Multivariate analyses & decoding

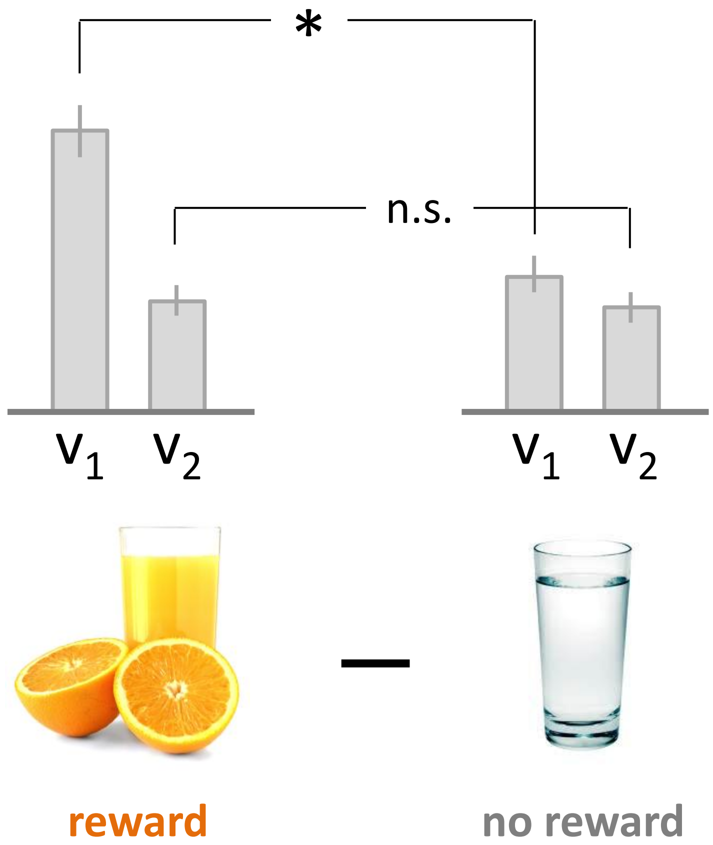
Kay H. Brodersen

Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
University of Zurich & ETH Zurich

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

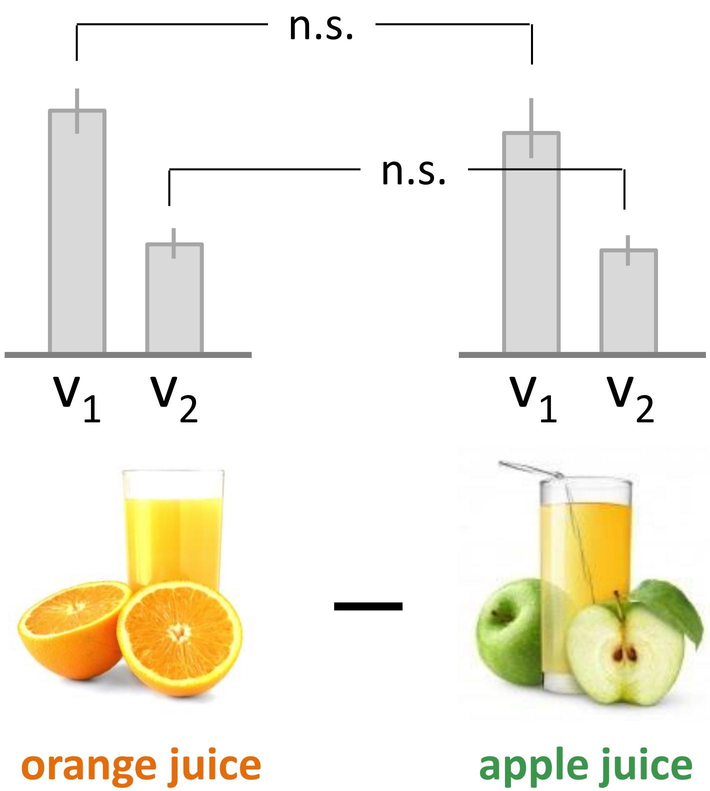
Why multivariate?

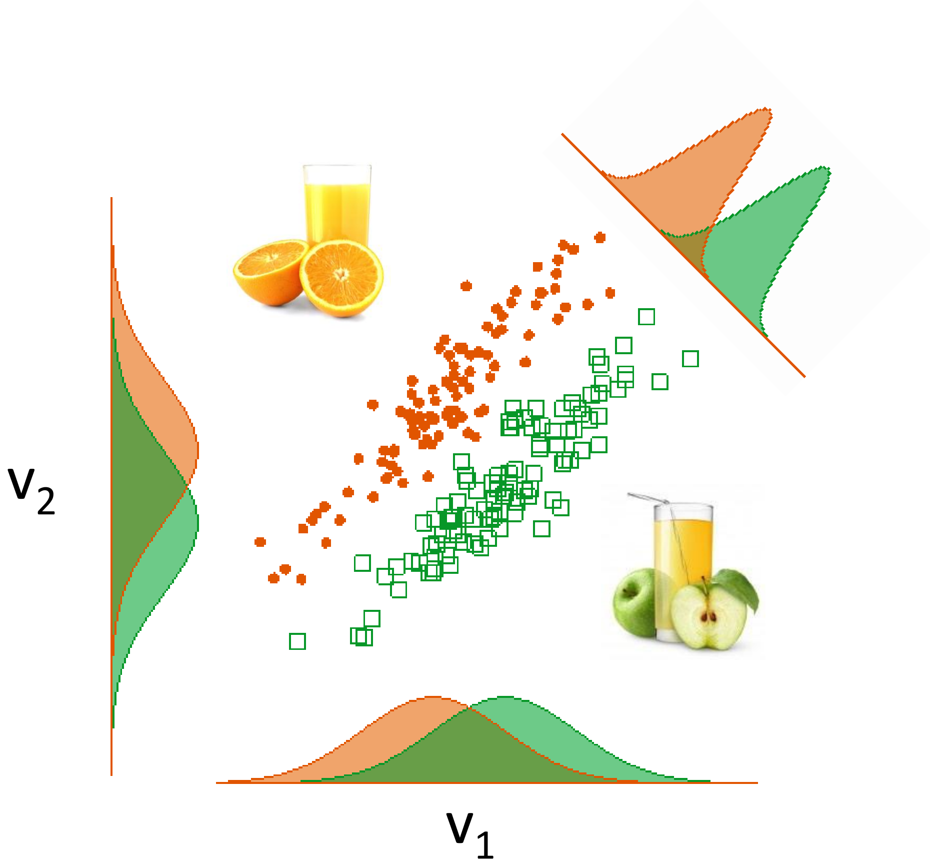
Univariate approaches are excellent for localizing activations in individual voxels.



Why multivariate?

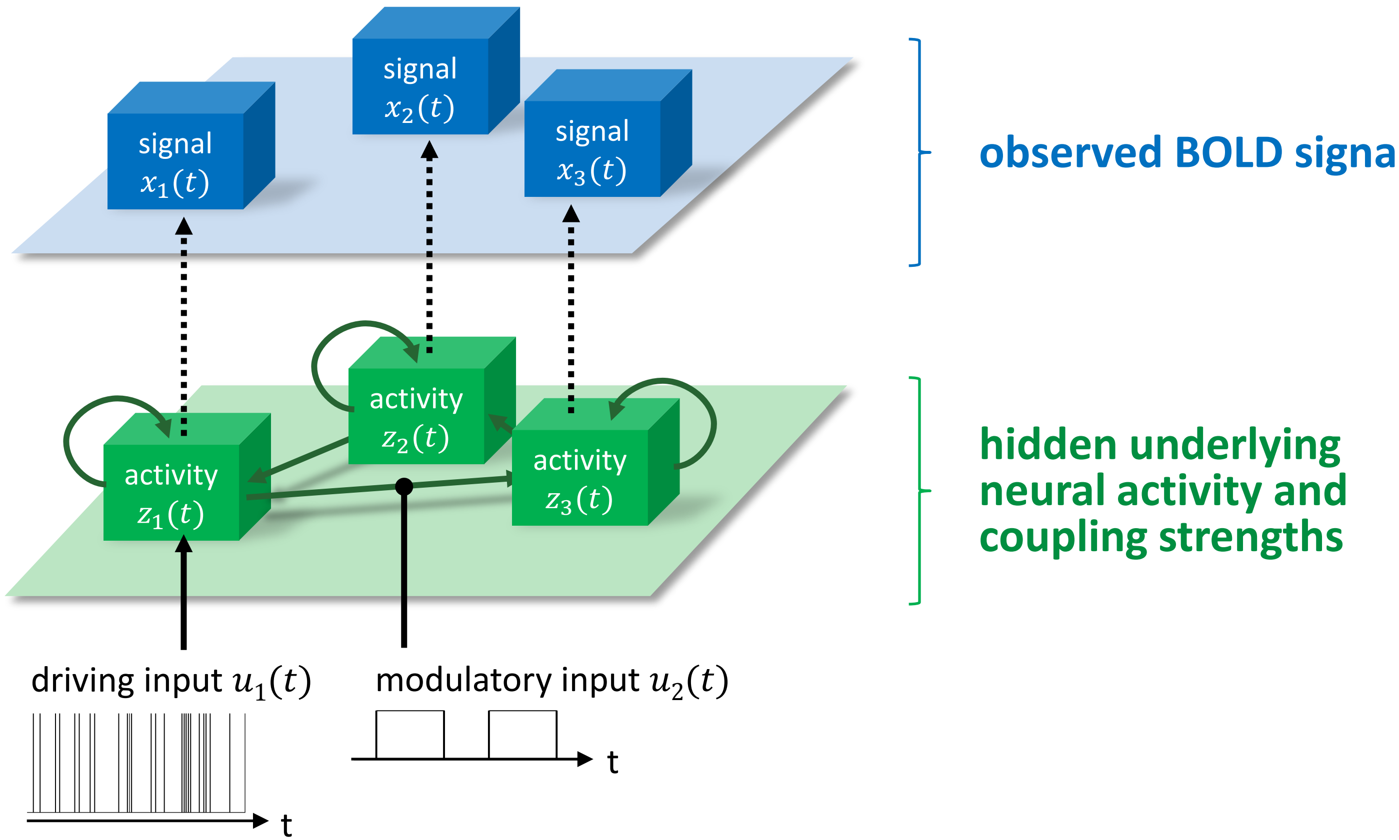
Multivariate approaches can be used to examine responses that are jointly encoded in multiple voxels.





Why multivariate?

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



Friston, Harrison & Penny (2003) *NeuroImage;* Stephan & Friston (2007) *Handbook of Brain Connectivity;* Stephan et al. (2008) *NeuroImage*

Overview

1 Modelling principles

2 Classification

3 Multivariate Bayes

4 Generative embedding

Encoding vs. decoding

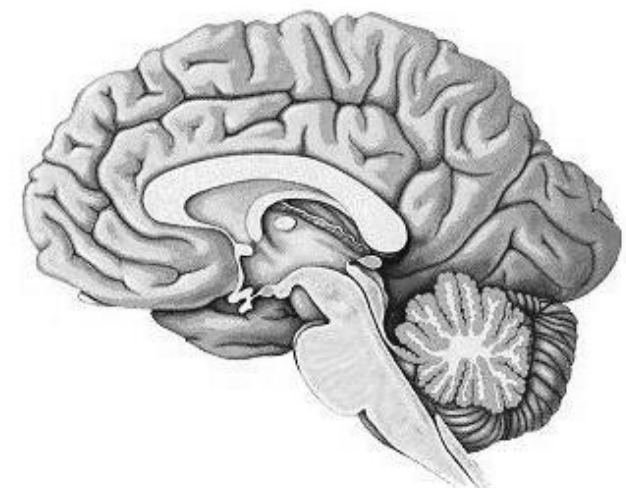
condition
stimulus
response
prediction error

encoding model

$g: X_t \to Y_t$

decoding model

$h: Y_t \to X_t$



context (cause or consequence)

$X_t \in \mathbb{R}^d$

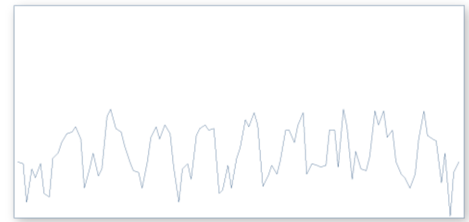
BOLD signal

$Y_t \in \mathbb{R}^v$

Regression vs. classification

Regression model

independent variables (regressors) → f → **continuous** dependent variable



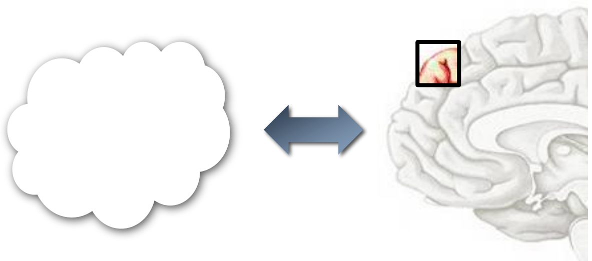
Classification model

independent variables (features) → f → **categorical** dependent variable (label)

vs.

Univariate vs. multivariate models

A univariate model considers a single voxel at a time.

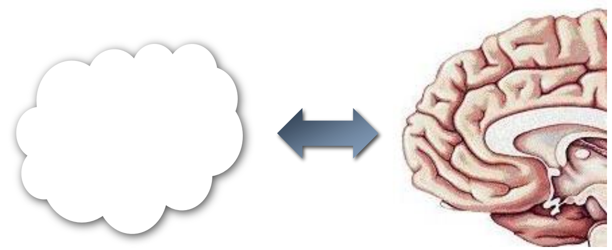


context
$X_t \in \mathbb{R}^d$

BOLD signal
$Y_t \in \mathbb{R}$

Spatial dependencies between voxels are only introduced afterwards, through random field theory.

A multivariate model considers many voxels at once.



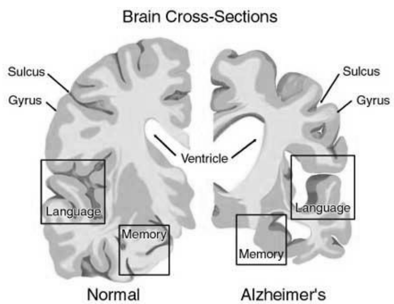
context
$X_t \in \mathbb{R}^d$

BOLD signal
$Y_t \in \mathbb{R}^v, \; v \gg 1$

Multivariate models enable inferences on distributed responses without requiring focal activations.

Prediction vs. inference

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



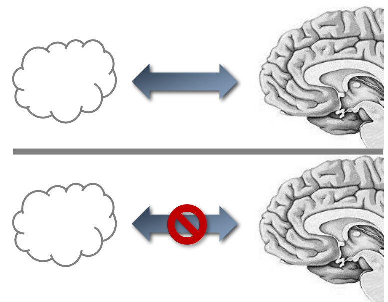
predicting a cognitive state using a brain-machine interface

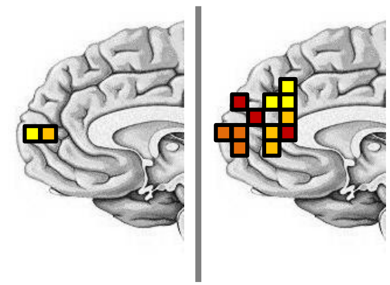
predicting a subject-specific diagnostic status

predictive density

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

The goal of **inference** is to decide between competing hypotheses.





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

weighing the evidence for sparse vs. distributed coding

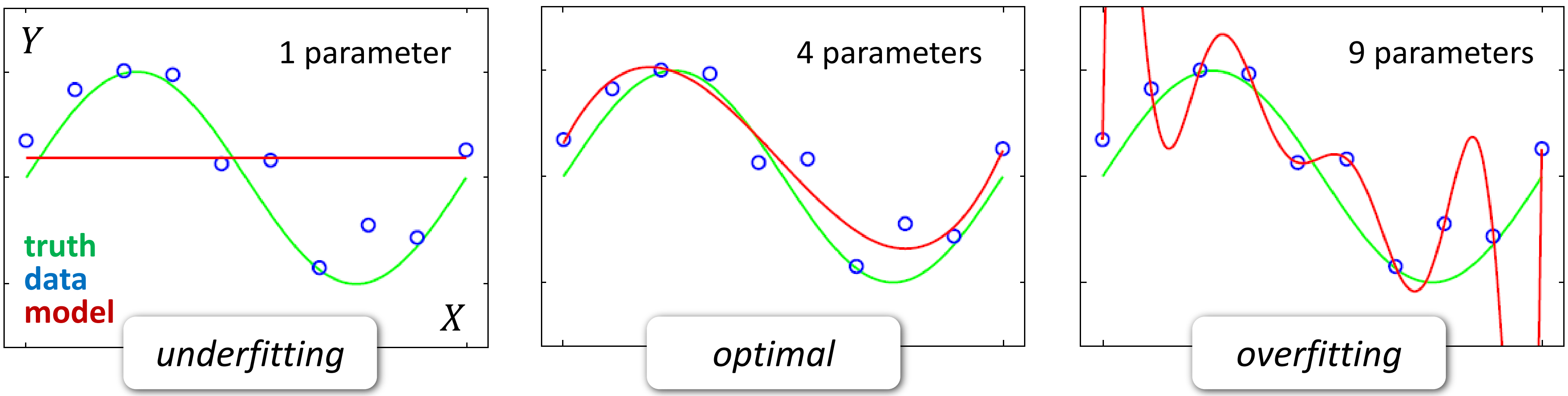
marginal likelihood (model evidence)

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

Goodness of fit vs. complexity

Goodness of fit is the degree to which a model explains observed data.

Complexity is the flexibility of a model (including, but not limited to, its number of parameters).



We wish to find the model that optimally trades off goodness of fit and complexity.

