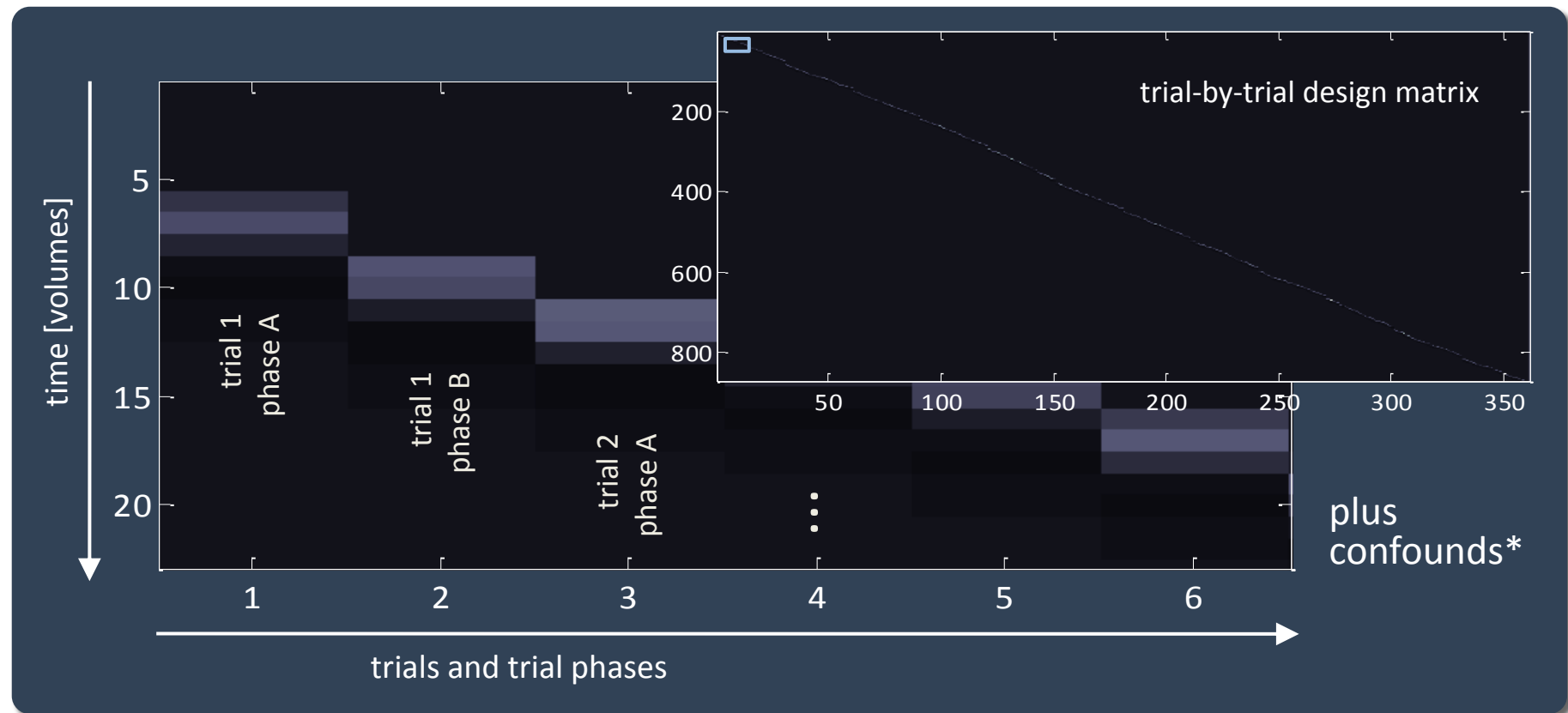
Inferring a representational space and extrapolation to novel classes



# Preprocessing for classification

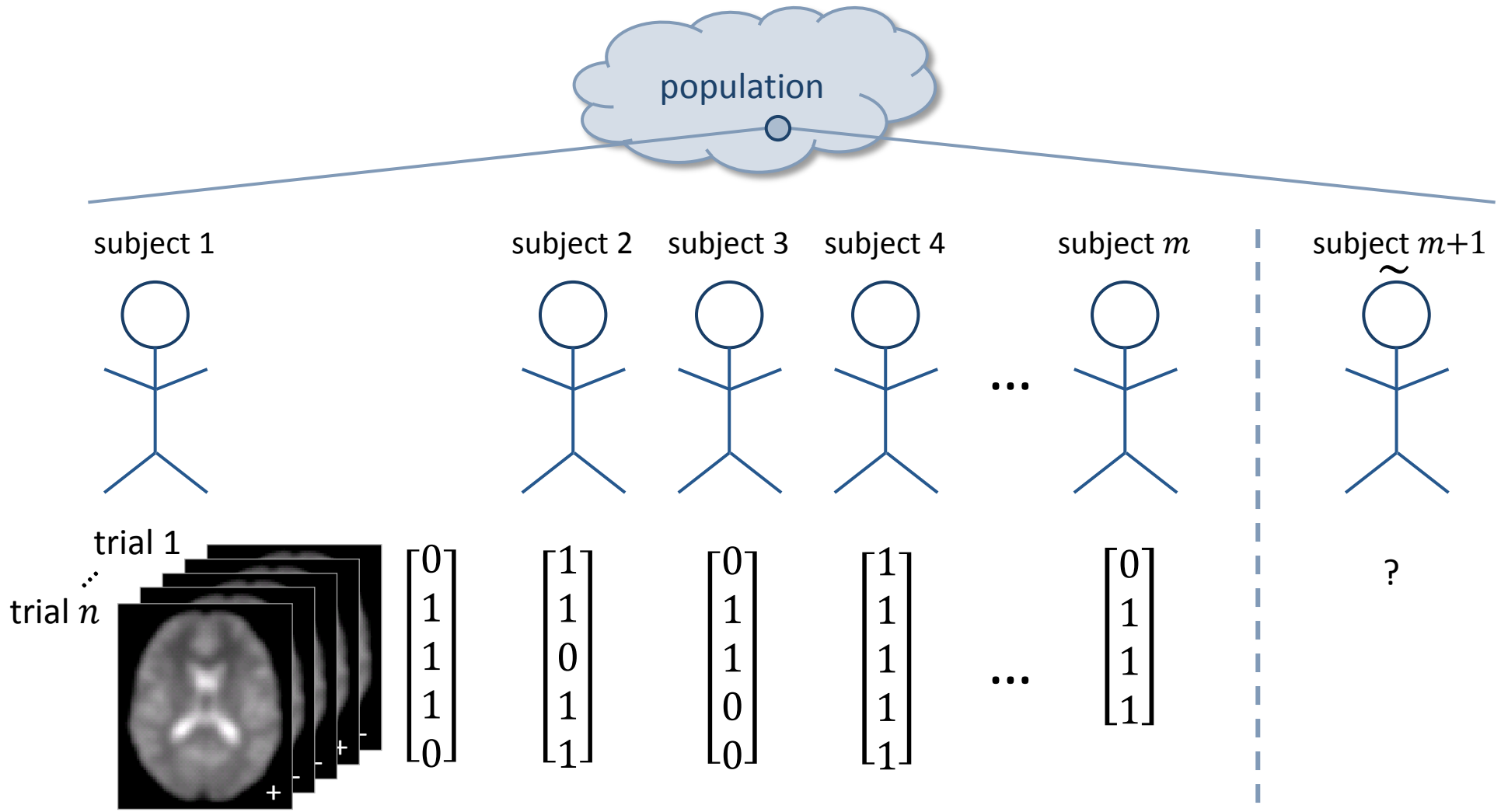
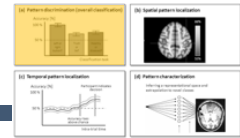


The most principled approach is to deconvolve the BOLD signal using a GLM.

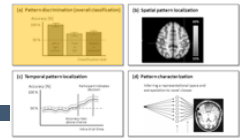


This approach results in one beta image per trial and phase.

# Performance evaluation



Brodersen et al. (in preparation)



## Single-subject study

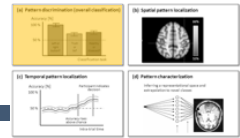
The most common approach is to assess how likely the obtained number of correctly classified trials could have occurred by chance.

$$p = P(X \geq k | H_0) = 1 - B(k | n, \pi_0)$$

$p$	probability of observing the obtained performance by chance
$k$	number of correctly classified trials
$n$	total number of trials
$\pi_0$	probability of getting a single result right by chance
$B$	binomial cumulative density function

In publications, this approach is referred to as a *binomial* test.

It is based on the assumption that, under the Null hypothesis, the classifier produces random predictions.



