Multivariate models and machine learning for fMRI

Methods and Models in fMRI, 15.11.2016

Jakob Heinzle
heinzle@biomed.ee.ethz.ch

Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering (IBT)
University and ETH Zürich

Many thanks to Sudhir Raman and Kay Brodersen for material

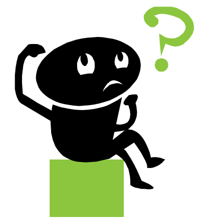
Translational Neuromodeling Unit

University of Zurich UZH

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Overview

Motivation

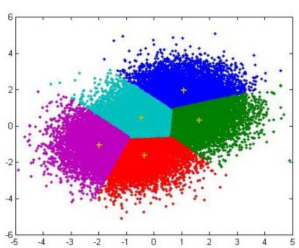


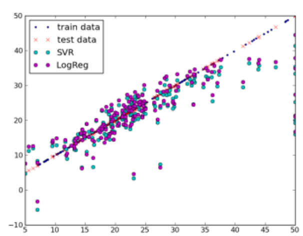
Modelling Terminology

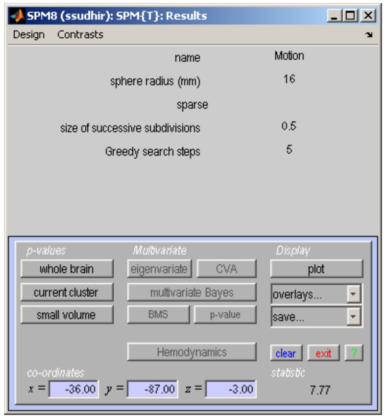
Learning from data

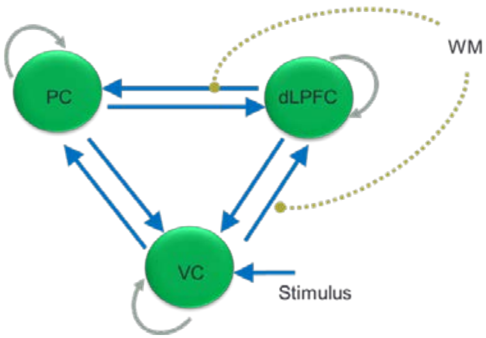
Multivariate Bayes in SPM

Generative Embedding



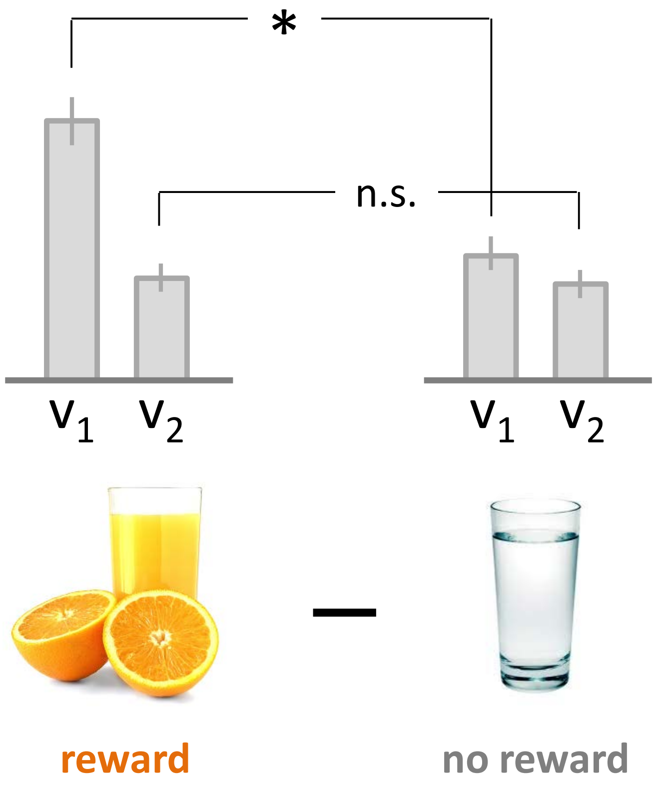






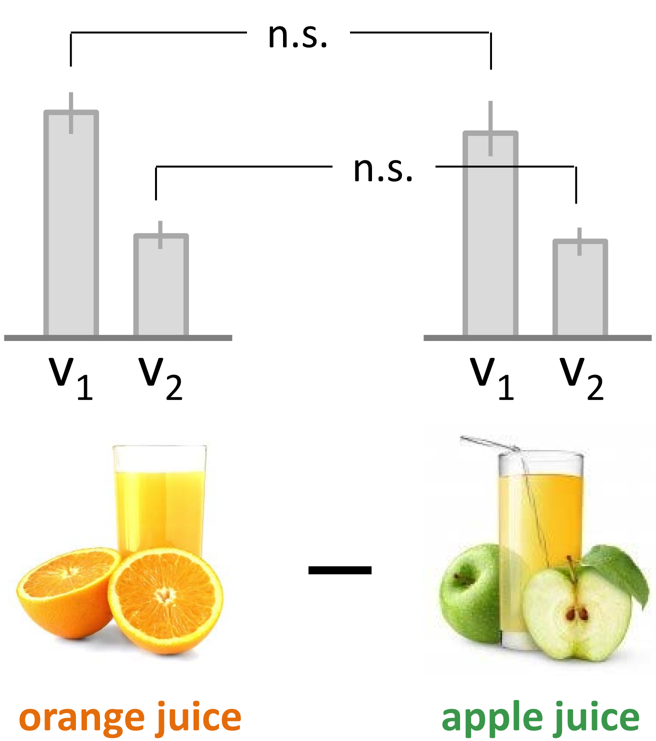
Why multivariate?