Bishop (2007) *PRML*

Summary of modelling terminology

General Linear Model (GLM)

- mass-univariate encoding model
- to explain brain activity from context and find clusters of similar effects

Dynamic Causal Modelling (DCM)

- multivariate encoding model
- to evaluate connectivity hypotheses

Classification

- multivariate decoding model
- to predict a categorical context label from brain activity

Multivariate Bayes (MVB)

- multivariate decoding model
- to evaluate anatomical and coding hypotheses

Overview

1 Modelling principles

2 Classification

3 Multivariate Bayes

4 Generative embedding

Constructing a classifier

A principled way of designing a classifier would be to adopt a probabilistic approach:

$Y_t \xrightarrow{\quad f \quad}$ that k which maximizes $p(X_t = k | Y_t, X, Y)$

In practice, classifiers differ in terms of how strictly they implement this principle.

Generative classifiers

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

- *Gaussian naïve Bayes*
- *Linear discriminant analysis*

Discriminative classifiers

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

- *Logistic regression*
- *Relevance vector machine*
- *Gaussian process classifier*

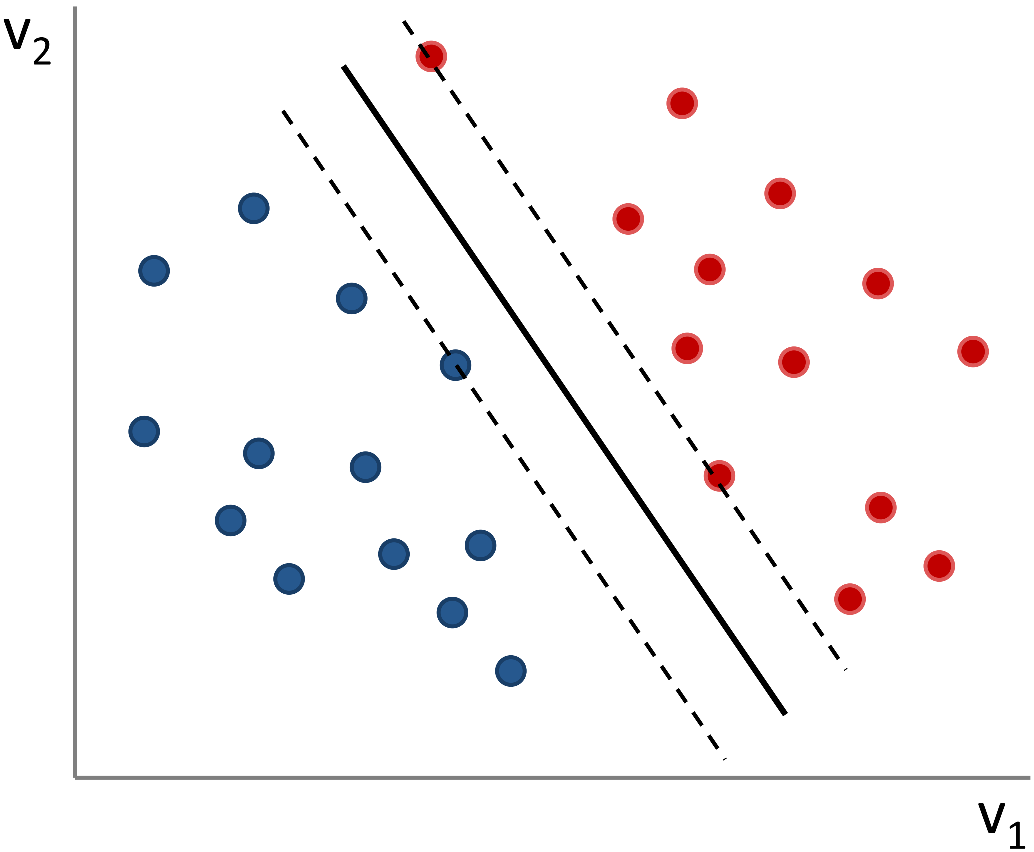
Discriminant classifiers

estimate $f(Y_t)$ directly

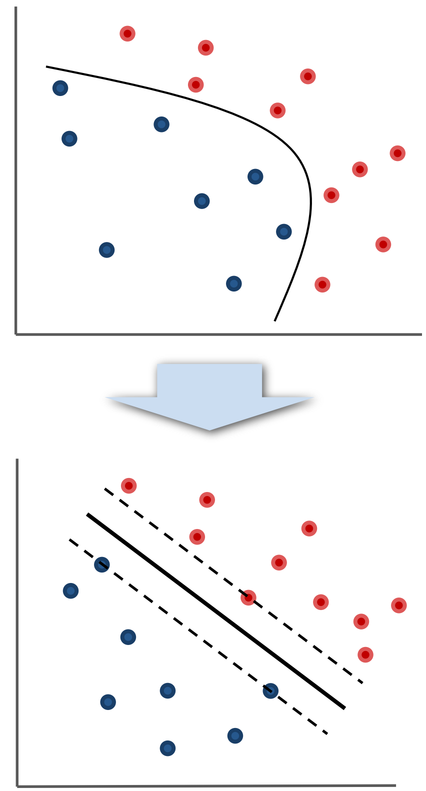
- *Fisher's linear discriminant*
- *Support vector machine*

Support vector machine (SVM)

Linear SVM

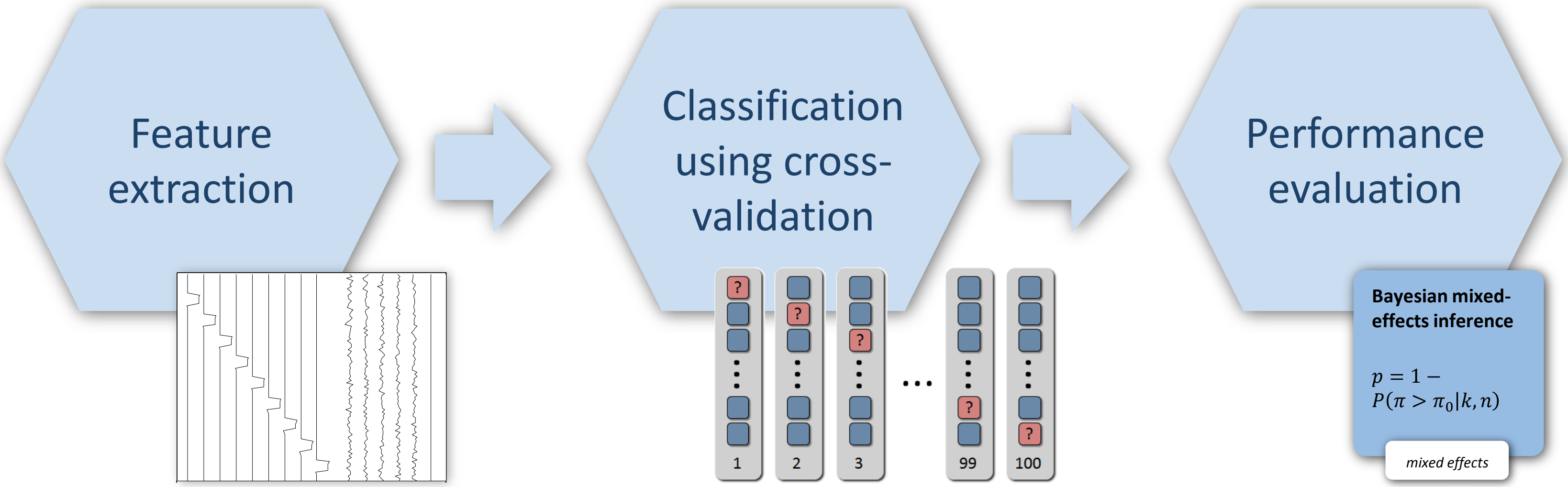


Nonlinear SVM



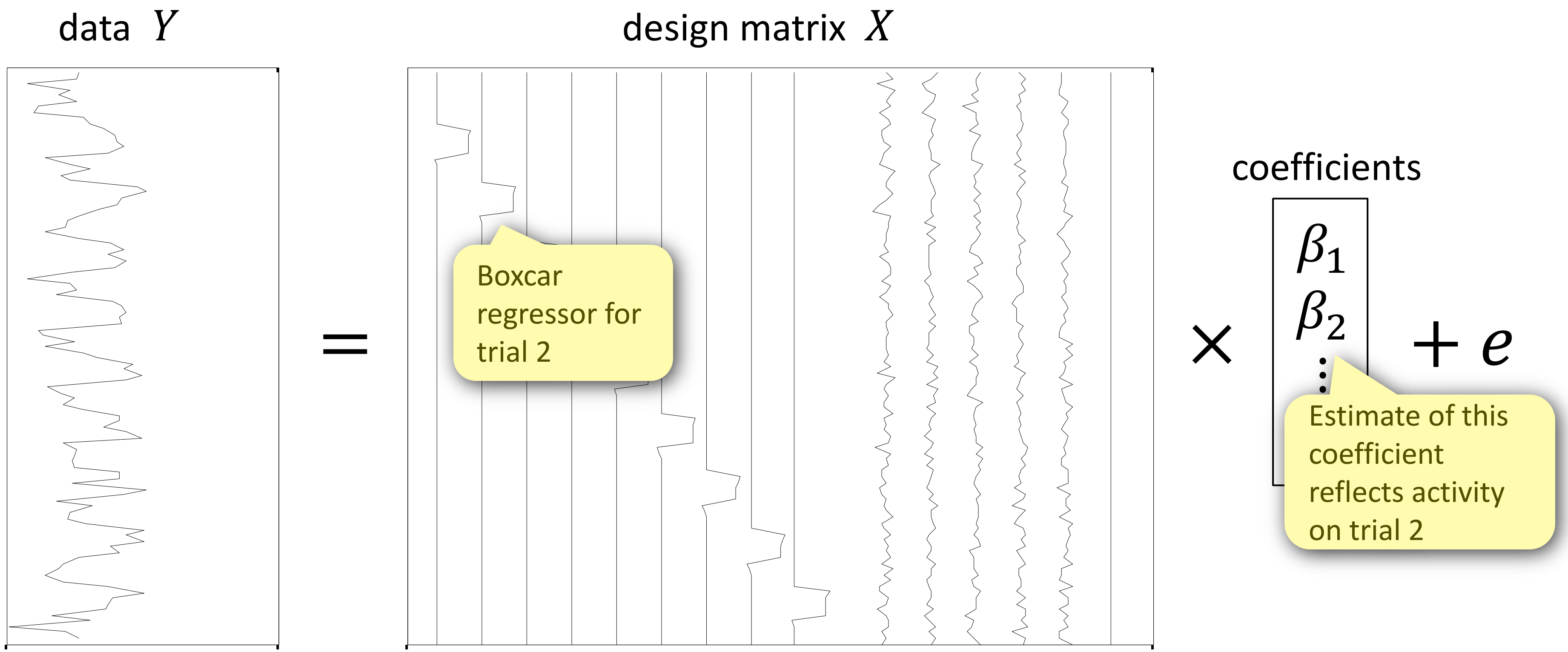
Vapnik (1999) Springer; Schölkopf et al. (2002) MIT Press

Stages in a classification analysis



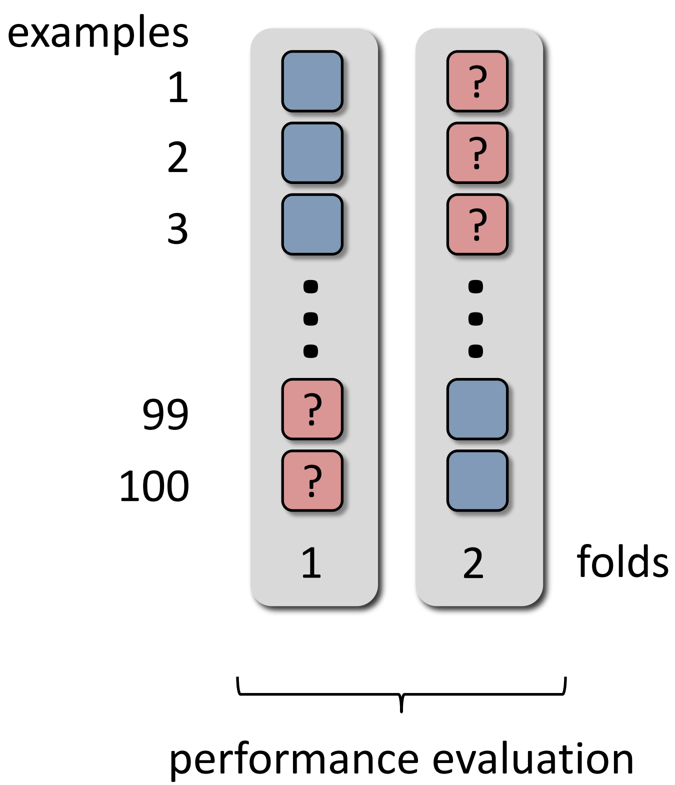
Feature extraction for trial-by-trial classification

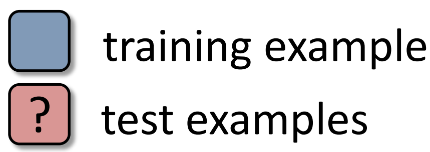
We can obtain trial-wise estimates of neural activity by filtering the data with a GLM.



Cross-validation

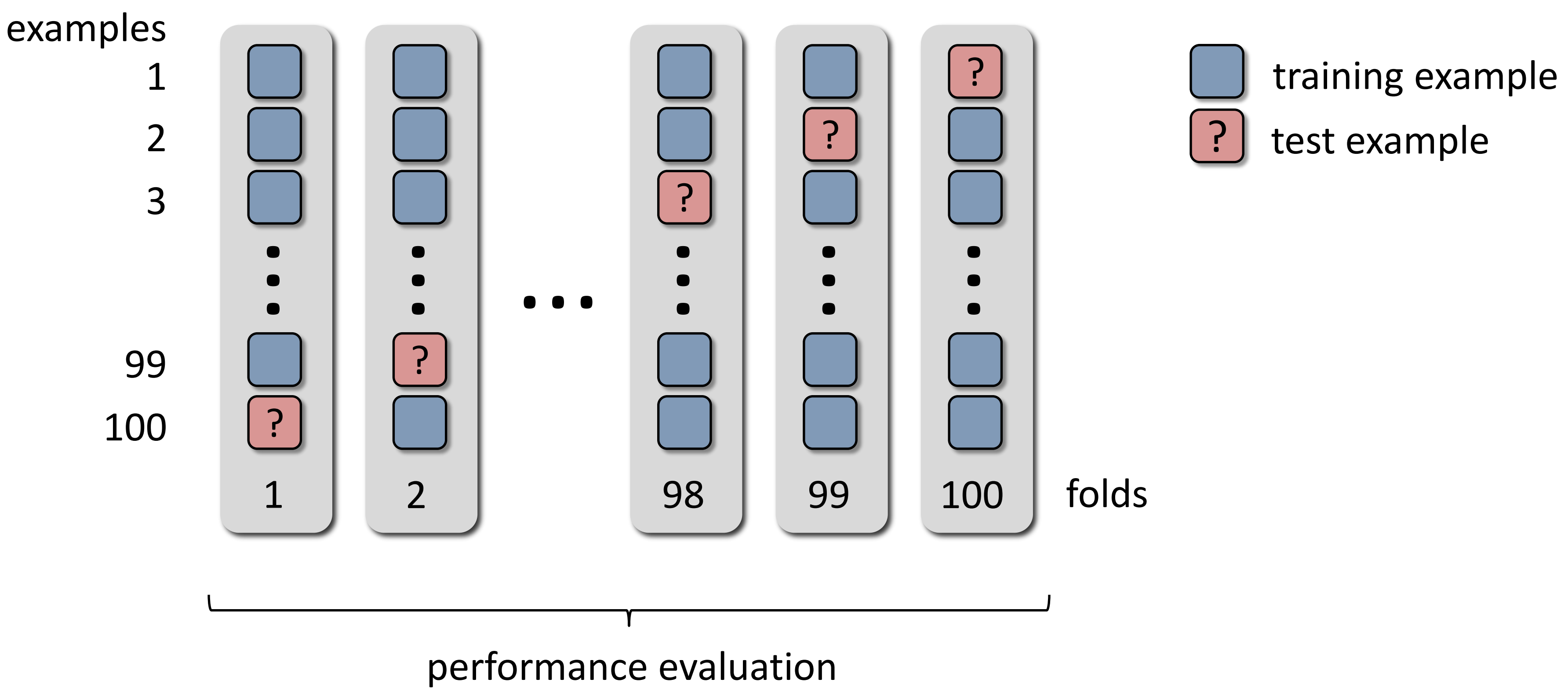
The generalization ability of a classifier can be estimated using a resampling procedure known as *cross-validation*. One example is 2-fold cross-validation:





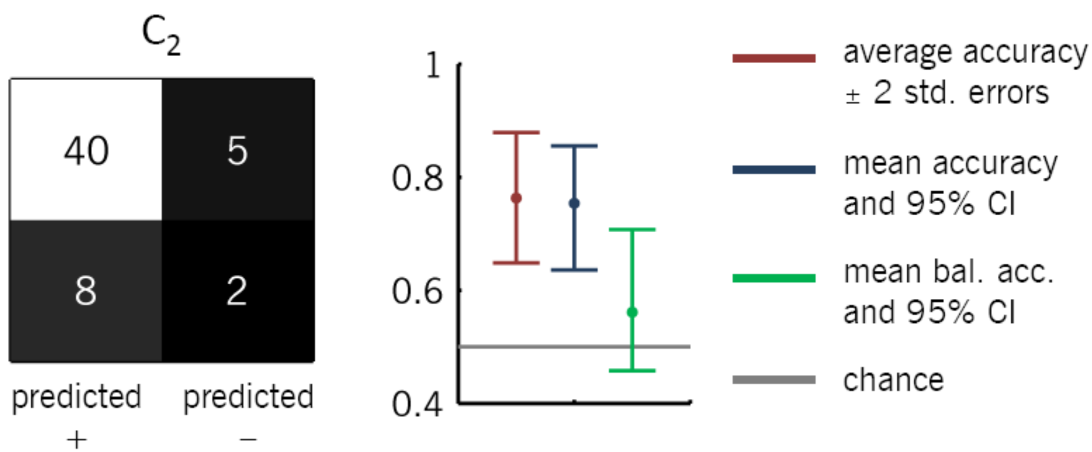
Cross-validation

Another commonly used variant is *leave-one-out* cross-validation.