## Group study

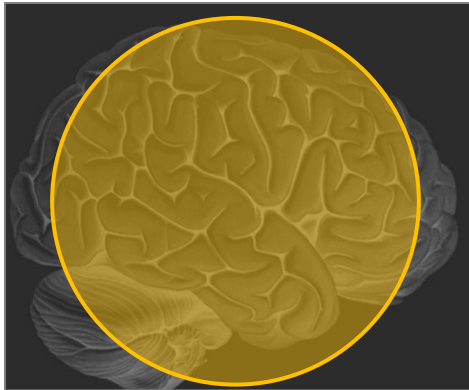
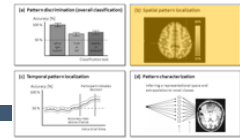
The most common approach is to assess the probability with which the observed subject-wise sample accuracies were sampled from a distribution with a mean equal to chance.

$$t = \sqrt{m} \frac{\bar{\pi} - \pi_0}{\hat{\sigma}_{m-1}}$$
$$p = 1 - t_{m-1}(t)$$

$p$	probability of observing the obtained performance by chance
$m$	number of subjects
$\bar{\pi}$	sample mean of subject-wise sample accuracies
$\hat{\sigma}_{m-1}$	sample standard deviation of subject-wise sample accuracies
$\pi_0$	probability of getting a single result right by chance
$t_{m-1}$	cumulative Student's $t$ -distribution with $m - 1$ d.o.f.

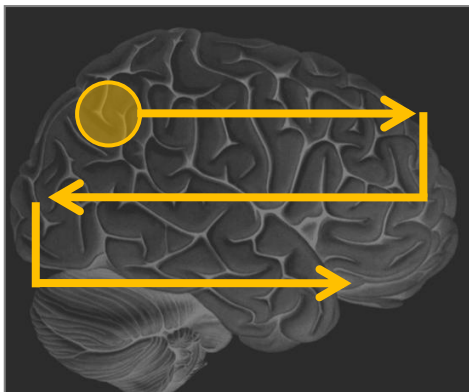
This approach represents a random-effects analysis of classification outcomes based on the additional assumption that the mean of sample accuracies is approximately Normal.





## Approach 1 – Consider the **entire brain**, and find out which voxels are jointly discriminative.

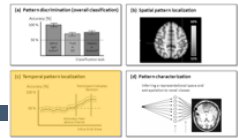
- ▣ based on a classifier with a constraint on sparseness in features  
Hampton & O’Doherty (2007); Grosenick et al. (2008, 2009)
- ▣ based on Gaussian Processes  
Marquand et al. (2010) NeuroImage; Lomakina et al. (*in preparation*)



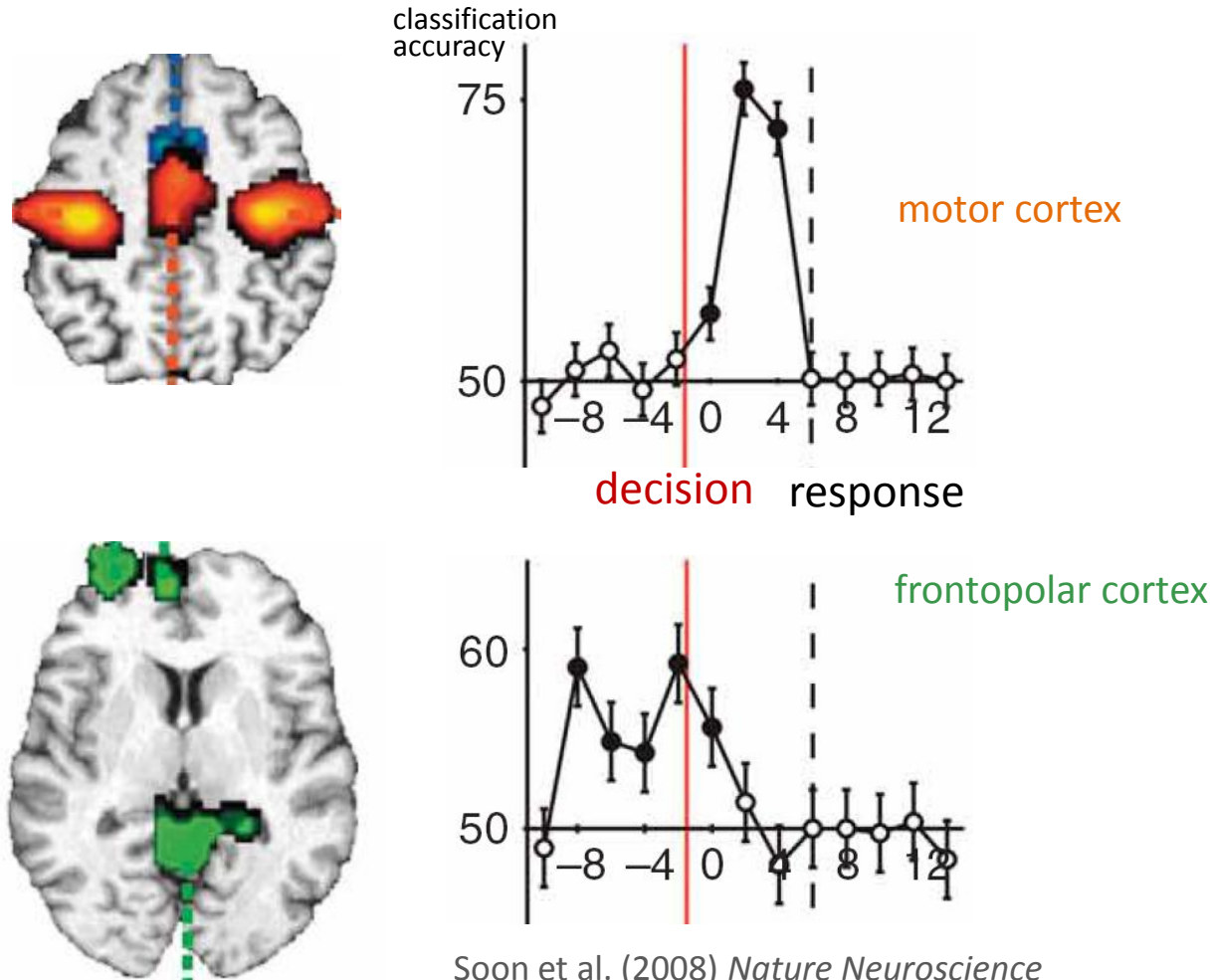
## Approach 2 – At each voxel, consider a **small local environment**, and compute a discriminability score.

- ▣ based on a CCA  
Nandy & Cordes (2003) *Magn. Reson. Med.*
- ▣ based on a classifier
- ▣ based on Euclidean distances
- ▣ based on Mahalanobis distances  
Kriegeskorte et al. (2006, 2007a, 2007b)  
Serences & Boynton (2007) *J Neuroscience*
- ▣ based on the mutual information

# Temporal evolution of discriminability

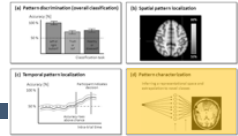


## Example – decoding which button the subject pressed

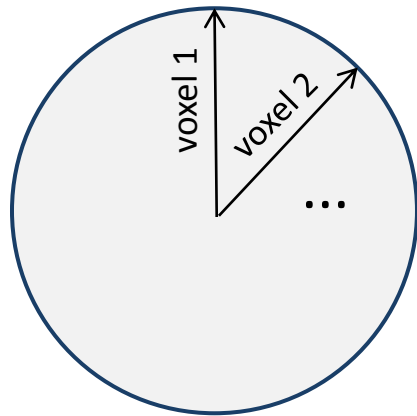


Soon et al. (2008) *Nature Neuroscience*

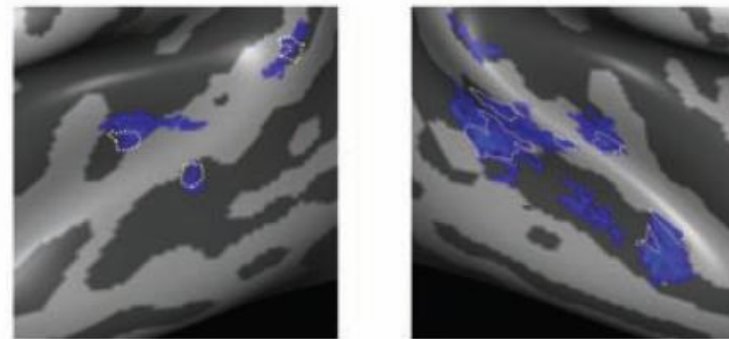
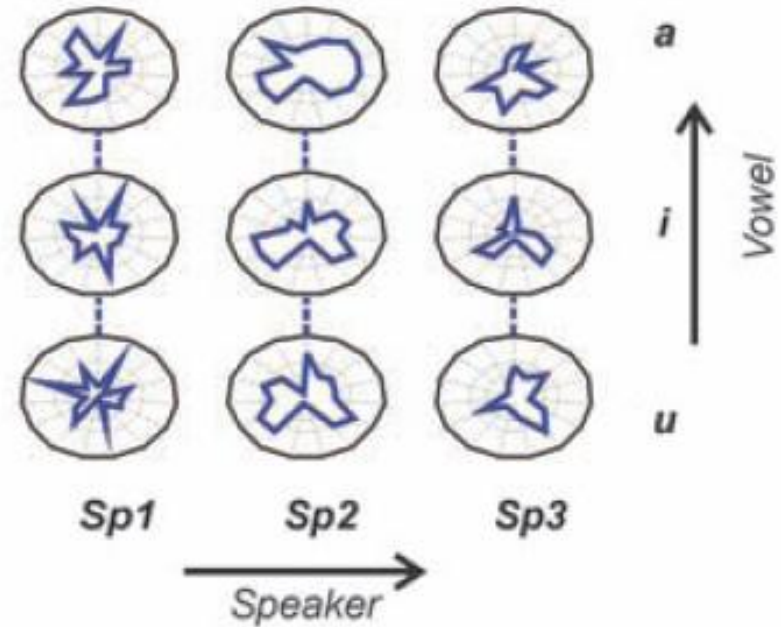
# Pattern characterization