Performance evaluation

- Evaluating the performance of a classification algorithm critically requires a measure of the degree to which unseen examples have been identified with their correct class labels.
- The procedure of averaging across accuracies obtained on individual cross-validation folds is flawed in two ways. First, it does not allow for the derivation of a meaningful confidence interval. Second,it leads to an optimistic estimate when a biased classifier is tested on an imbalanced dataset.
- Both problems can be overcome by replacing the conventional point estimate of accuracy by an estimate of the posterior distribution of the *balanced accuracy*.

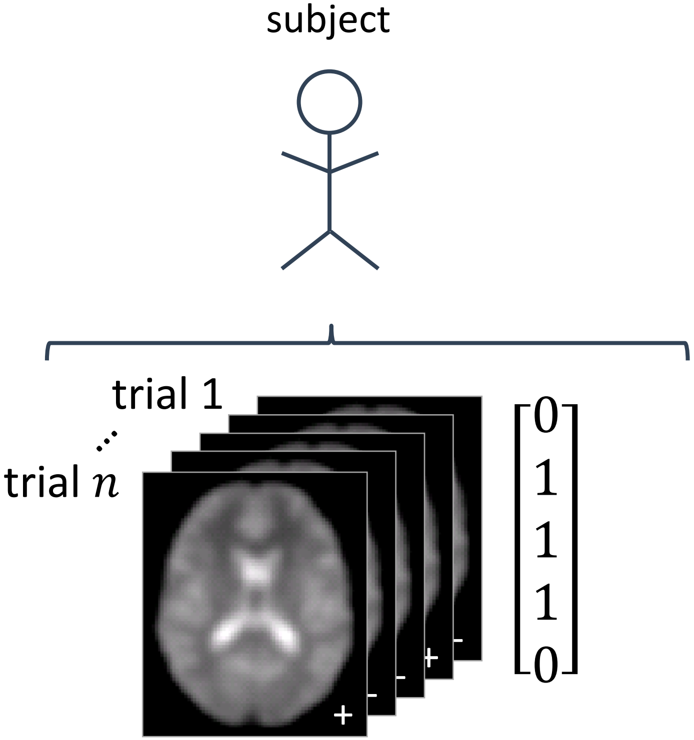


Brodersen, Ong, Buhmann, Stephan (2010) *ICPR*

Performance evaluation

Single-subject study with n trials

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



Binomial test

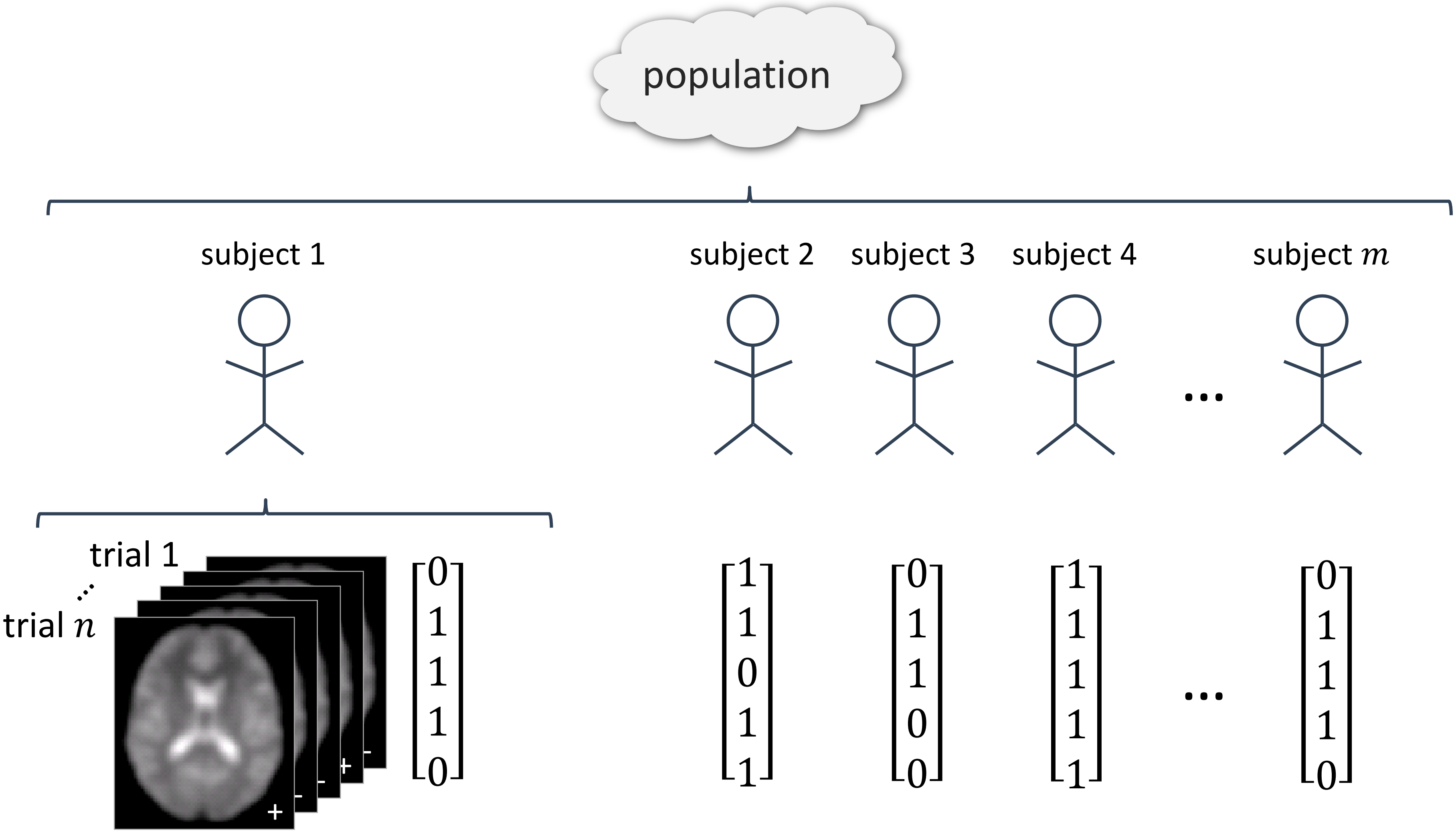
$$p = P(X \geq k | H_0) = 1 - B(k | n, \pi_0)$$

In MATLAB:

```
p = 1 - binocdf(k,n,pi_0)
```

k	number of correctly classified trials
n	total number of trials
π_0	chance level (typically 0.5)
B	binomial cumulative density function

Performance evaluation



Performance evaluation

Group study with m subjects, n trials each

In a group setting, we must account for both within-subjects (fixed-effects) and between-subjects (random-effects) variance components.

Binomial test on concatenated data

$p = 1 -$
$B(\sum k | \sum n, \pi_0)$

fixed effects

Binomial test on averaged data

$p = 1 -$
$B\left(\frac{1}{m}\sum k \middle| \frac{1}{m}\sum n, \pi_0\right)$

fixed effects

t-test on summary statistics

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

$p = 1 - t_{m-1}(t)$

random effects

Bayesian mixed-effects inference

$p = 1 -$
$P(\pi > \pi_0 | k, n)$

available for MATLAB and R

mixed effects

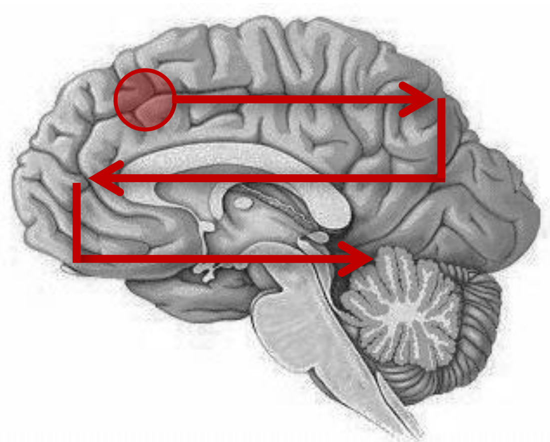
$\bar{\pi}$	sample mean of sample accuracies	π_0	chance level (typically 0.5)
$\hat{\sigma}_{m-1}$	sample standard deviation	t_{m-1}	cumulative Student's t-distribution

Brodersen, Mathys, Chumbley, Daunizeau, Ong, Buhmann, Stephan (2012) JMLR
Brodersen, Daunizeau, Mathys, Chumbley, Buhmann, Stephan *(under review)*

Spatial deployment of informative regions

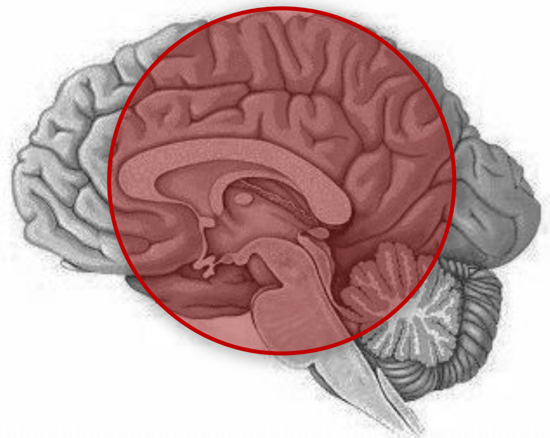
Which brain regions are jointly informative of a cognitive state of interest?

Searchlight approach



A sphere is passed across the brain. At each location, the classifier is evaluated using only the voxels in the current sphere → map of t-scores.

Whole-brain approach



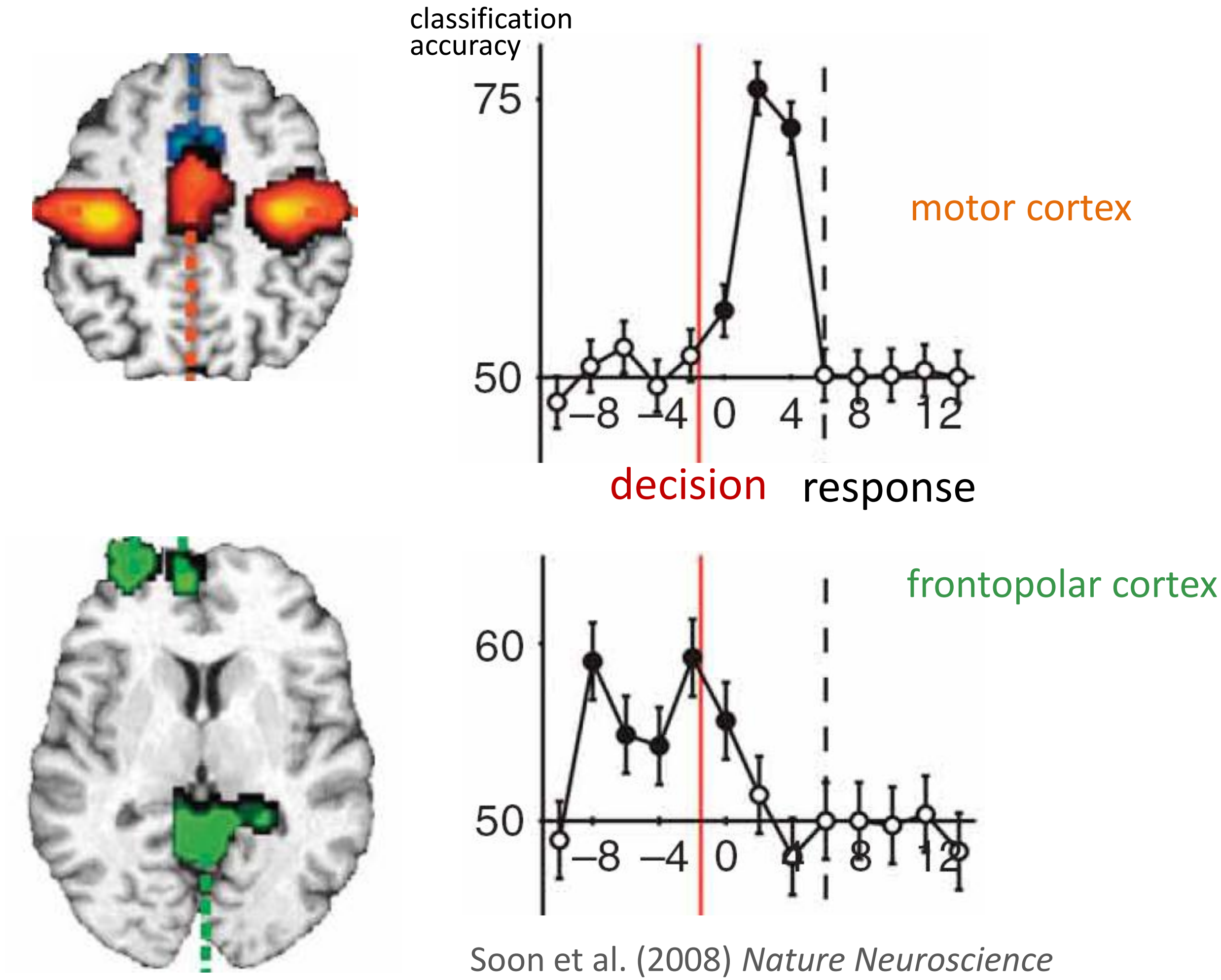
A constrained classifier is trained on whole-brain data. Its voxel weights are related to their empirical null distributions using a permutation test → map of t-scores.

Nandy & Cordes (2003) *MRM*
Kriegeskorte et al. (2006) *PNAS*

Mourao-Miranda et al. (2005) *NeuroImage*

Temporal evolution of discriminability

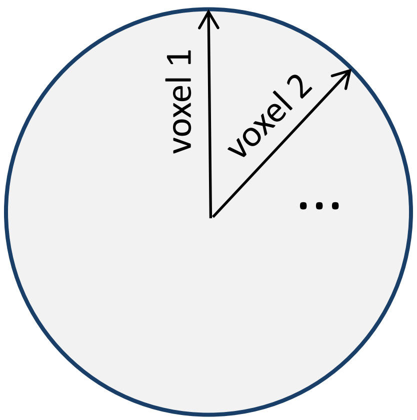
Example – decoding which button the subject pressed



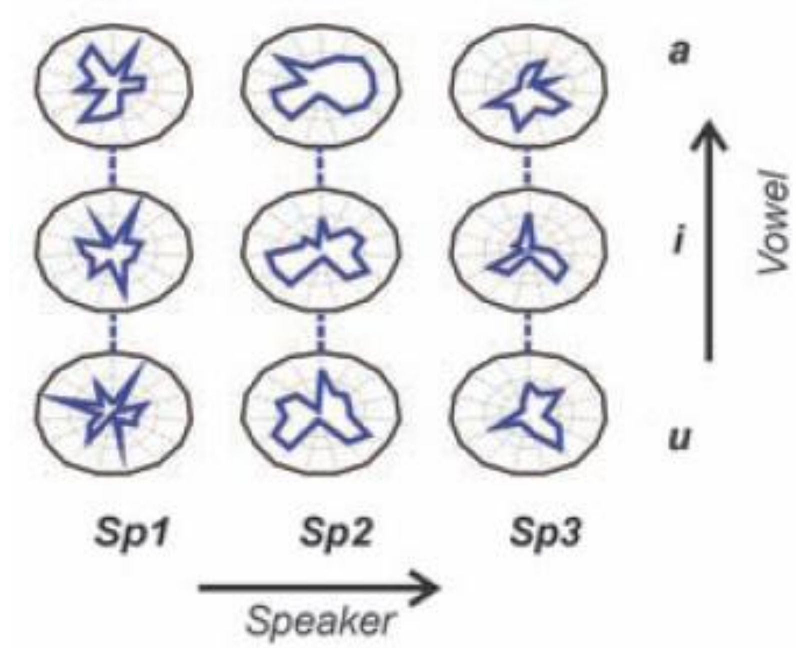
Soon et al. (2008) *Nature Neuroscience*

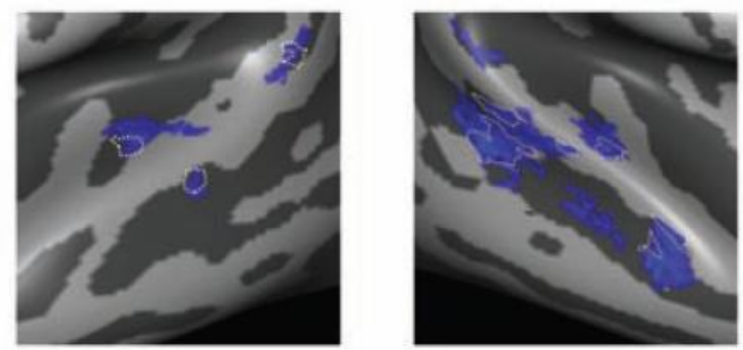
Pattern characterization

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



fingerprint plot
(one plot per class)

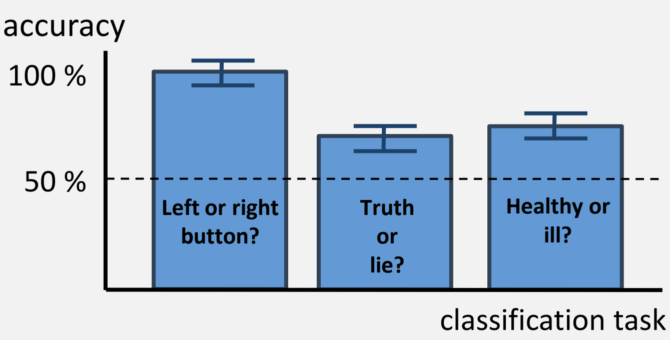




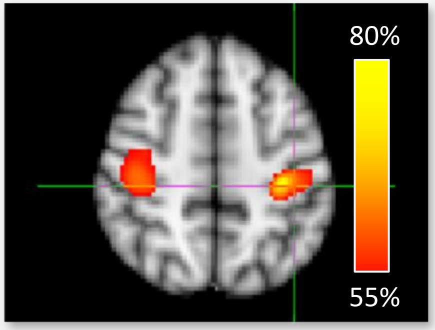
Formisano et al. (2008) *Science*

Research questions for classification

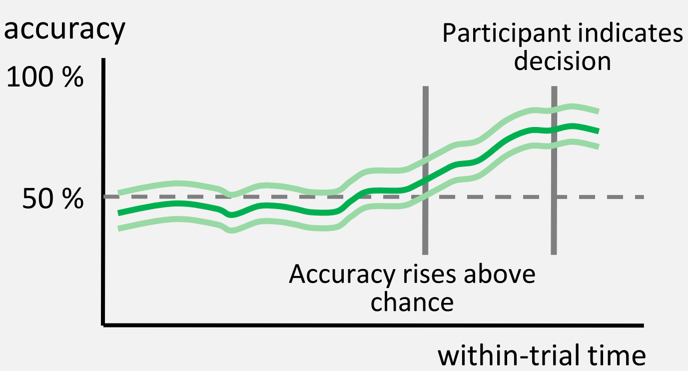
Overall classification accuracy



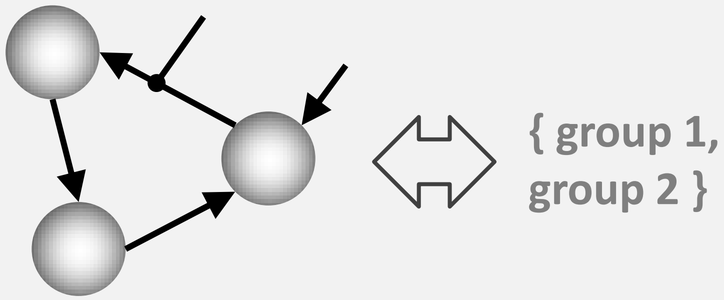
Spatial deployment of discriminative regions



Temporal evolution of discriminability



Model-based classification



Pereira et al. (2009) *NeuroImage*, Brodersen et al. (2009) *The New Collection*

Overview

1 Modelling principles

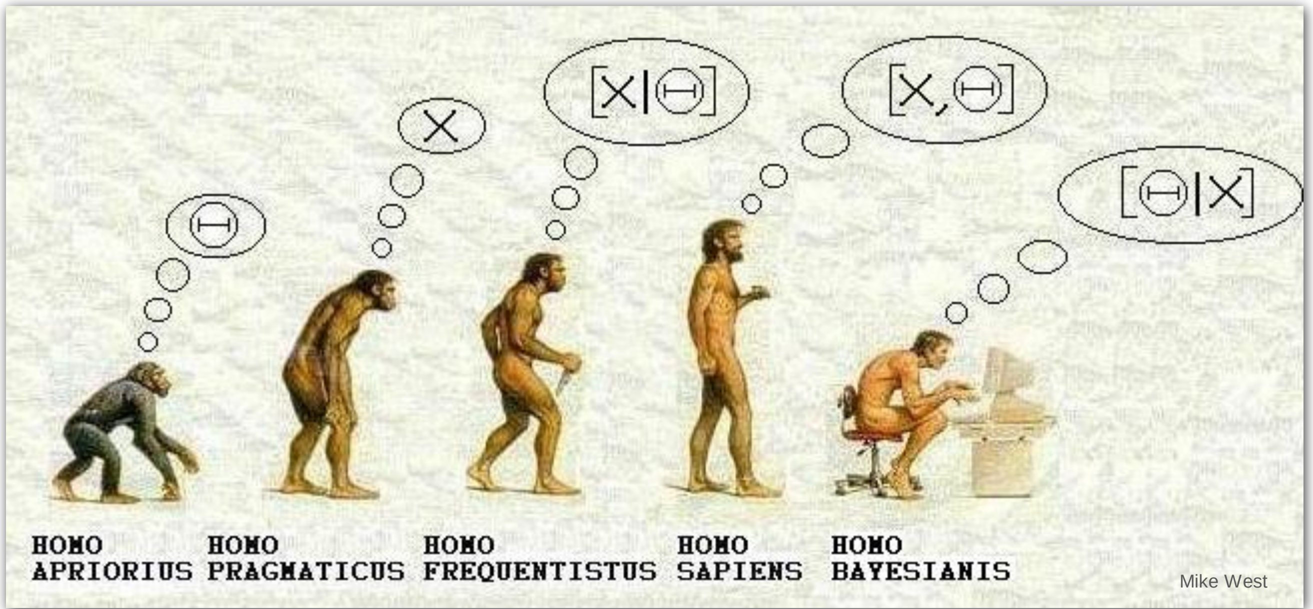
2 Classification

3 Multivariate Bayes

4 Generative embedding

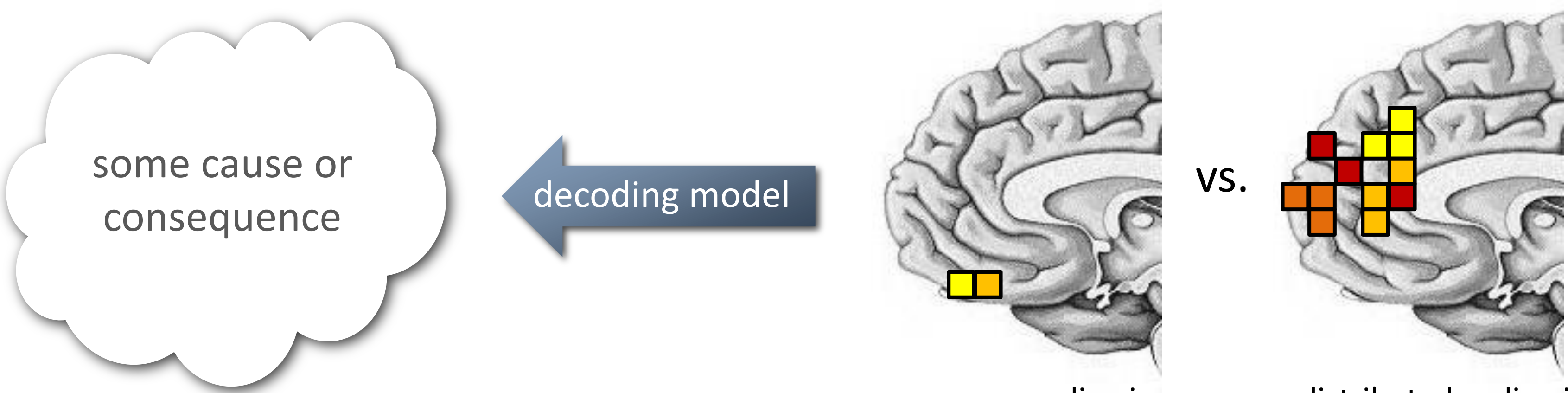
Multivariate Bayes

SPM brings multivariate analyses into the conventional inference framework of Bayesian hierarchical models and their inversion.



Multivariate Bayes

Multivariate analyses in SPM rest on the central notion that inferences about how the brain represents things can be reduced to model comparison.

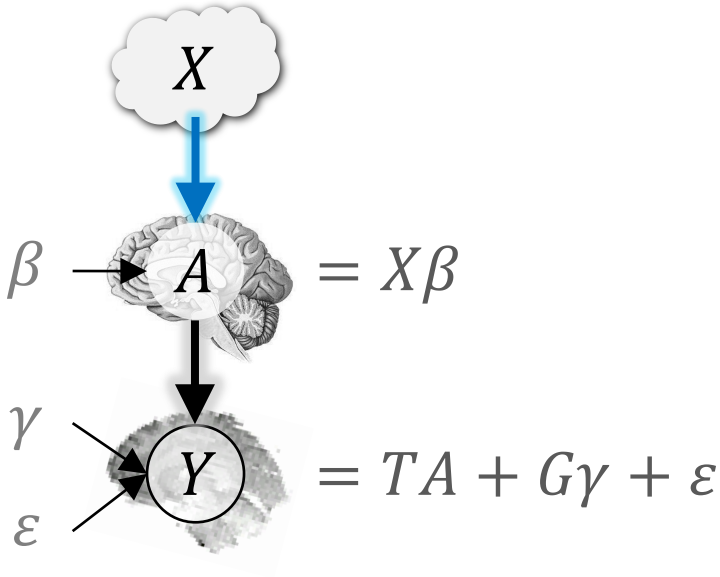


sparse coding in orbitofrontal cortex

distributed coding in prefrontal cortex

From encoding to decoding

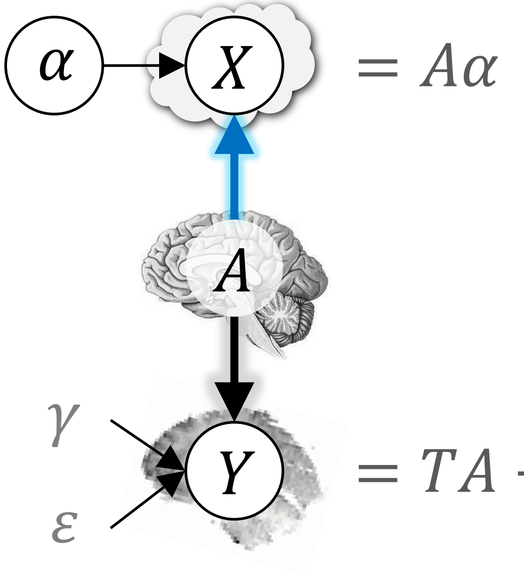
Encoding model: GLM

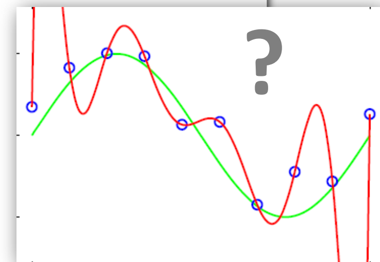


In summary:

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

Decoding model: MVB





In summary:

$TX = Y\alpha - G\gamma\alpha - \varepsilon\alpha$

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 α (the probability of falsely rejecting H_0, i.e., $1 -$ 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 α is: to reject H_0 when the likelihood ratio Λ exceeds a criticial 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 things 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 Λ and is therefore suboptimal.

Neyman & Person (1933) *Phil Trans Roy Soc London*
Kass & Raftery (1995) *J Am Stat Assoc*
Friston et al. (2009) *NeuroImage*

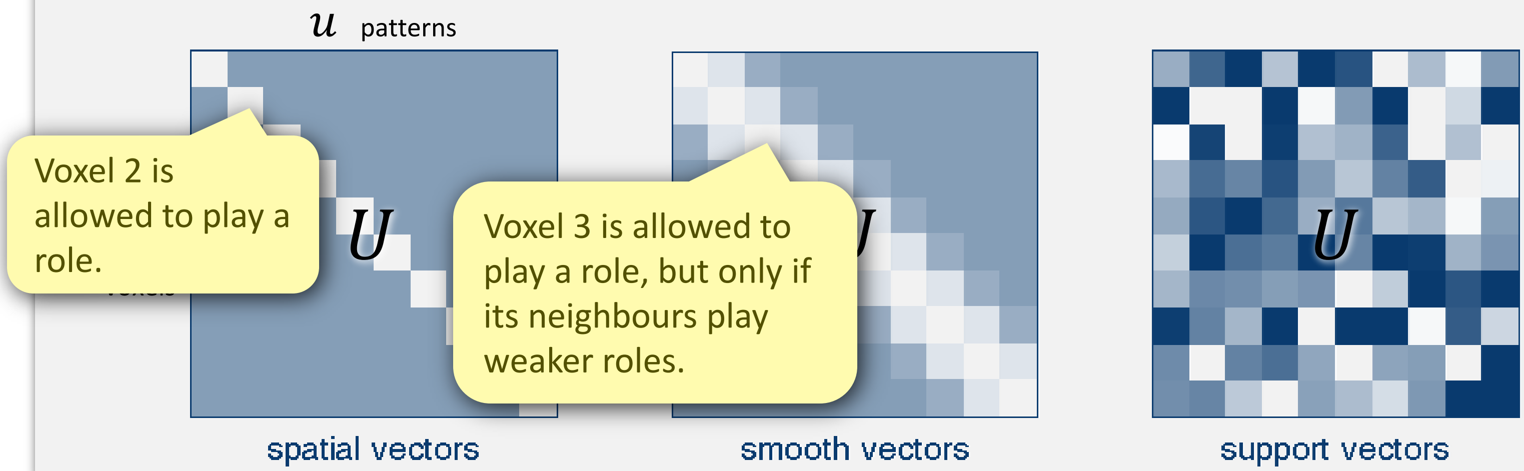
Specifying the prior for MVB

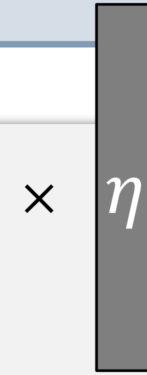
To make the ill-posed regression problem tractable, MVB uses a prior on voxel weights. Different priors reflect different anatomical and/or coding hypotheses.

Friston et al. 2008 *NeuroImage*

Specifying the prior for MVB

1st level – spatial coding hypothesis U





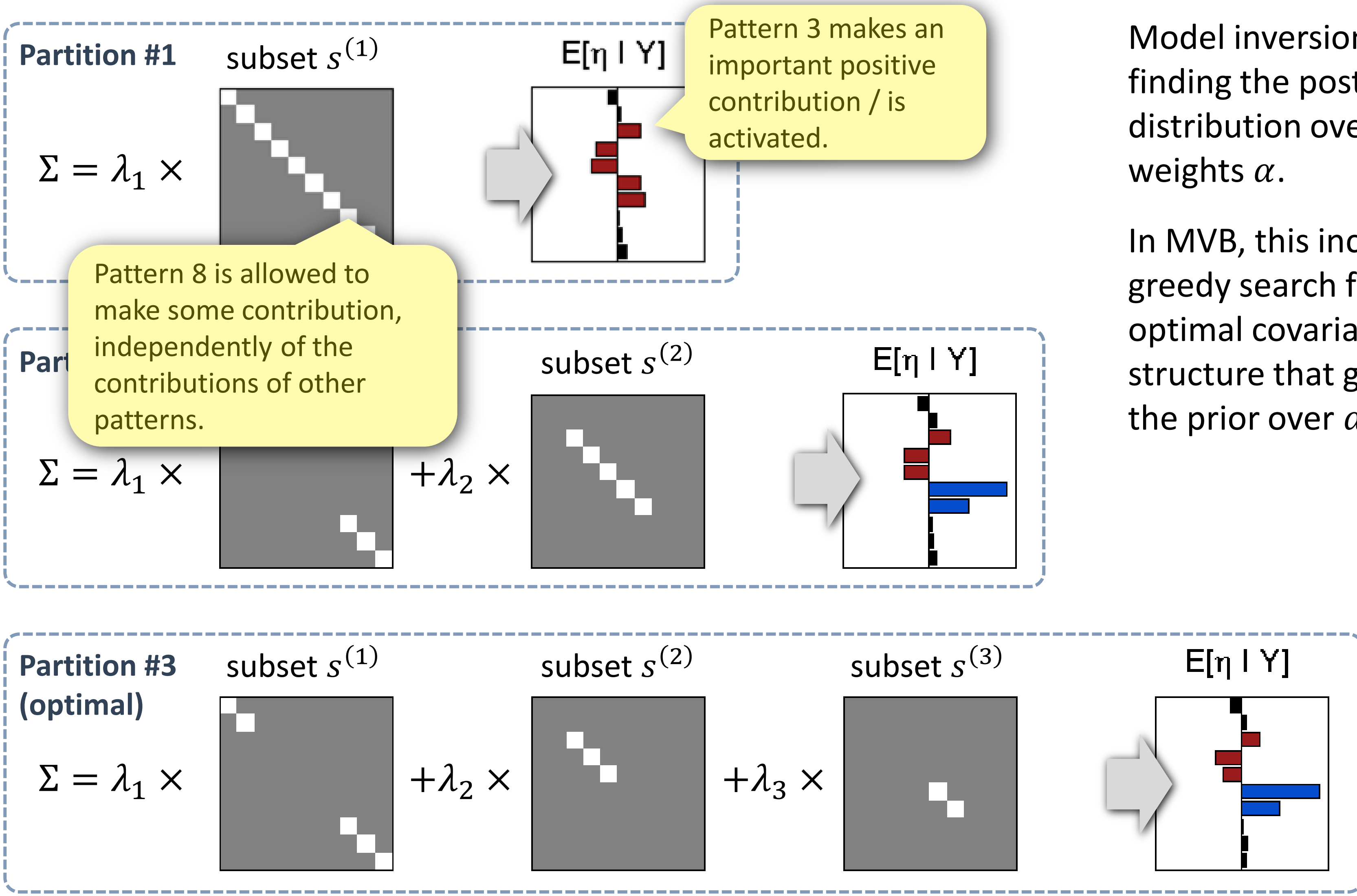
2nd level – pattern covariance structure Σ

$$p(\eta) = \mathcal{N}(\eta|0, \Sigma)$$

$$\Sigma = \sum_i \lambda_i s^{(i)}$$

Thus: $p(\alpha|\lambda) = \mathcal{N}_n(\alpha|0, U\Sigma U^T)$ and $p(\lambda) = \mathcal{N}(\lambda|\pi, \Pi^{-1})$

Inverting the model



Model inversion involves finding the posterior distribution over voxel weights α.