**Example** – decoding the identity of the person speaking to the subject in the scanner



fingerprint plot  
(one plot per class)



Formisano et al. (2008) *Science*

# Issues to be aware of (as researcher or reviewer)

---

- ❑ Classification induces constraints on the experimental design.
  - When estimating trial-wise Beta values, we need longer ITIs (typically 8 – 15 s).
  - At the same time, we need many trials (typically 100+).
  - Classes should be balanced. If they are imbalanced, we can resample the training set, constrain the classifier, or report the balanced accuracy.
- ❑ Construction of examples
  - Estimation of Beta images is the preferred approach.
  - Covariates should be included in the trial-by-trial design matrix.
- ❑ Temporal autocorrelation
  - In trial-by-trial classification, exclude trials around the test trial from the training set.
- ❑ Avoiding double-dipping
  - Any feature selection and tuning of classifier settings should be carried out on the training set only.
- ❑ Performance evaluation
  - Do random-effects or mixed-effects inference.
  - Correct for multiple tests.

# Outline

---

- 1 Foundations
- 2 Classification
- 3 Multivariate Bayes**
- 4 Further model-based approaches

# Multivariate Bayes

Multivariate analyses in SPM are not framed in terms of classification problems. Instead, SPM brings multivariate analyses into the conventional inference framework of hierarchical models and their inversion.

Multivariate Bayes (MVB) can be used to address two questions:

## **Is there a link between $X$ and $Y$ ?**

- using cross-validation (as seen earlier)
- using model comparison (new)

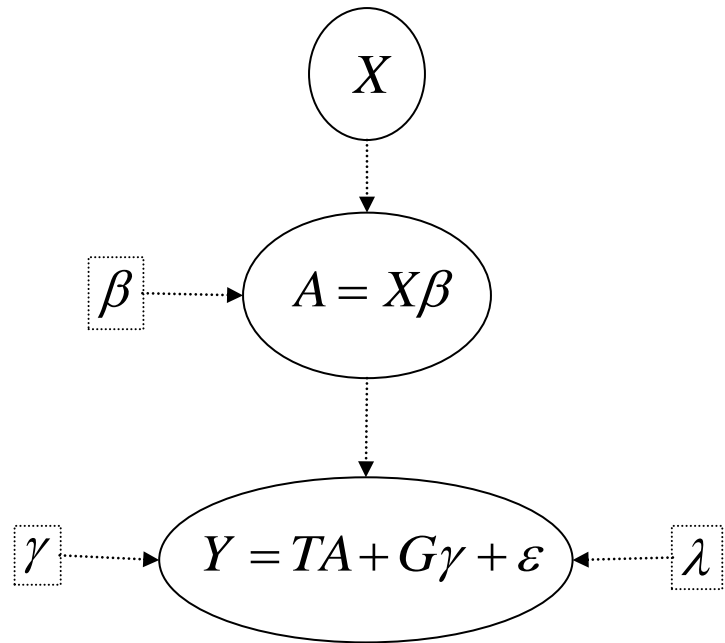
## **What is the form of the link between $X$ and $Y$ ?**

- smooth or sparse coding?  
(many voxels vs. few voxels)
- category-specific representations that are functionally selective or functionally segregated?

# Conventional inference framework

## Classical encoding model

$X$  as a cause of  $Y$

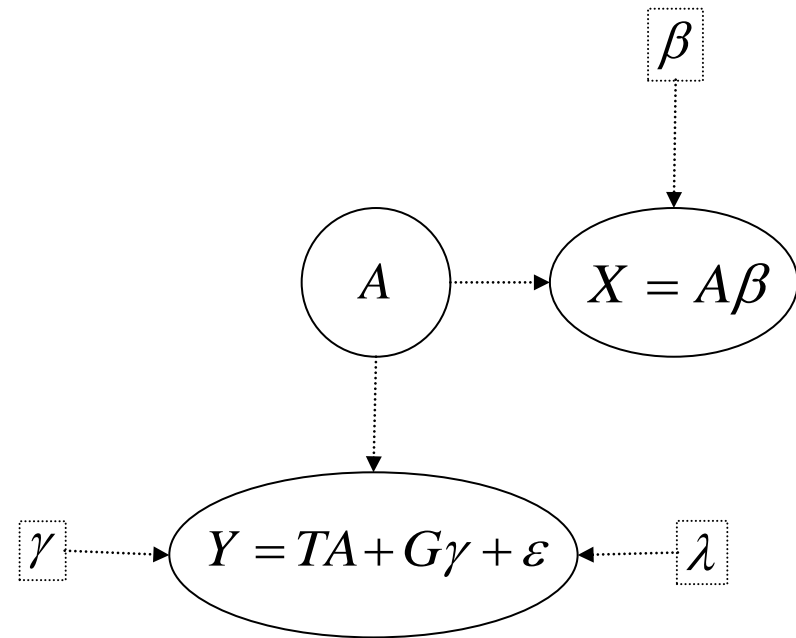


$$g(\theta) : X \rightarrow Y$$

$$Y = TX\beta + G\gamma + \varepsilon$$

## Bayesian decoding model

$X$  as a consequence of  $Y$



$$h(\theta) : Y \rightarrow X$$

Friston et al. (2008) *NeuroImage*

# Lessons from the Neyman-Pearson lemma

## Is there a link between $X$ and $Y$ ?

To test for a statistical dependency between a contextual variable  $X$  and the BOLD signal  $Y$ , we compare

- ▣  $H_0$ : there is no dependency
- ▣  $H_a$ : there is some dependency

## Which statistical test?

1. define a test size  $\alpha$   
(the probability of falsely rejecting  $H_0$ , i.e.,  $1 - \text{specificity}$ ),
2. choose the test with the highest power  $1 - \beta$   
(the probability of correctly rejecting  $H_0$ , i.e., sensitivity).

## The Neyman-Pearson lemma

The most powerful test of size  $\alpha$  is: to reject  $H_0$  when the likelihood ratio  $\Lambda$  exceeds a critical value  $u$ ,

$$\Lambda(Y) = \frac{p(Y|X)}{p(Y)} = \frac{p(X|Y)}{p(X)} \geq u$$

with  $u$  chosen such that

$$P(\Lambda(Y) \geq u | H_0) = \alpha.$$

The null distribution of the likelihood ratio  $p(\Lambda(Y) | H_0)$  can be determined non-parametrically or under parametric assumptions.

This lemma underlies both classical statistics and Bayesian statistics (where  $\Lambda(Y)$  is known as a Bayes factor).

Neyman & Person (1933) *Phil Trans Roy Soc London*



# Lessons from the Neyman-Pearson lemma

---

## In summary

1. Inference about how the brain represents context variables reduces to model comparison.
2. To establish that a link exists between some context  $X$  and activity  $Y$ , the direction of the mapping is not important.
3. Testing the accuracy of a classifier is not based on  $\Lambda$  is therefore suboptimal.

Neyman & Person (1933) *Phil Trans Roy Soc London*

Kass & Raftery (1995) *J Am Stat Assoc*

Friston et al. (2009) *NeuroImage*

# Priors help to regularize the inference problem

Mapping brain activity onto a context variable is ill-posed: there is an infinite number of equally likely solutions. We therefore require constraints (priors) to estimate the voxel weights  $\beta$ .

SPM comes with several alternative coding hypotheses, specified in terms of spatial priors on voxel weights,  $p(\tilde{\beta})$ , after transformations  $\tilde{Y} = YU$  and  $\tilde{\beta} = \beta U$ .