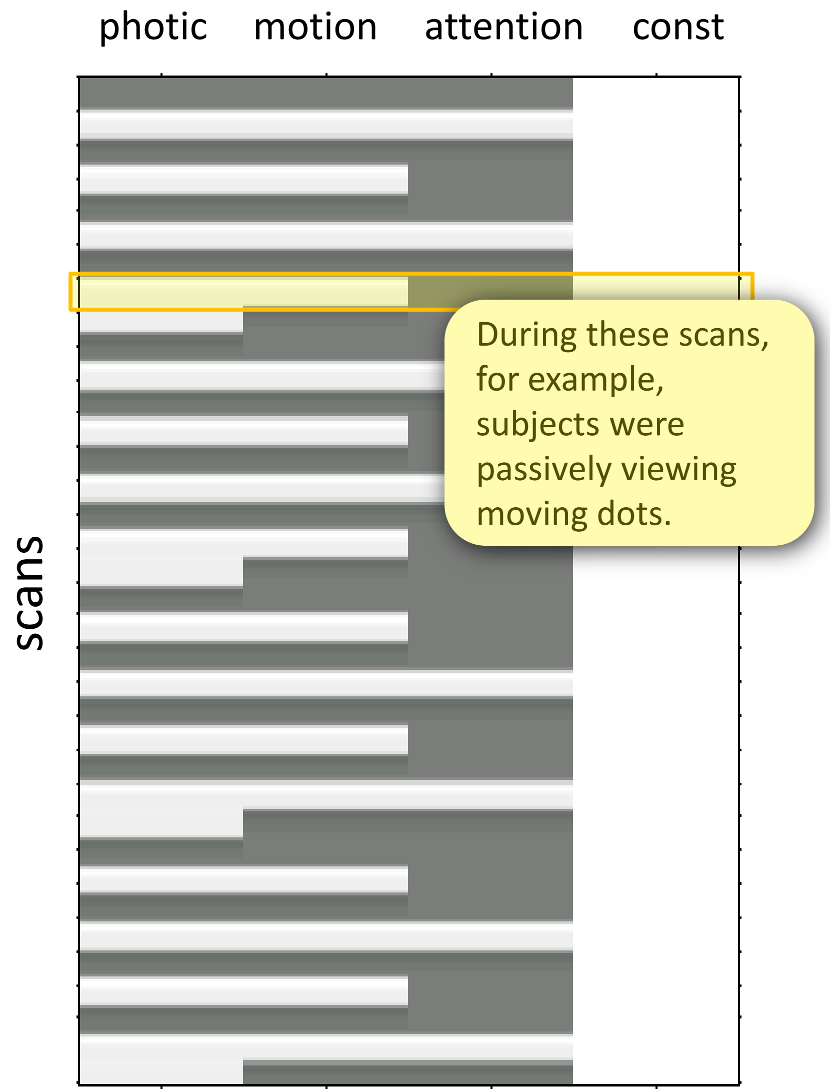
In MVB, this includes a greedy search for the optimal covariance structure that governs the prior over α.

Example: decoding motion from visual cortex

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

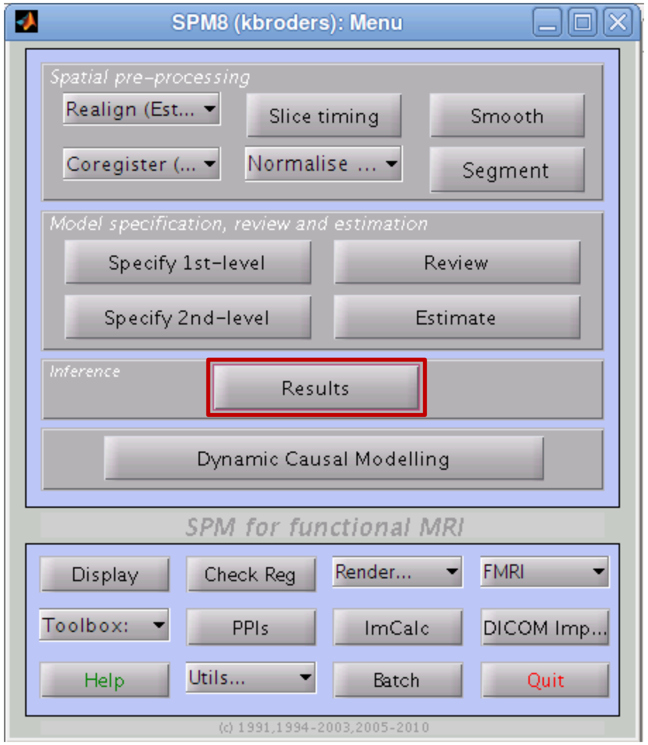
This dataset is based on a simple block design. There are three experimental factors:

- **photic** – display shows random dots
- **motion** – dots are moving
- **attention** – subjects asked to pay attention



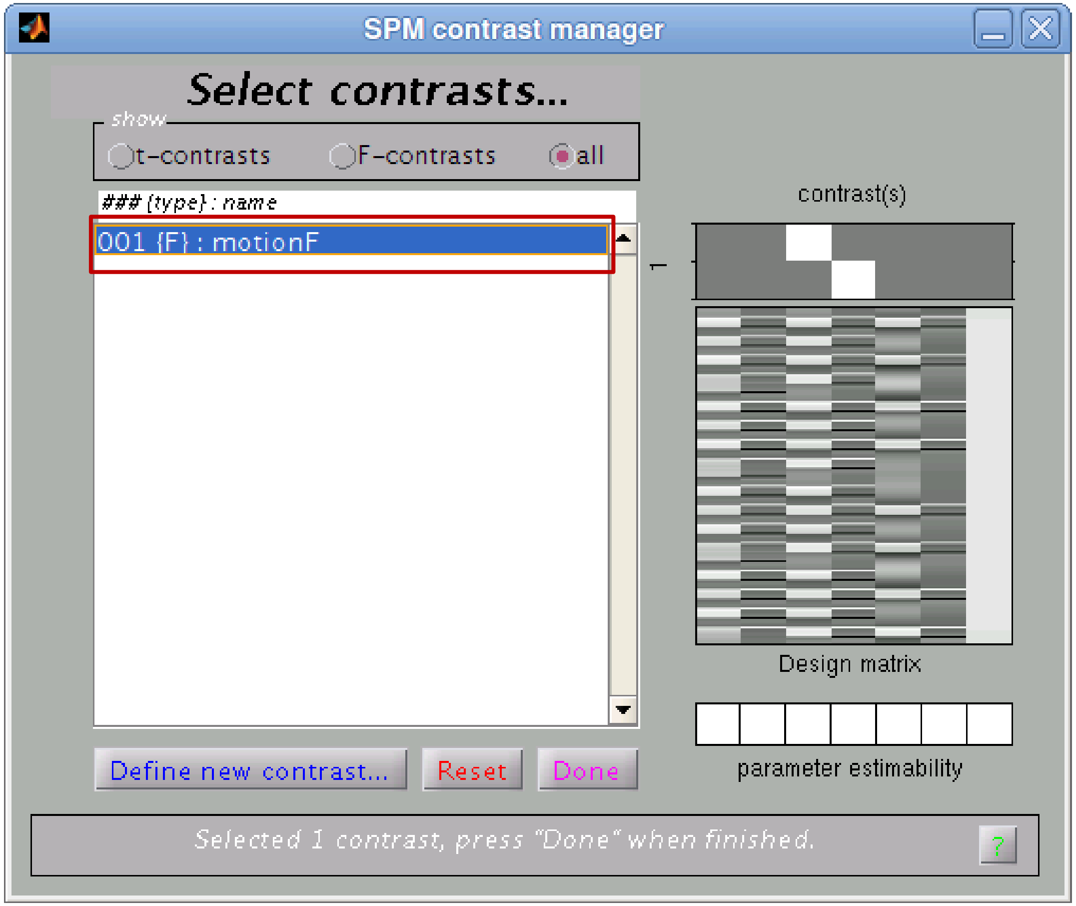
Buechel & Friston 1999 *Cerebral Cortex*
Friston et al. 2008 *NeuroImage*

Multivariate Bayes in SPM



Step 1

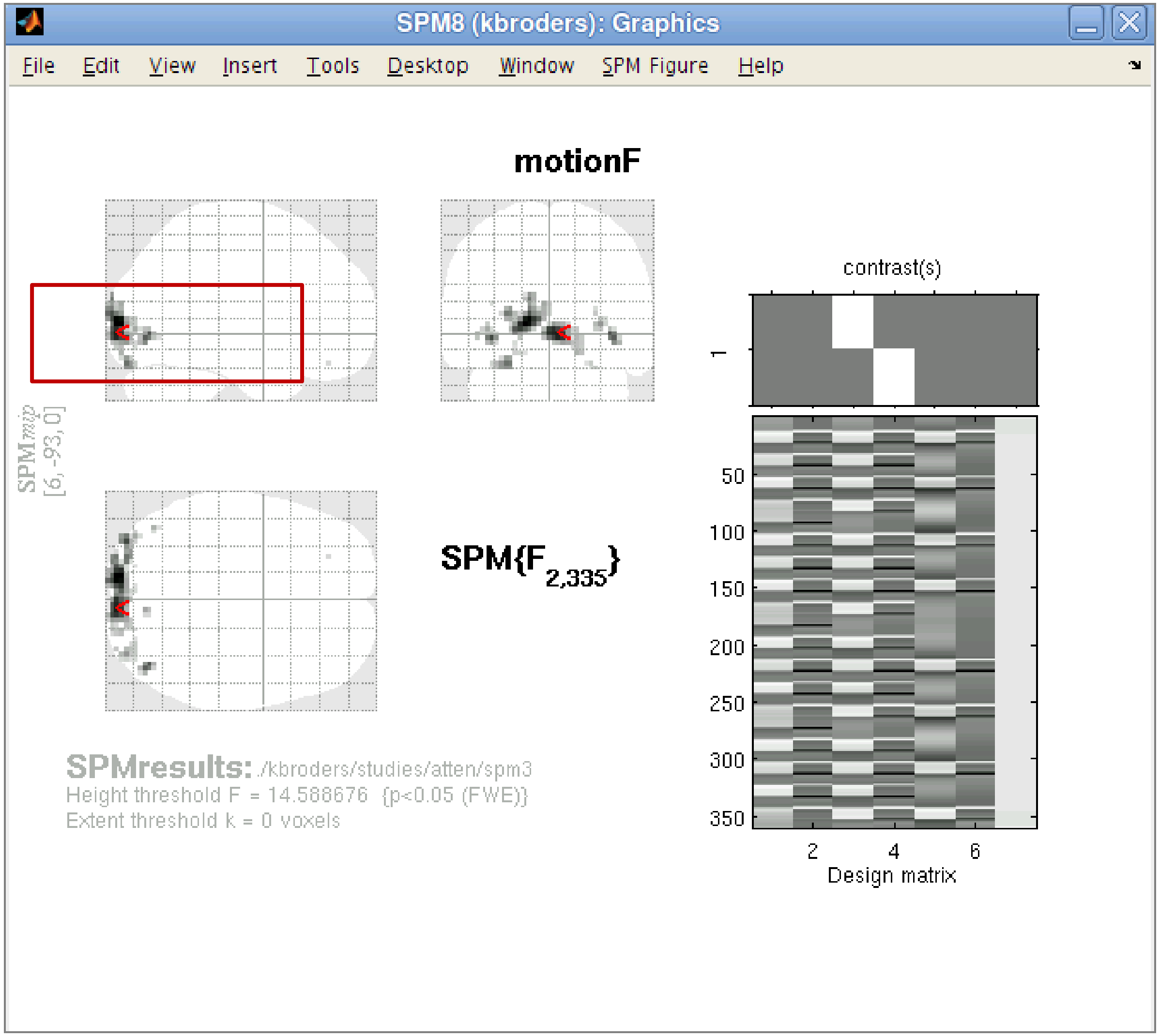
After having specified and estimated a model, use the *Results* button.



Step 2

Select the contrast to be decoded.

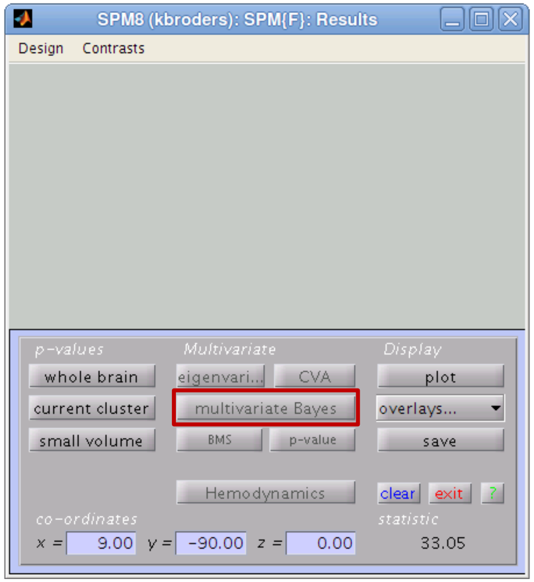
Multivariate Bayes in SPM



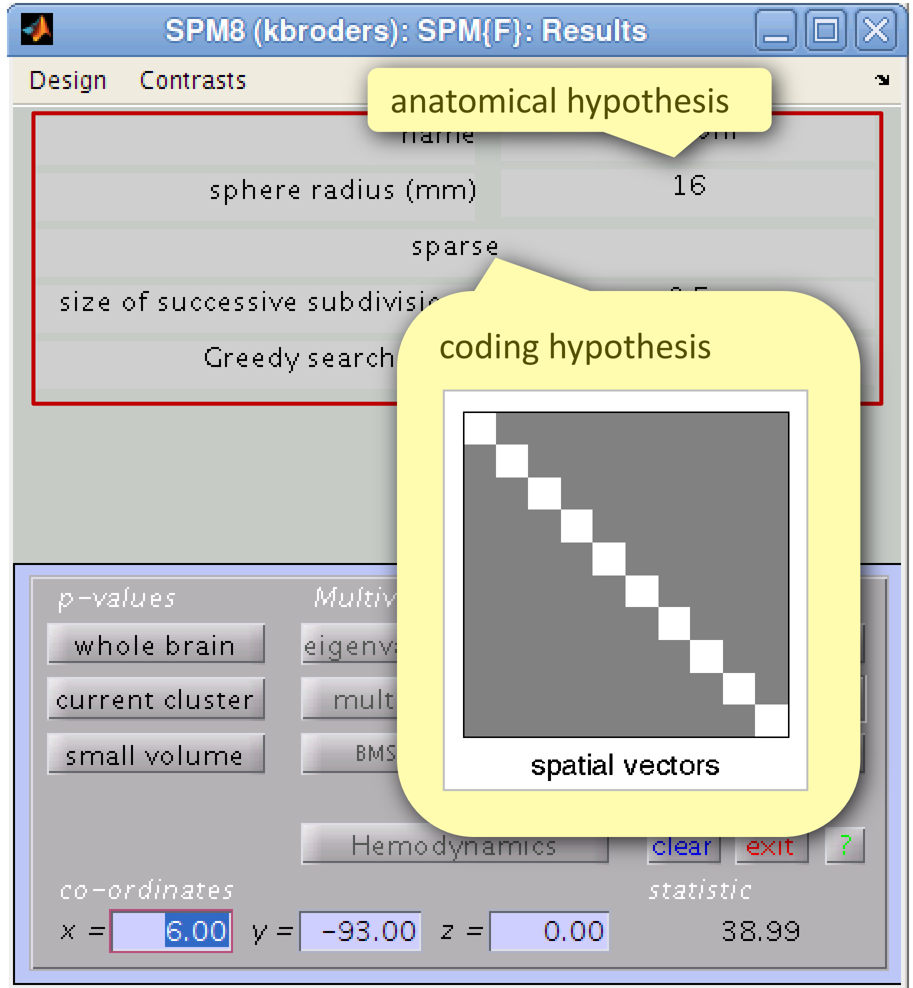
Step 3

Pick a region of interest.

Multivariate Bayes in SPM

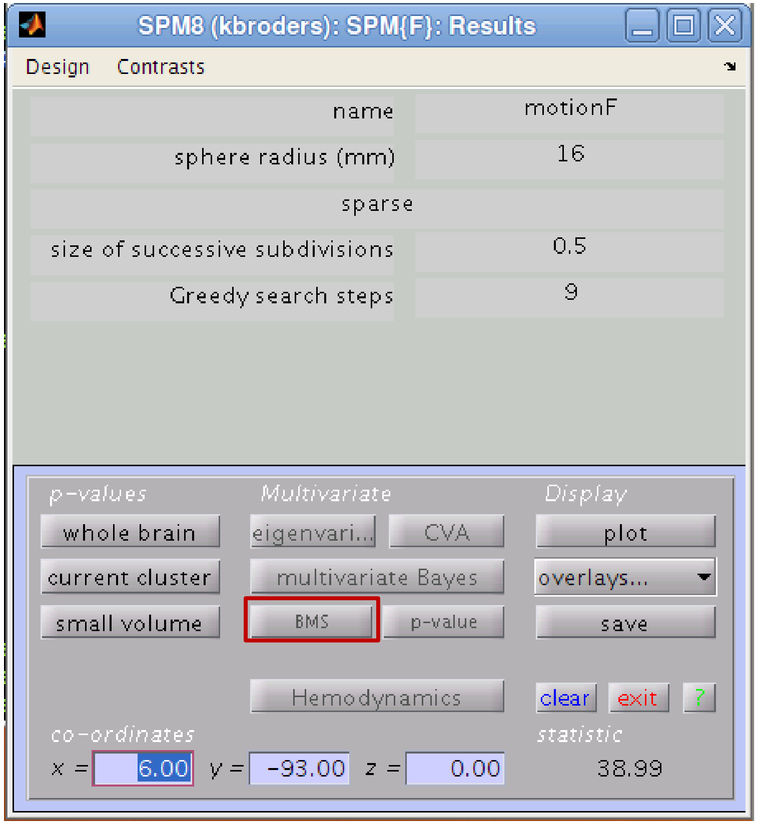


Step 4
Multivariate Bayes can be invoked from within the Multivariate section.



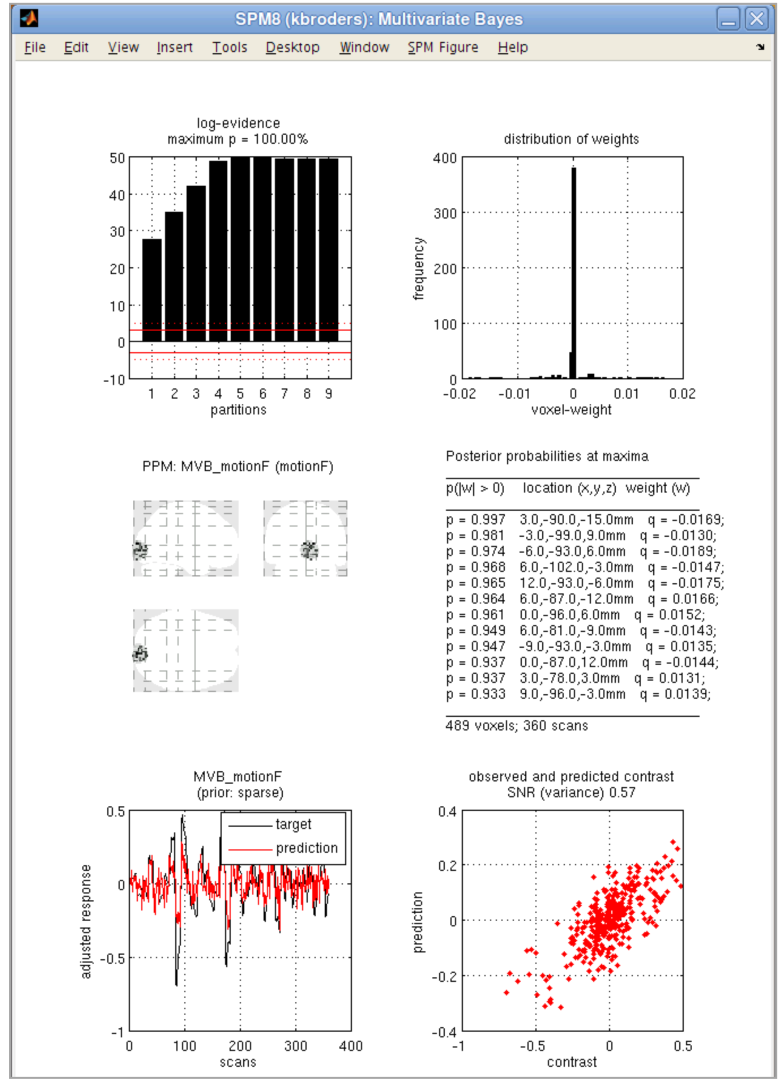
Step 5
Here, the region of interest is specified as a sphere around the cursor. The spatial prior implements a *sparse* coding hypothesis.

Multivariate Bayes in SPM

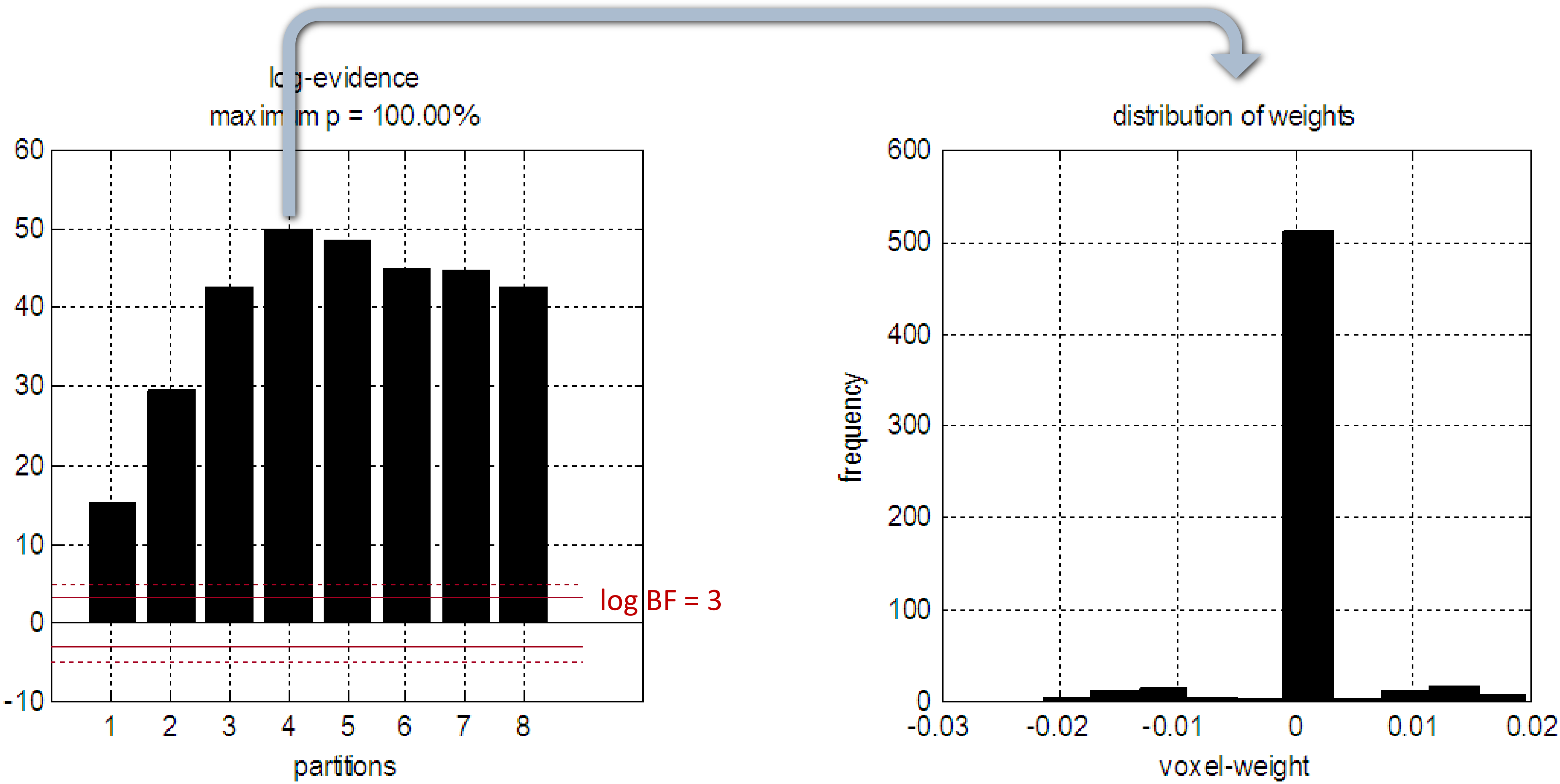


Step 6

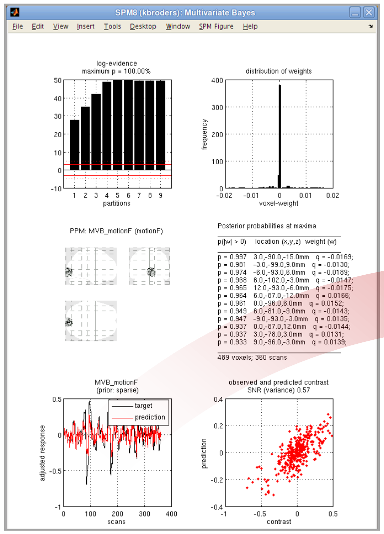
Results can be displayed using the BMS button.

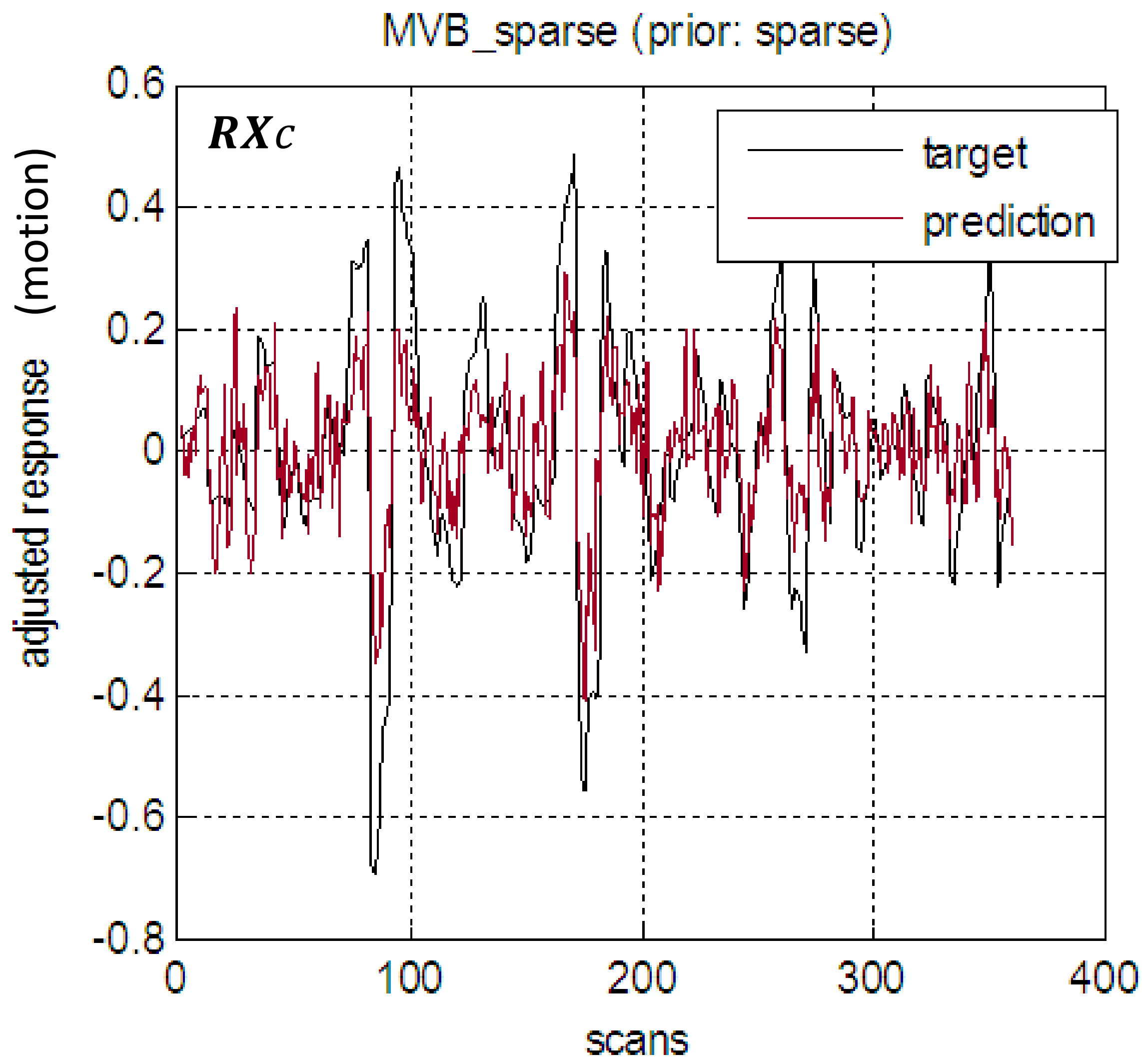


Model evidence and voxel weights

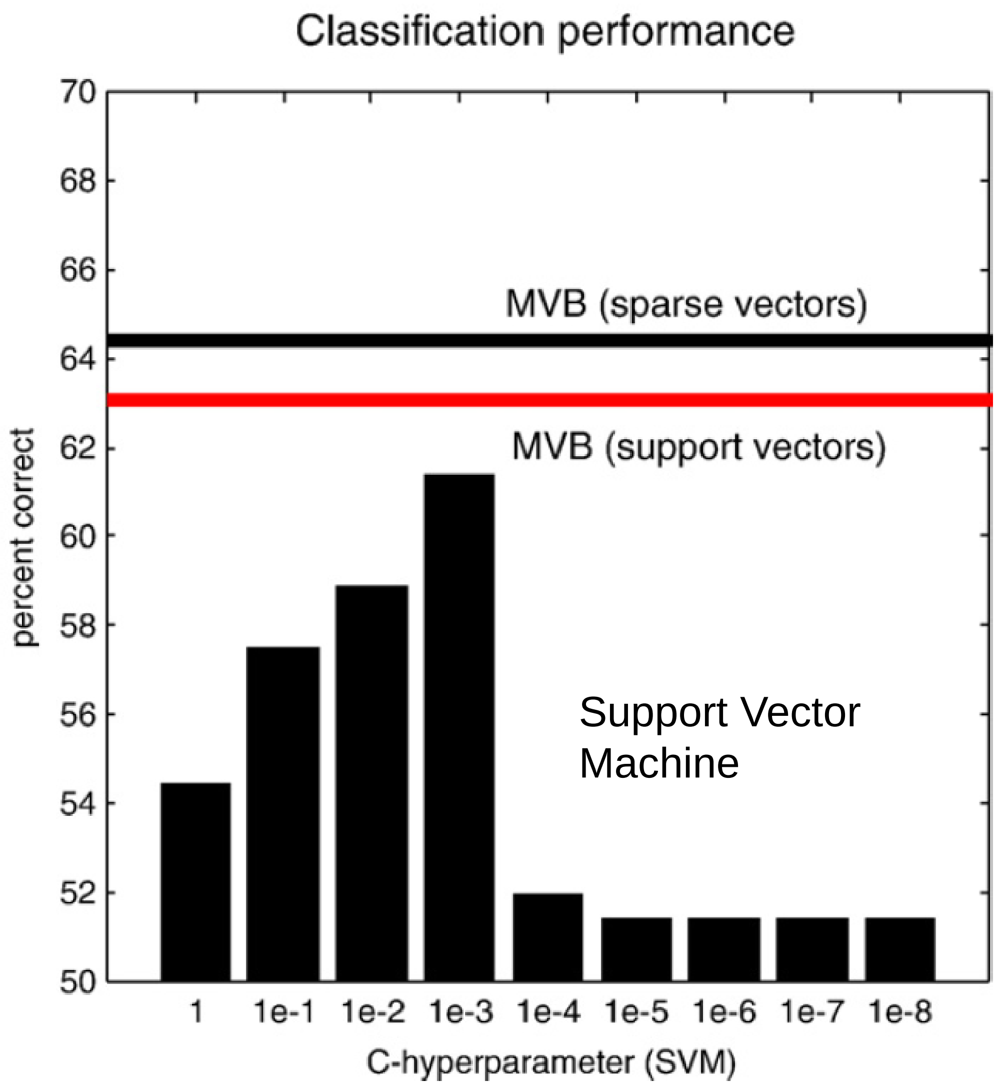


Observations vs. predictions





Using MVB for point classification

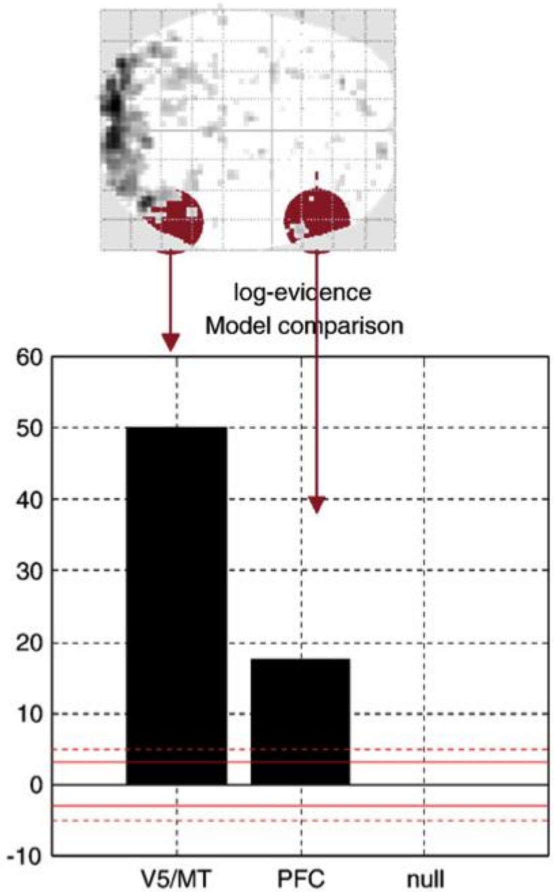


MVB may outperform conventional point classifiers when using a more appropriate coding hypothesis.