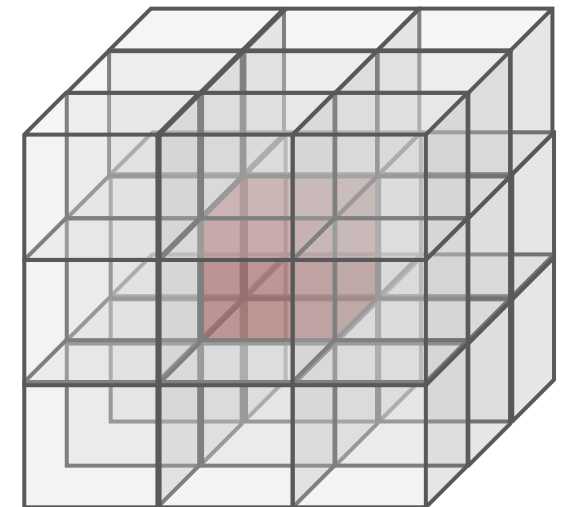
Null:  $U = \emptyset$

Spatial vectors:  $U = I$

Smooth vectors:  $U(\vec{x}_i, \vec{x}_j) = \exp(-\frac{1}{2}(\vec{x}_i - \vec{x}_j)^2 \sigma^{-2})$

Singular vectors:  $UDV^T = RY^T$

Support vectors:  $U = RY^T$



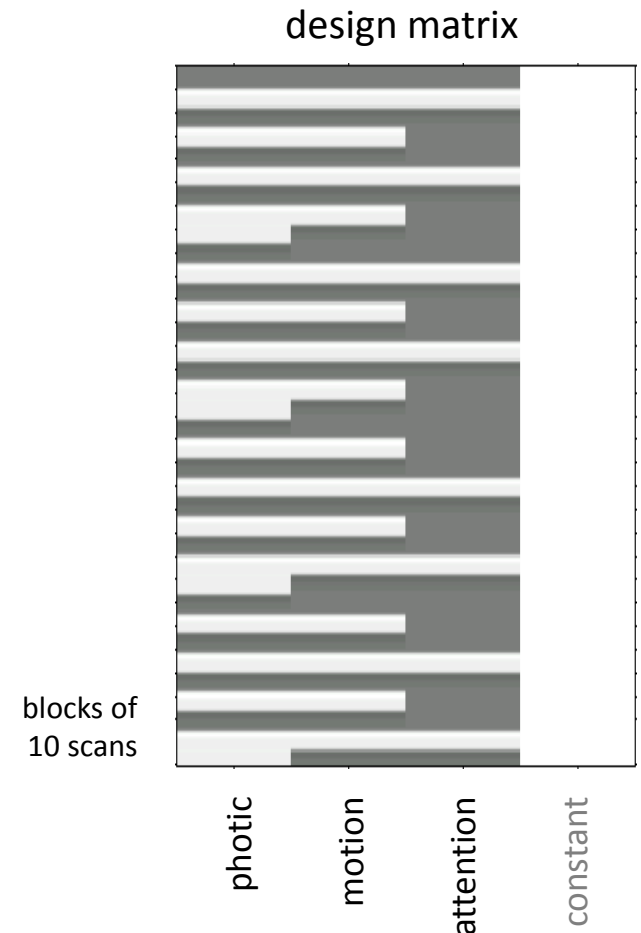
# Multivariate Bayes: example

- MVB can be illustrated using SPM's attention-to-motion example dataset.

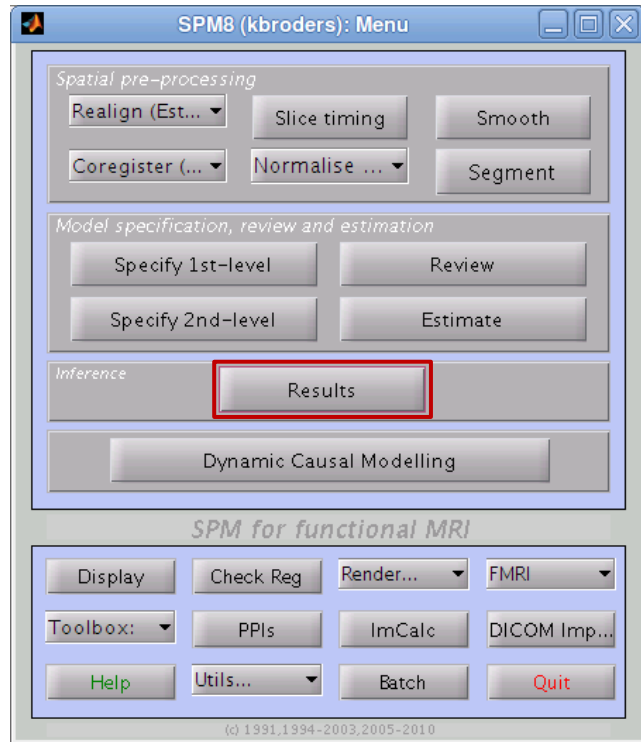
Buechel & Friston 1999 *Cerebral Cortex*

Friston et al. 2008 *NeuroImage*

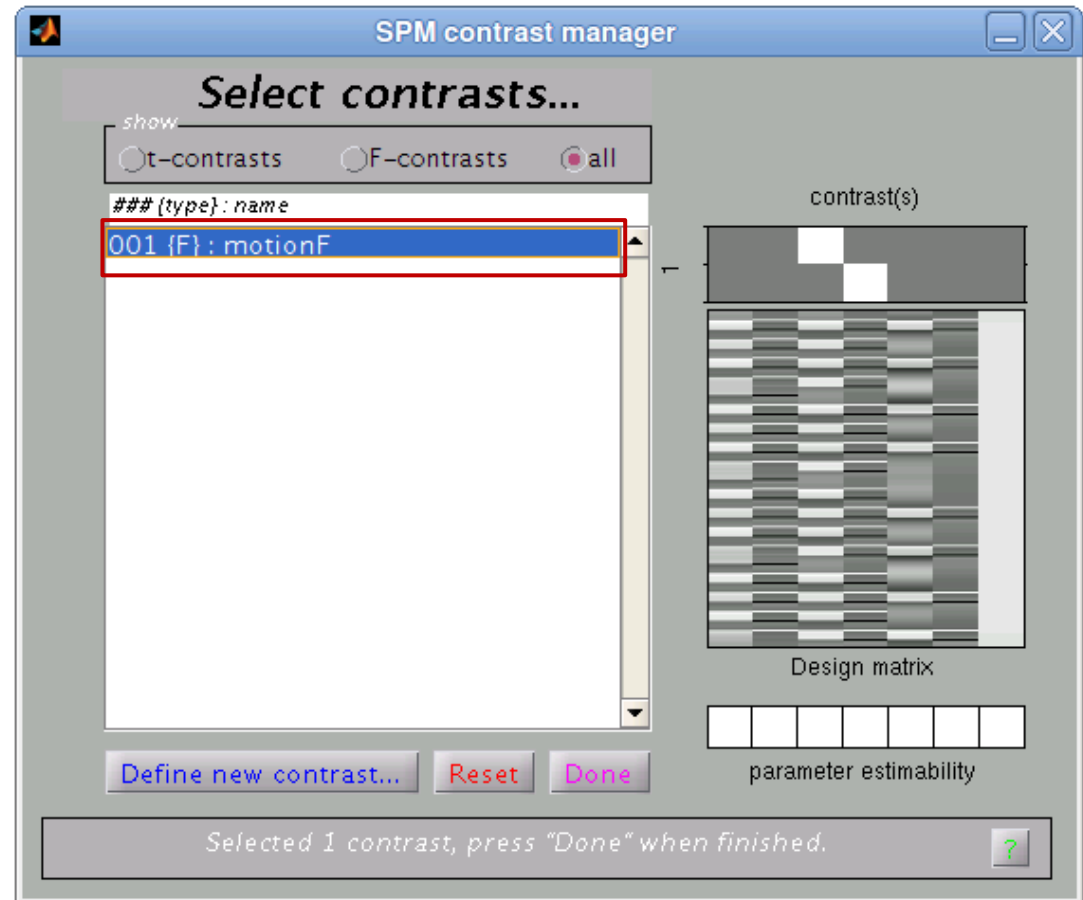
- This dataset is based on a simple block design. Each block is a combination of some of the following three factors:
  - photic – there is some visual stimulus
  - motion – there is motion
  - attention – subjects are paying attention
- We form a design matrix by convolving box-car functions with a canonical haemodynamic response function.