Summary: research questions for MVB

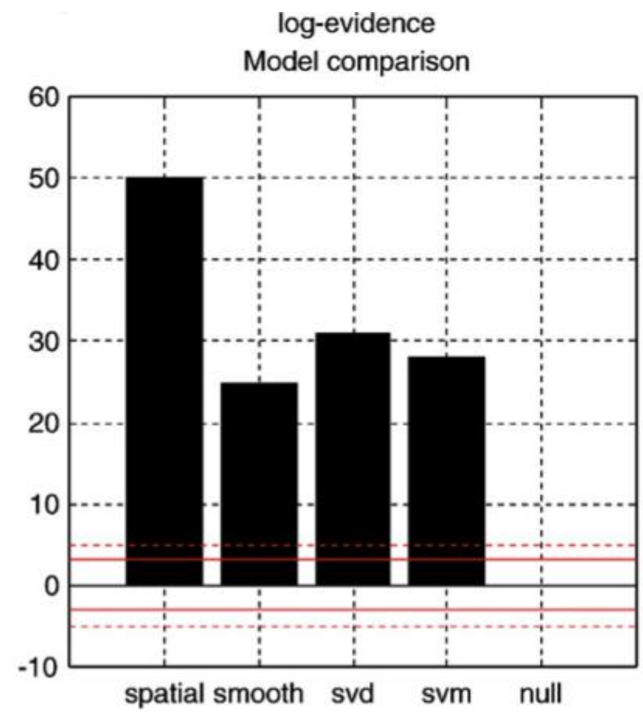
Where does the brain represent things?

Evaluating competing anatomical hypotheses



How does the brain represent things?

Evaluating competing coding hypotheses



Recent MVB studies

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Decoding episodic memory in ageing: A Bayesian analysis of activity patterns predicting memory

Alexa M. Morcom [a,b,c,*], Karl J. Friston [c]

[a] Psychology and Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK
[b] Centre for Cognitive and Neural Systems, University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK
[c] The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London, WC1N 3BG, UK

The Journal of Neuroscience, November 14, 2012 • 32(46):16417–16423 • **16417**

Behavioral/Systems/Cognitive

Action-Specific Value Signals in Reward-Related Regions of the Human Brain

Thomas H. B. FitzGerald, Karl J. Friston, and Raymond J. Dolan
Wellcome Trust Centre for Neuroimaging, London WC1N 3BG, United Kingdom

Overview

1 Modelling principles

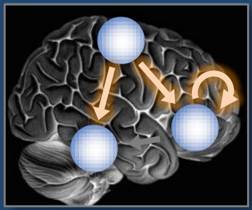
2 Classification

3 Multivariate Bayes

4 Generative embedding

Model-based analyses by data representation

Model-based analyses



How do patterns of hidden quantities (e.g., connectivity among brain regions) differ between groups?

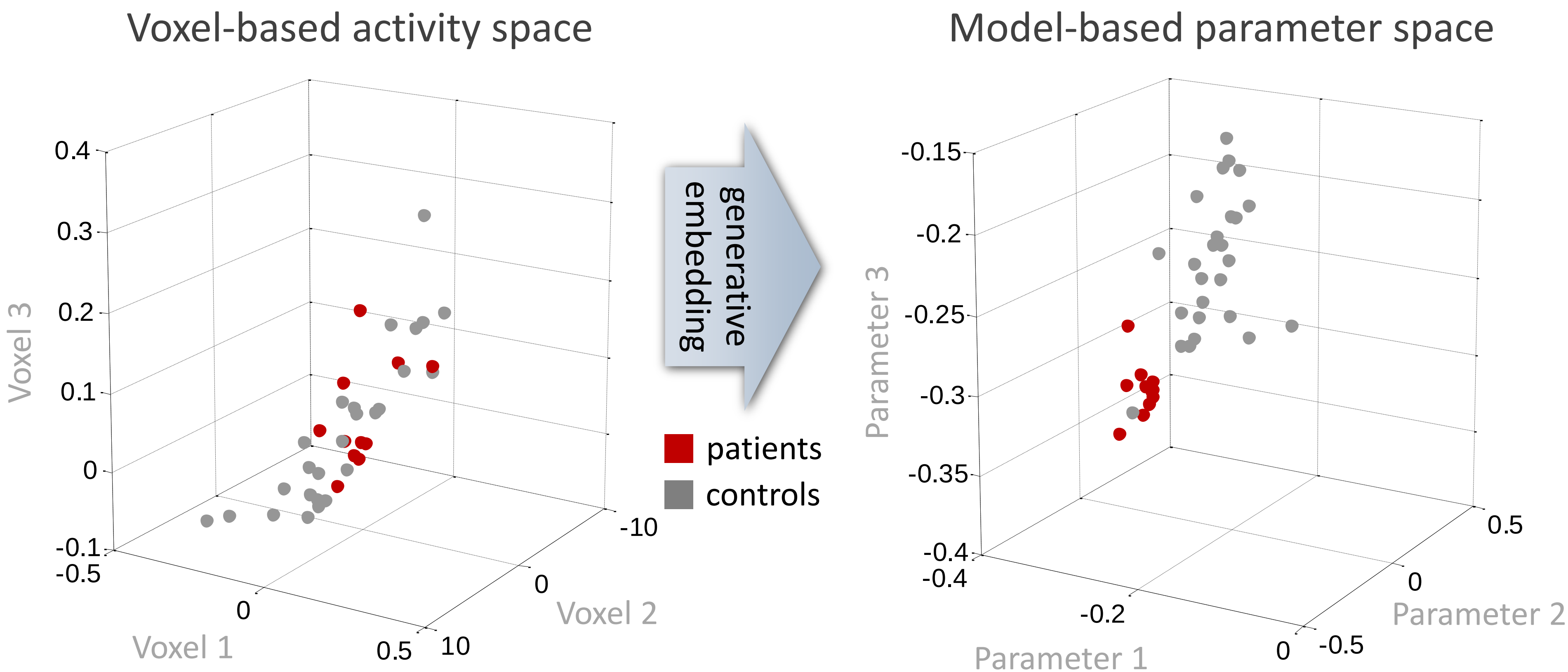
Activation-based analyses

Which functional differences allow us to separate groups?

Structure-based analyses

Which anatomical structures allow us to separate patients and healthy controls?

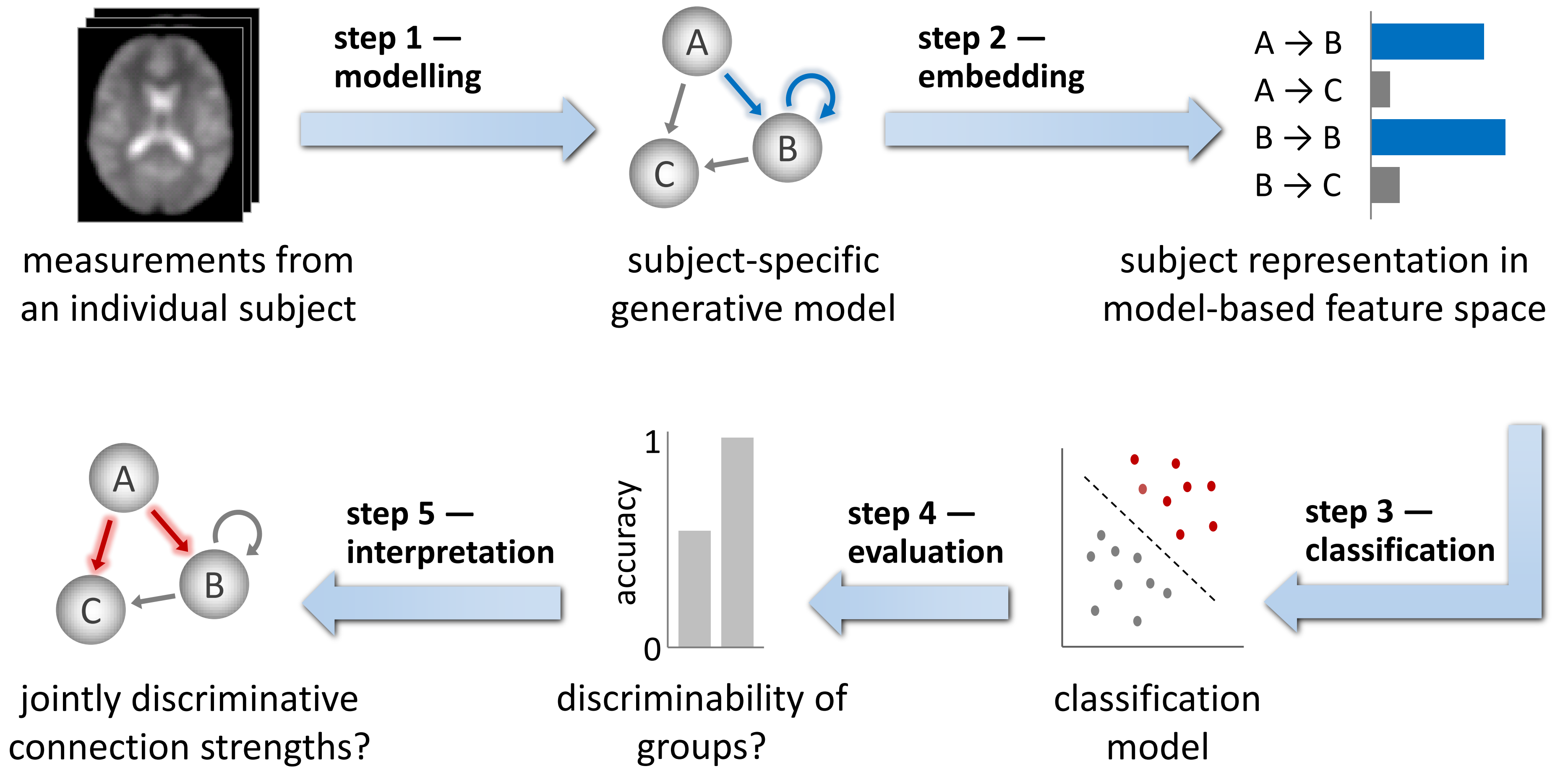
Generative embedding



classification accuracy
75%

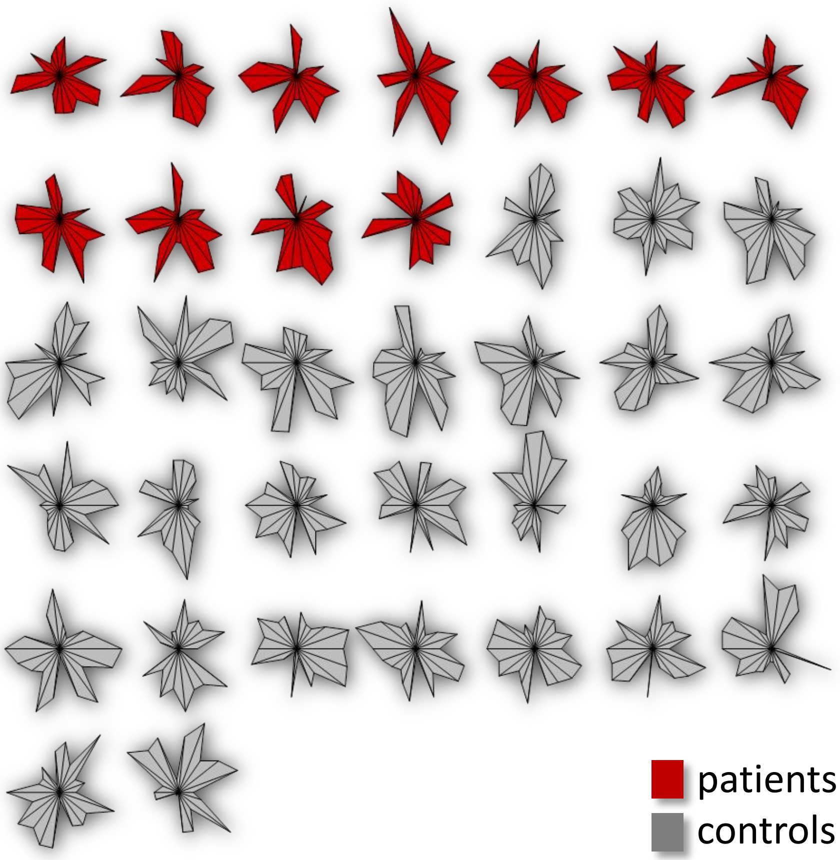
classification accuracy
98%

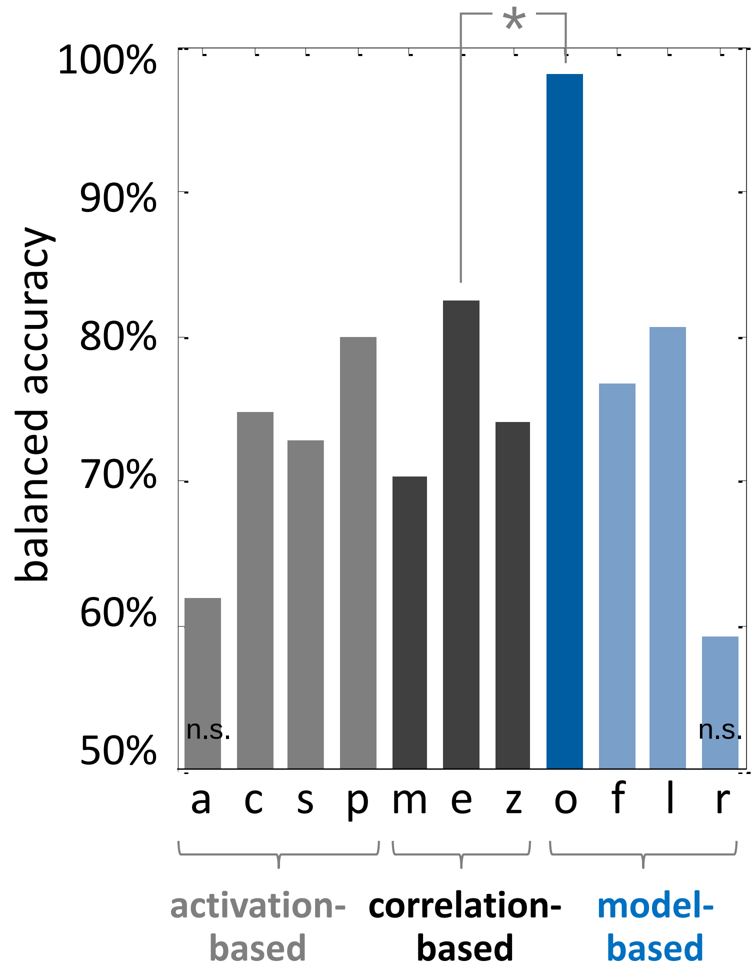
Model-based classification



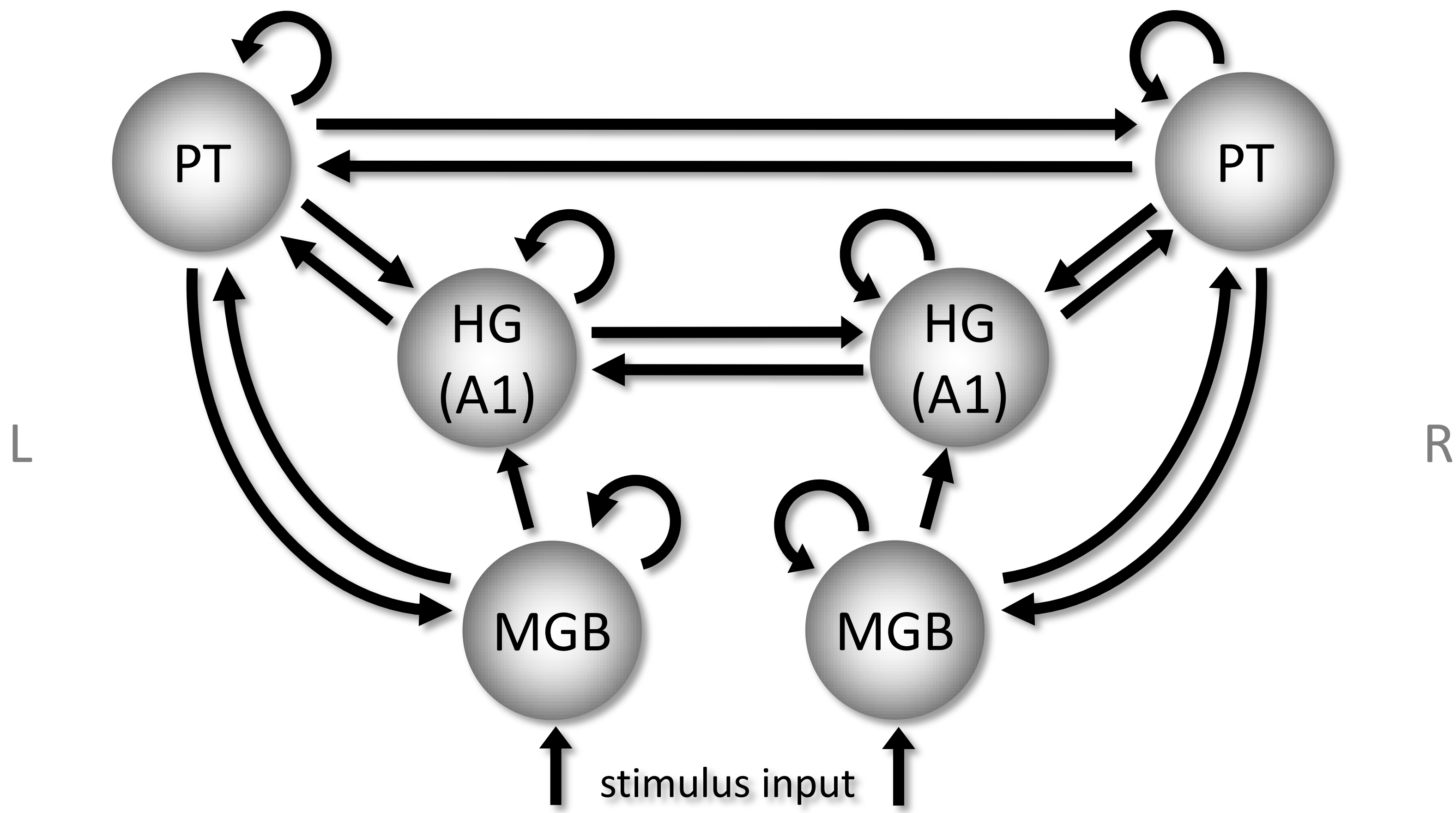
Brodersen, Haiss, Ong, Jung, Tittgemeyer, Buhmann, Weber, Stephan (2011) *NeuroImage*
Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann, Stephan (2011) *PLoS Comput Biol*