# Multivariate Bayes: example

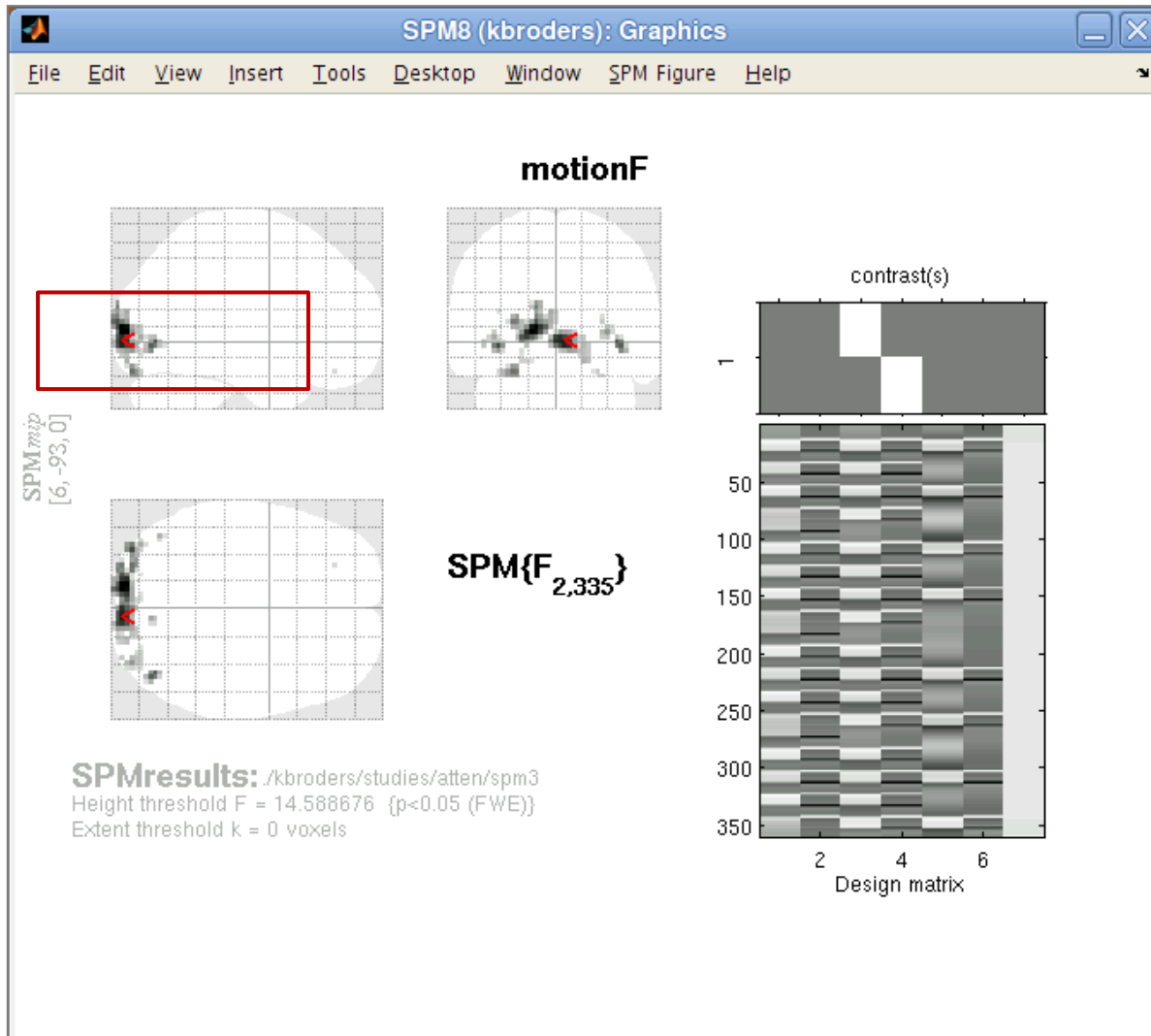


After having specified and estimated a design, we use the *Results* button.



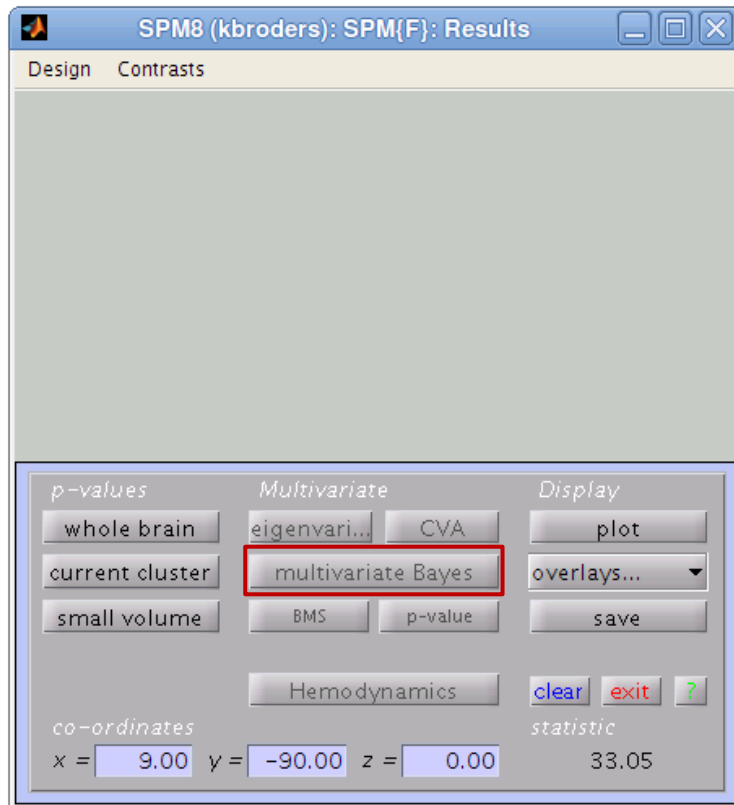
Next, we select the contrast of interest.

# Multivariate Bayes: example

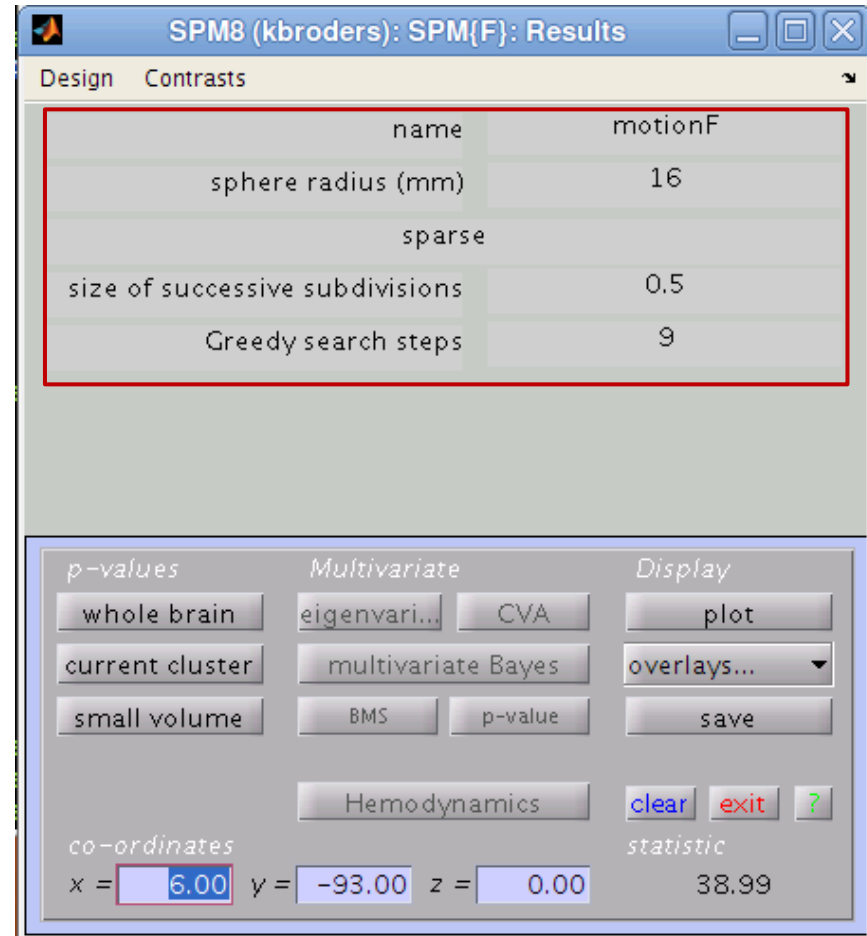


We place the cursor onto the region of interest.

# Multivariate Bayes: example

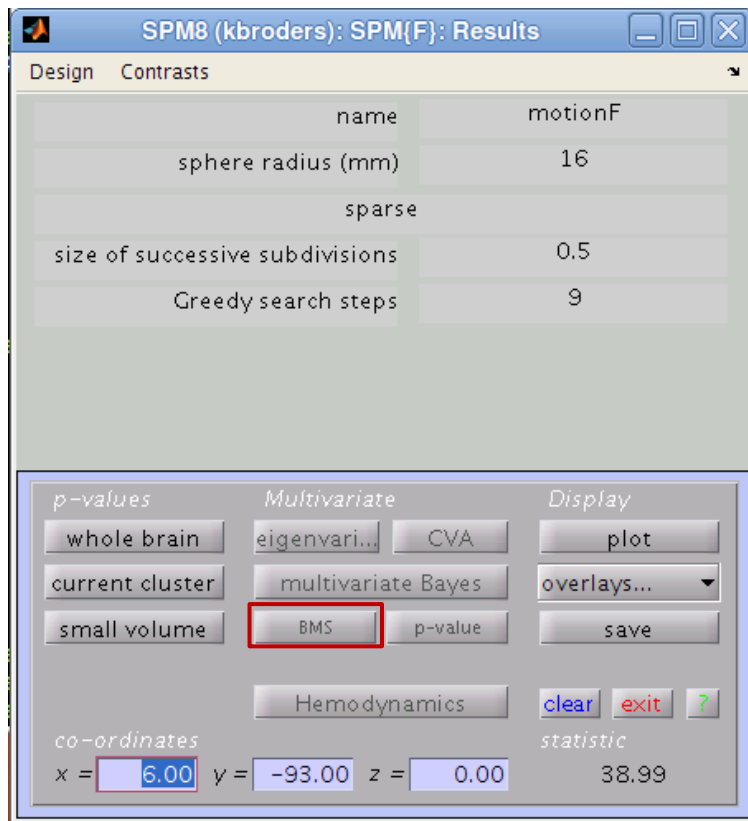


Multivariate Bayes can be invoked from within the Multivariate section.

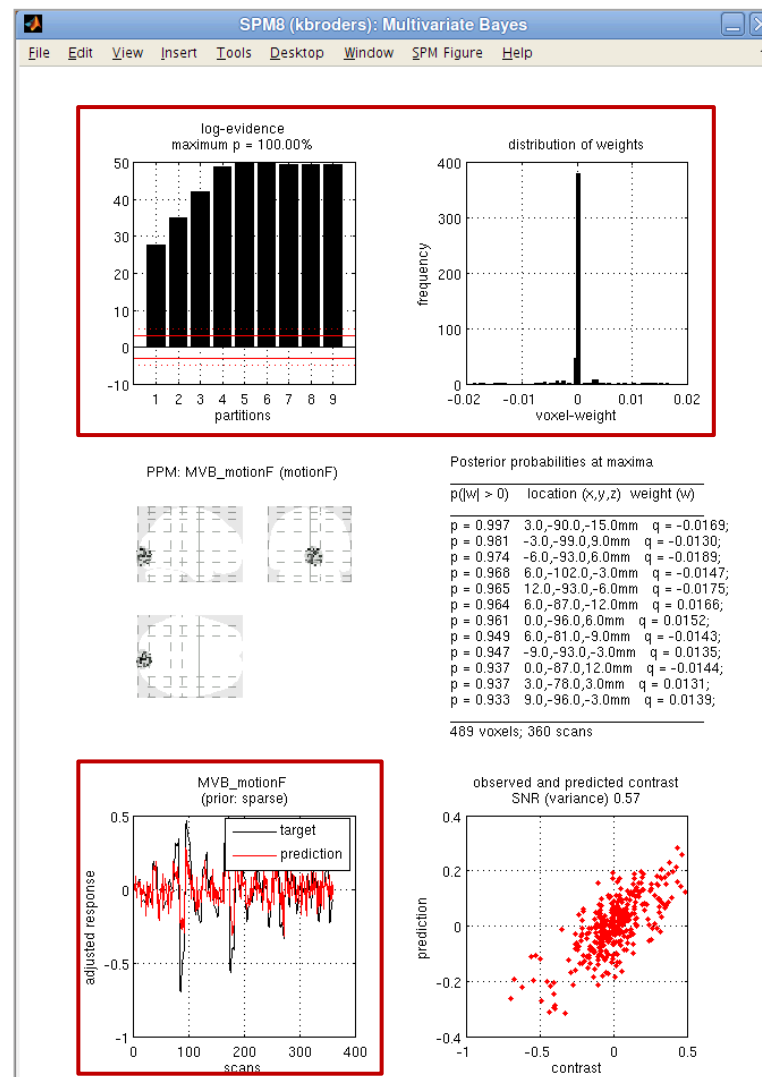


We specify the region of interest as a sphere around the cursor. We examine the *sparse* coding hypothesis.

# Multivariate Bayes: example

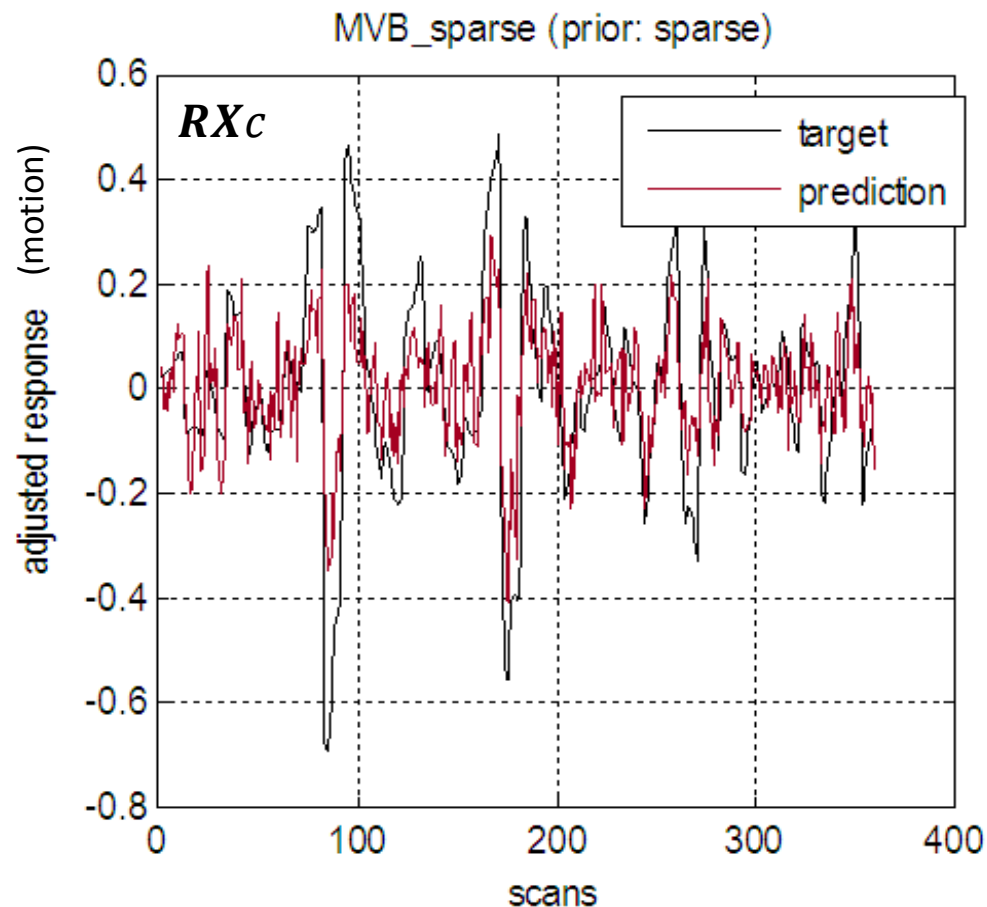


To display results, we use the button for Bayesian model selection (BMS).