Model-based classification

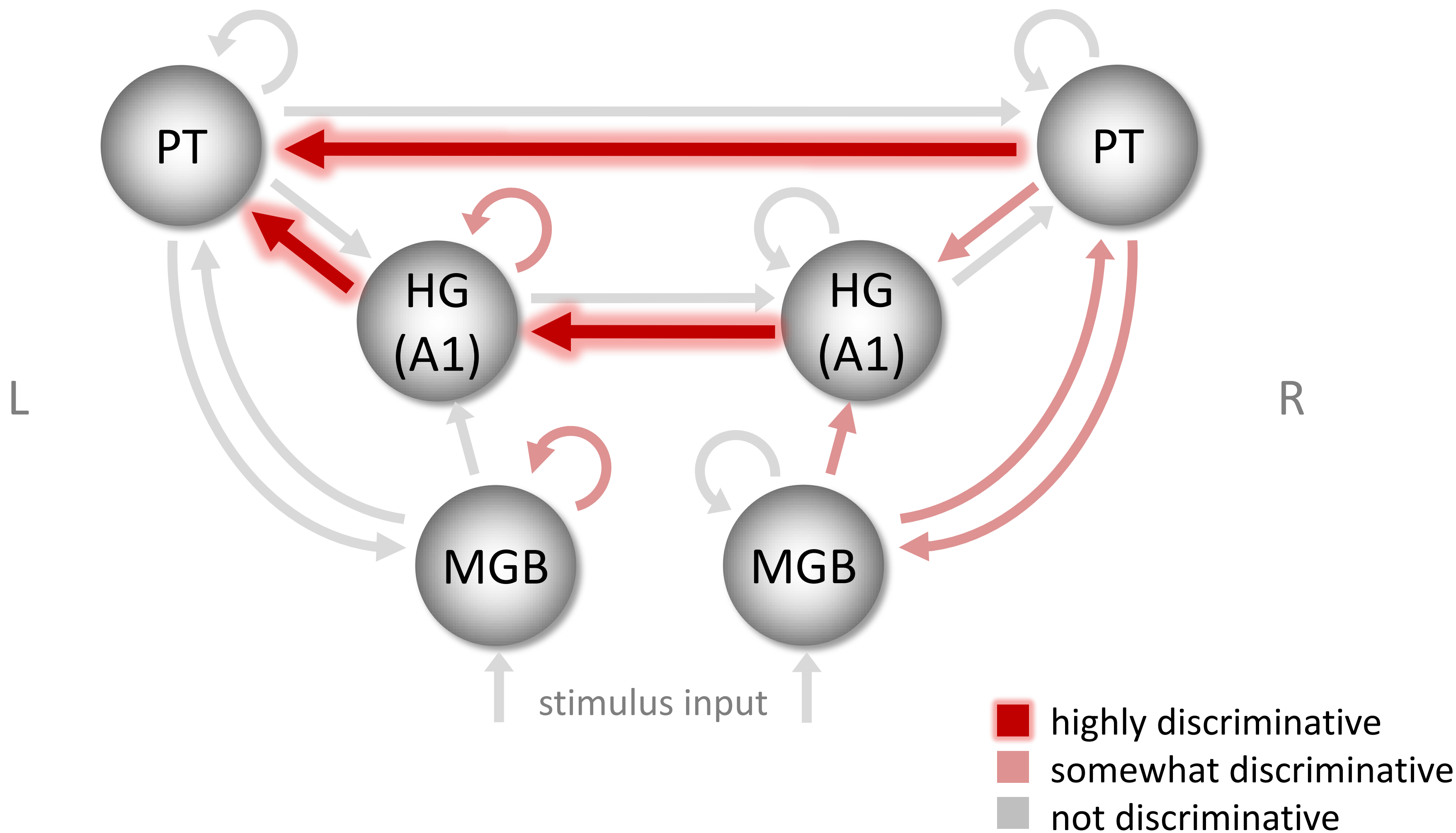




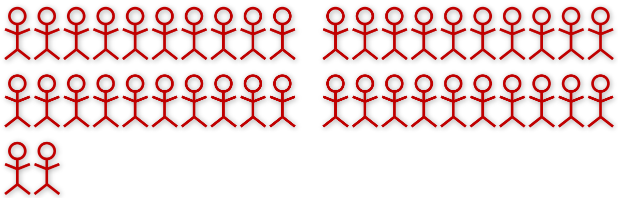
Model-based classification: interpretation



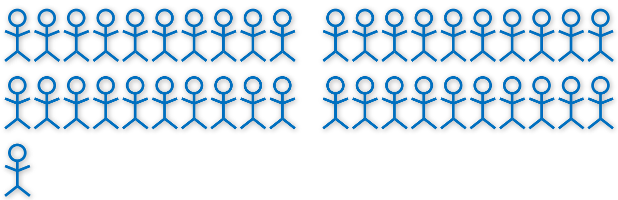
Model-based classification: interpretation



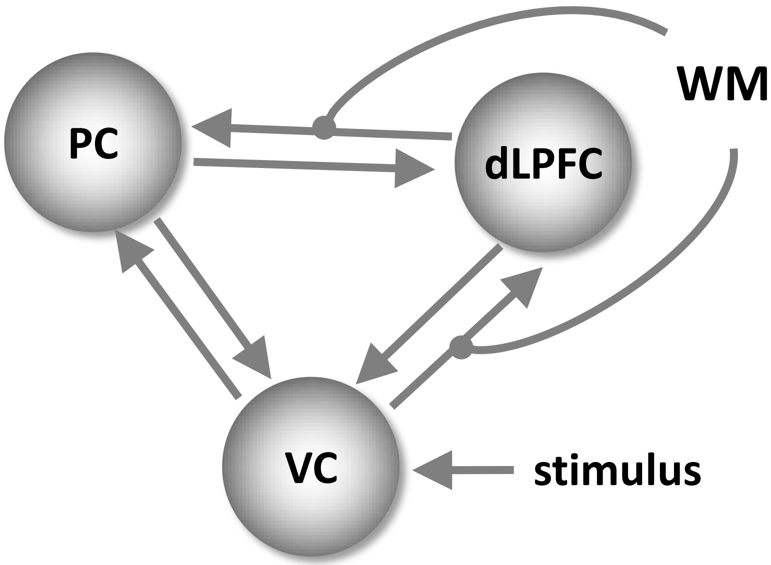
42 patients diagnosed with schizophrenia



41 healthy controls



fMRI data acquired during working-memory task & modelled using a three-region DCM

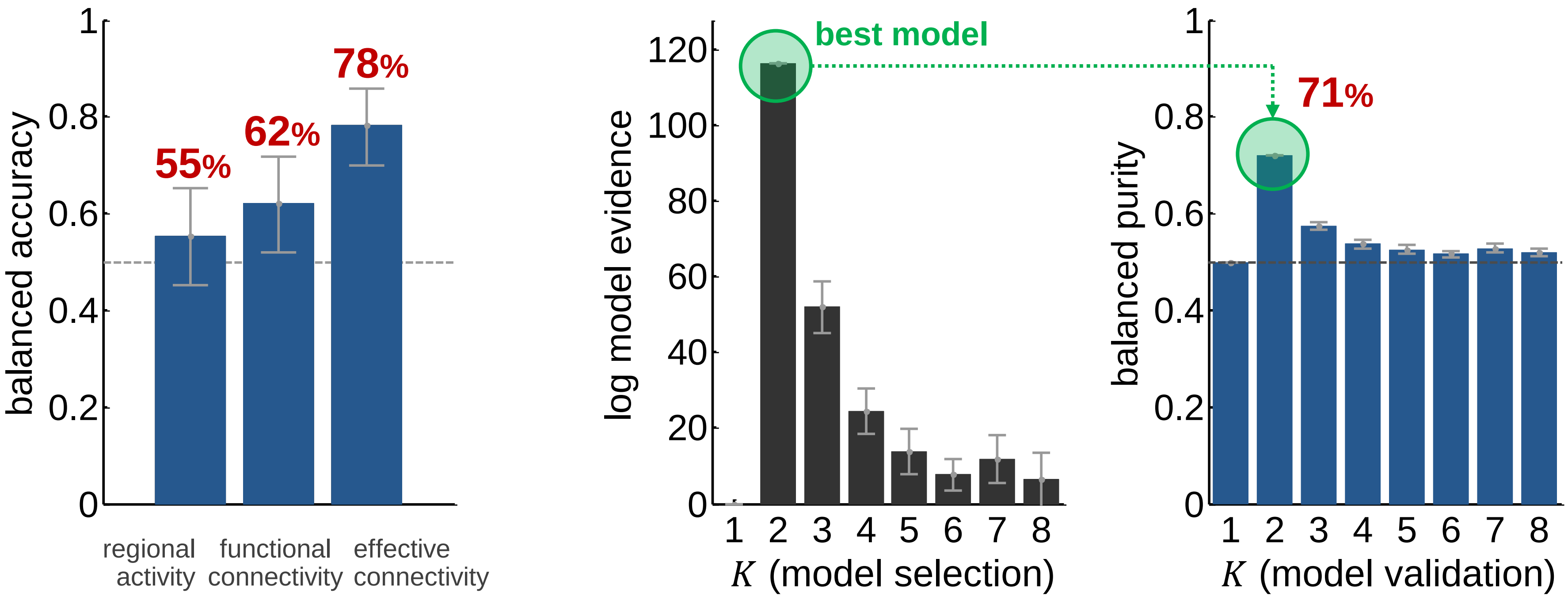


Deserno, Sterzer, Wüstenberg, Heinz, & Schlagenhauf (2012) *J Neurosci*

Model-based clustering

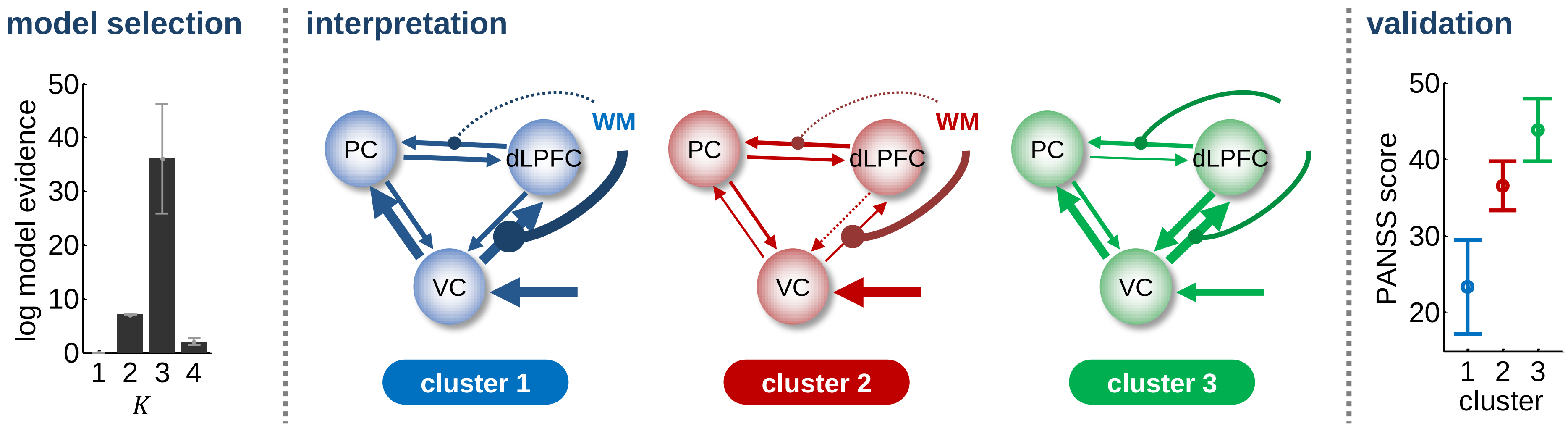
supervised learning:
SVM classification

unsupervised learning:
GMM clustering (using effective connectivity)



Brodersen, Deserno, Schlagenhauf, Penny, Lin, Gupta, Buhmann, Stephan *(in preparation)*

Model-based clustering



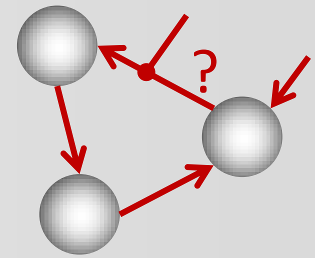
Brodersen et al. *(in preparation)*

Generative embedding and DCM

Question 1 – What do the data tell us about hidden processes in the brain?

⇨ compute the posterior

$$p(\theta|y,m) = \frac{p(y|\theta,m)p(\theta|m)}{p(y|m)}$$

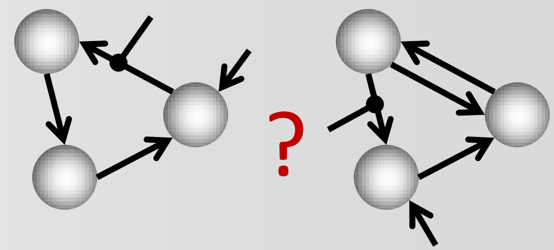


Question 2 – Which model is best w.r.t. the observed fMRI data?

⇨ compute the model evidence

$$p(m|y) \propto p(y|m)p(m)$$

$$= \int p(y|\theta,m)p(\theta|m)d\theta$$

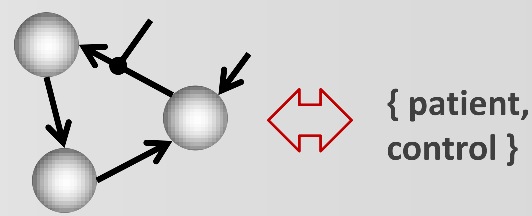


Question 3 – Which model is best w.r.t. an external criterion?

⇨ compute the classification accuracy

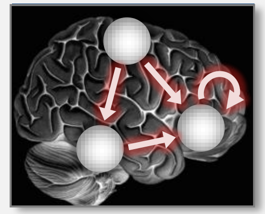
$$p(h(y) = x|y)$$

$$= \iiint p\big(h(y) = x|y, y_{\text{train}}, x_{\text{train}}\big)\, p(y)\, p\big(y_{\text{train}}\big)\, p\big(x_{\text{train}}\big)\ dy\ dy_{\text{train}}\ x_{\text{train}}$$



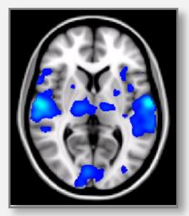
Model-based classification using DCM

model-based classification

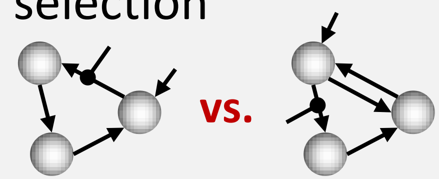


⇔ { group 1, group 2 }

activation-based classification

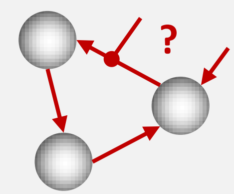


model selection

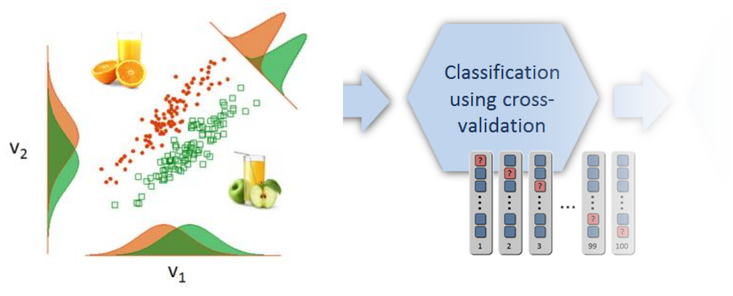


structure-based classification

inference on model parameters

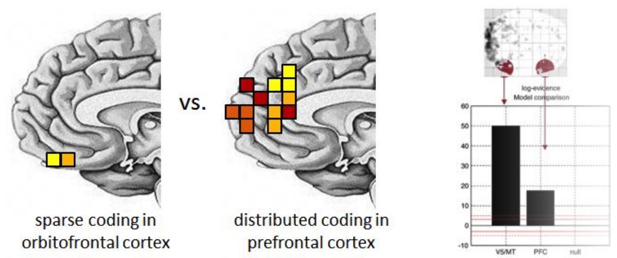


Summary



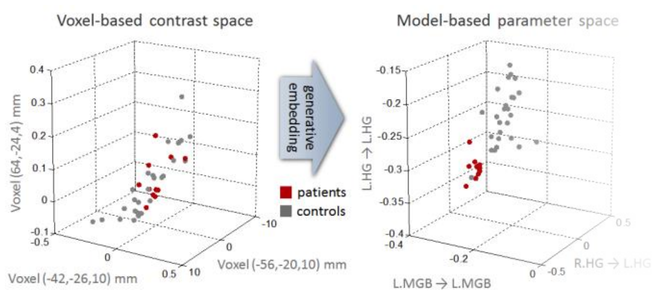
Classification

- to assess whether a cognitive state is linked to patterns of activity
- to visualize the spatial deployment of discriminative activity



Multivariate Bayes

- to evaluate competing anatomical hypotheses
- to evaluate competing coding hypotheses



Generative embedding

- to assess whether groups differ in terms of patterns of connectivity
- to generate mechanistic subgroup hypotheses