# Multivariate Bayes: example

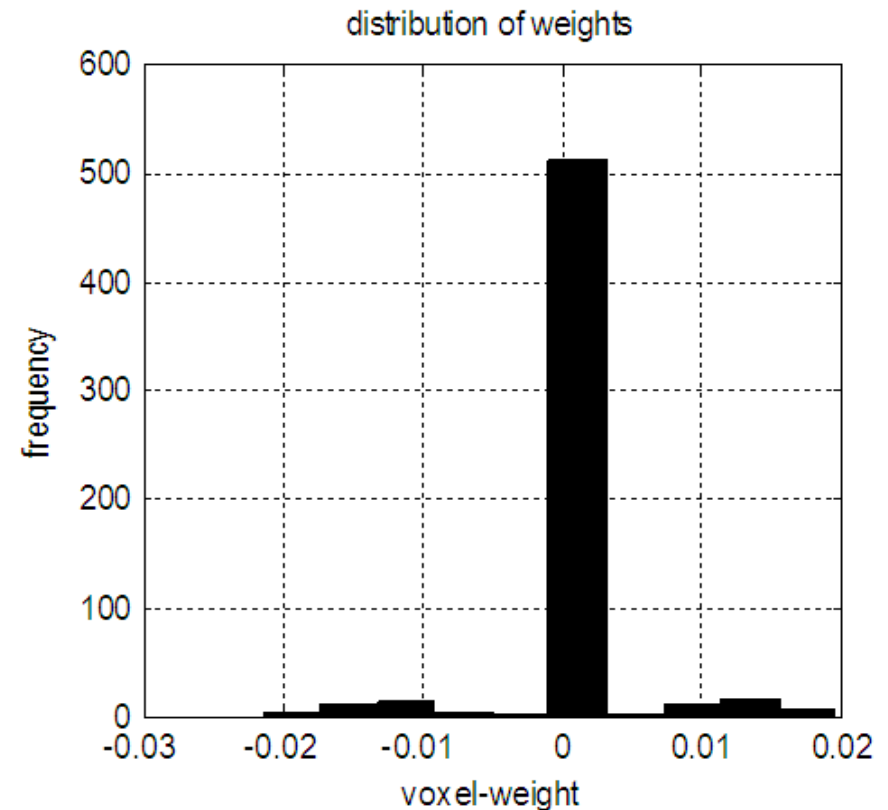
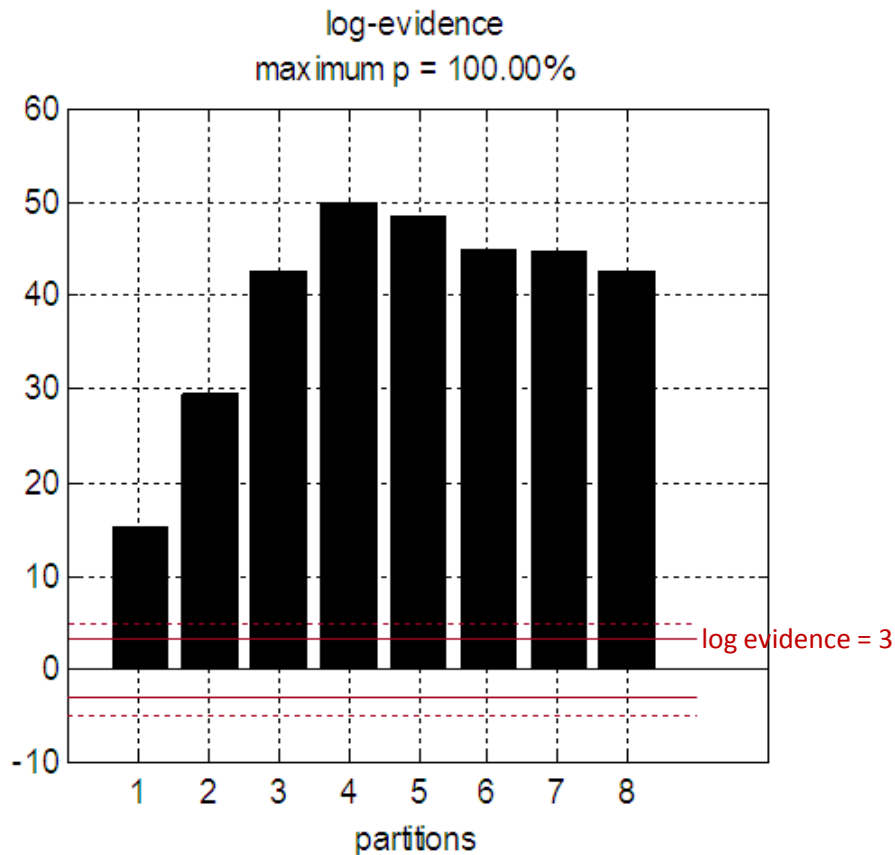
MVB-based predictions closely match the observed responses. But crucially, they don't perfectly match them. Perfect match would indicate overfitting.





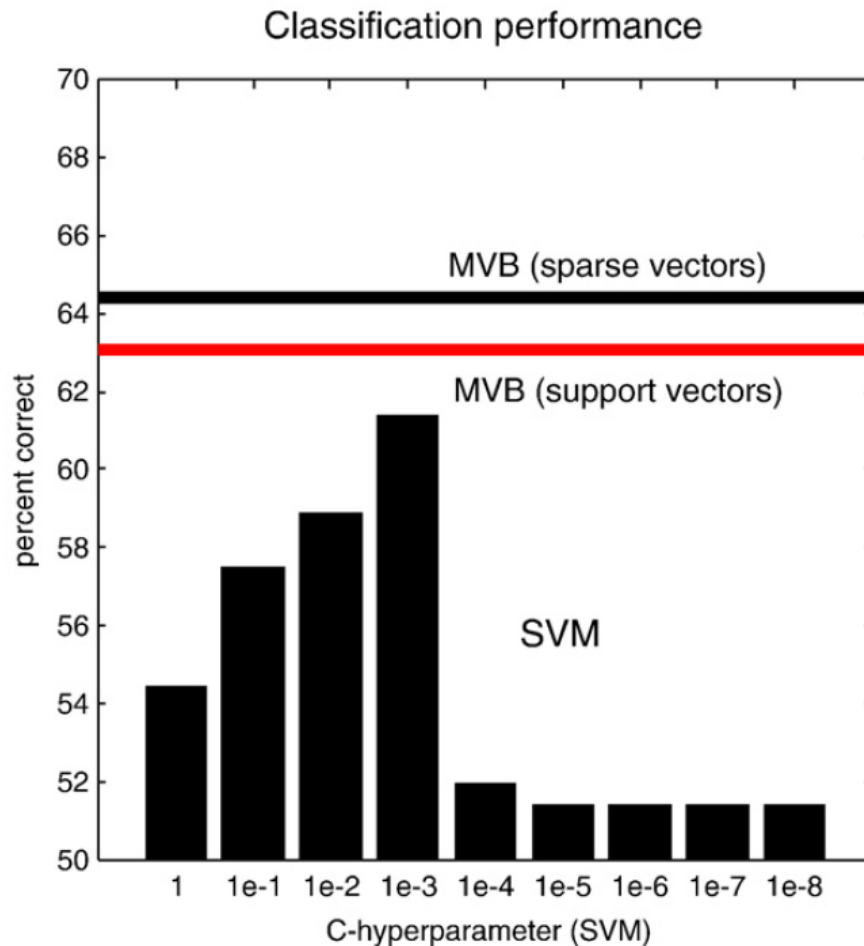
# Multivariate Bayes: example

The weights attributed to each voxel in the sphere are sparse and multimodal. This suggests sparse coding.



# Multivariate Bayes: example

MVB may outperform conventional point classifiers when using a more appropriate coding hypothesis.



# Outline

---

- 1 Foundations
- 2 Classification
- 3 Multivariate Bayes
- 4 Further model-based approaches**

# Recall: challenges for multivariate approaches

---

## 1 Variable selection

Given tens of thousands of voxels and very few trials of data, how do we find those brain regions that are jointly informative of some variable of interest?

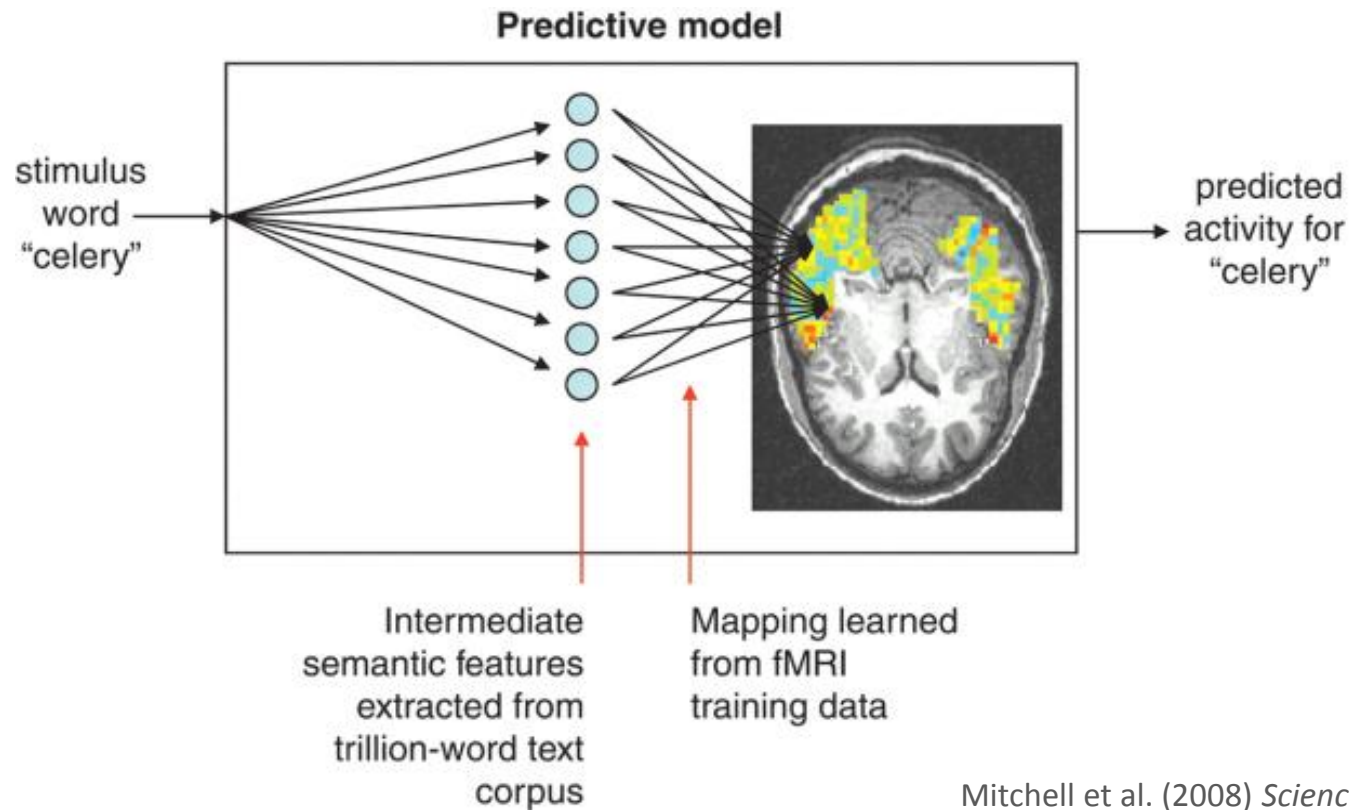
## 2 Neurobiological interpretability

How can we obtain results that are mechanistically interpretable in the context of the underlying neurobiological system?

# Identification / inferring a representational space

## Approach

1. estimation of an encoding model
2. nearest-neighbour classification or voting

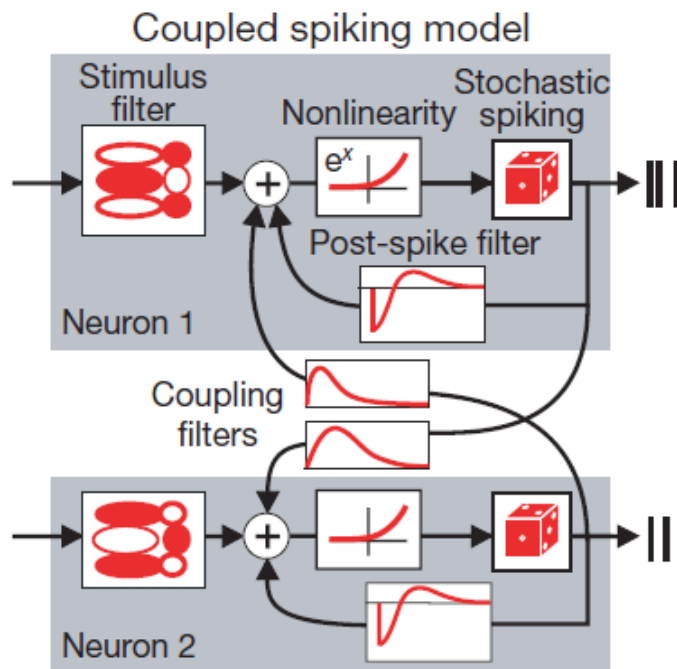


Mitchell et al. (2008) *Science*

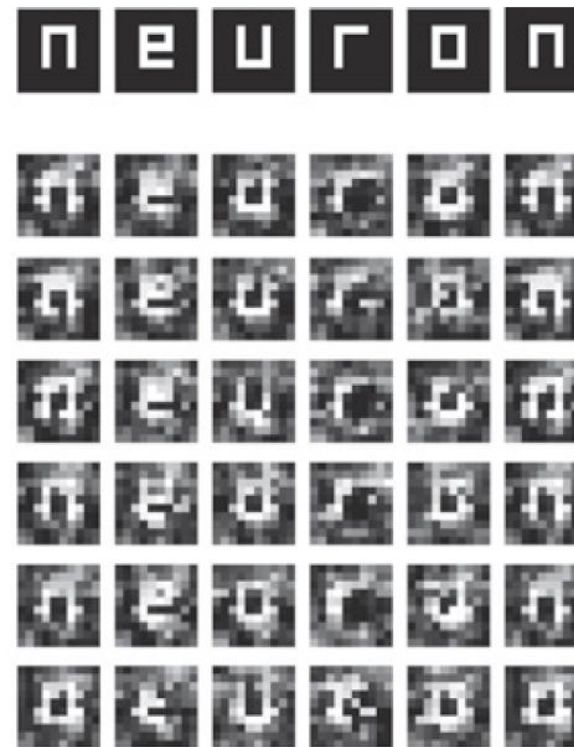
# Reconstruction / optimal decoding

## Approach

1. estimation of an encoding model
2. model inversion

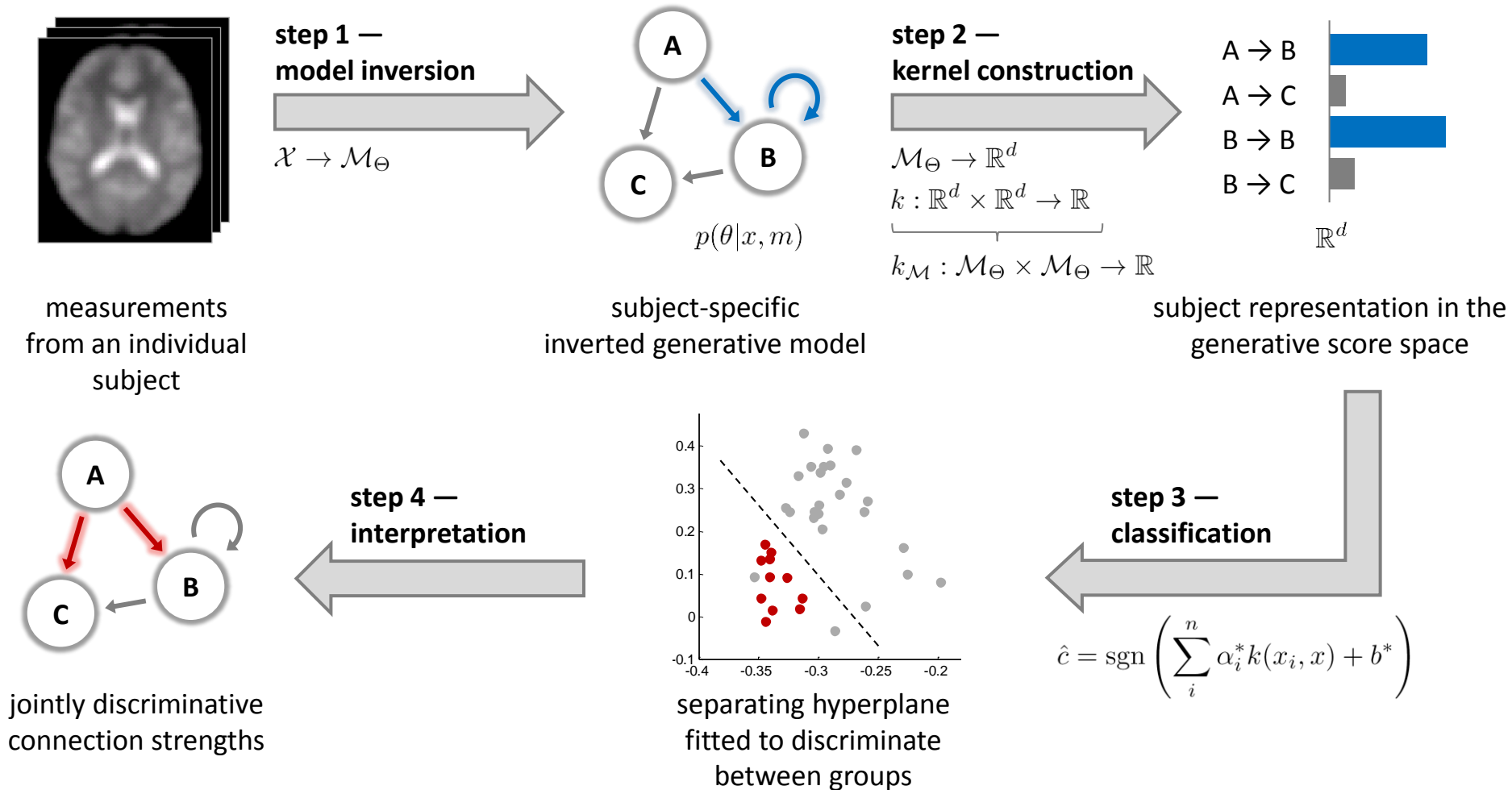


Paninski et al. (2007) *Progr Brain Res*  
Pillow et al. (2008) *Nature*



Miyawaki et al. (2009) *Neuron*

# Model-based classification by generative embedding



Brodersen, Haiss, Ong, Jung, Tittgemeyer, Buhmann, Weber, Stephan (2011) *NeuroImage*  
 Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann, Stephan (2011) *PLoS Comput Biol*

# Summary

---

- 1. Foundations.** Multivariate methods can uncover and exploit information jointly encoded by multiple voxels. Remember the distinction between prediction and inference, encoding and decoding, univariate and multivariate, and classification and regression.
- 2. Classification.** Classification studies typically aim to examine (i) overall discriminability, (ii) the spatial deployment of informative regions, (iii) the temporal evolution of discriminative activity, and (iv) the nature of the distributed activity.
- 3. Multivariate Bayes.** Multivariate Bayes offers an alternative scheme that maps multivariate patterns of activity onto brain states within the conventional statistical framework of hierarchical models and their inversion.
- 4. Model-based approaches.** Model-based approaches aim to augment previous methods by neurobiological interpretability and are likely to become very fruitful in the future.