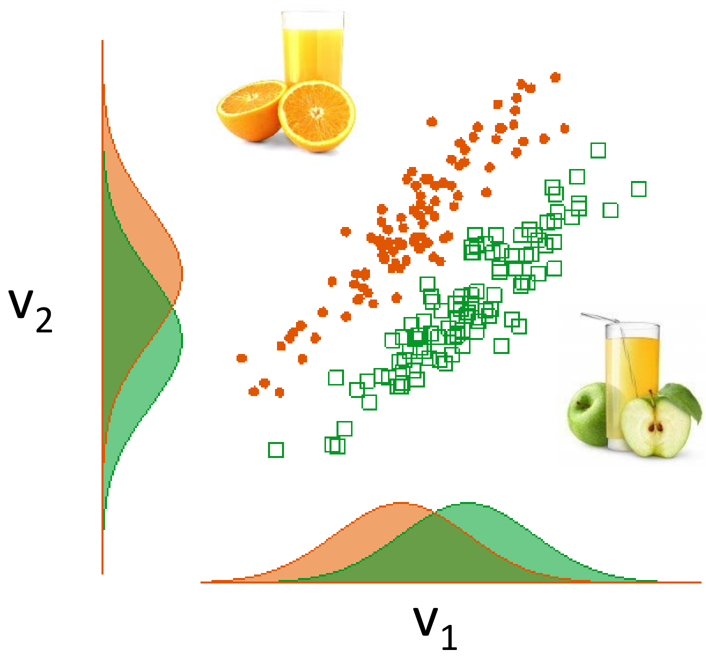
Univariate approaches are excellent for localizing activations in individual voxels.



Why multivariate?

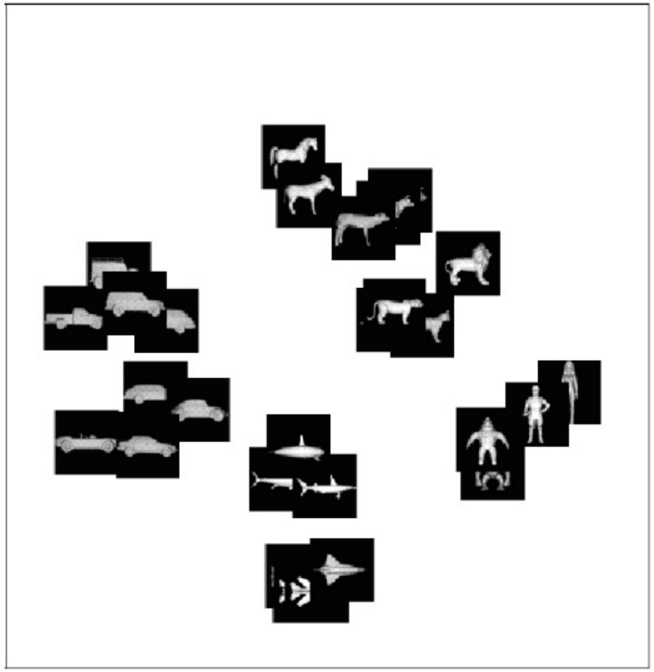
Multivariate approaches can be used to examine responses that are jointly encoded in multiple voxels.



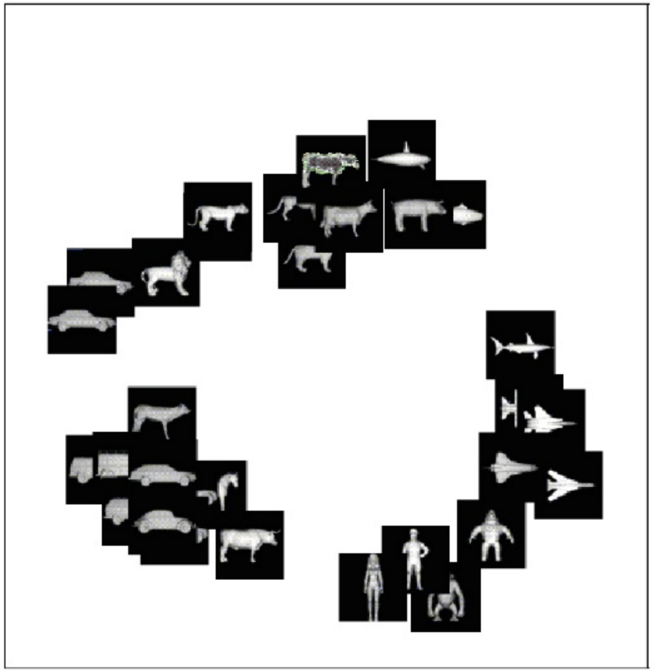


A bit of history – Multidymensional scaling

Psychophysical rating



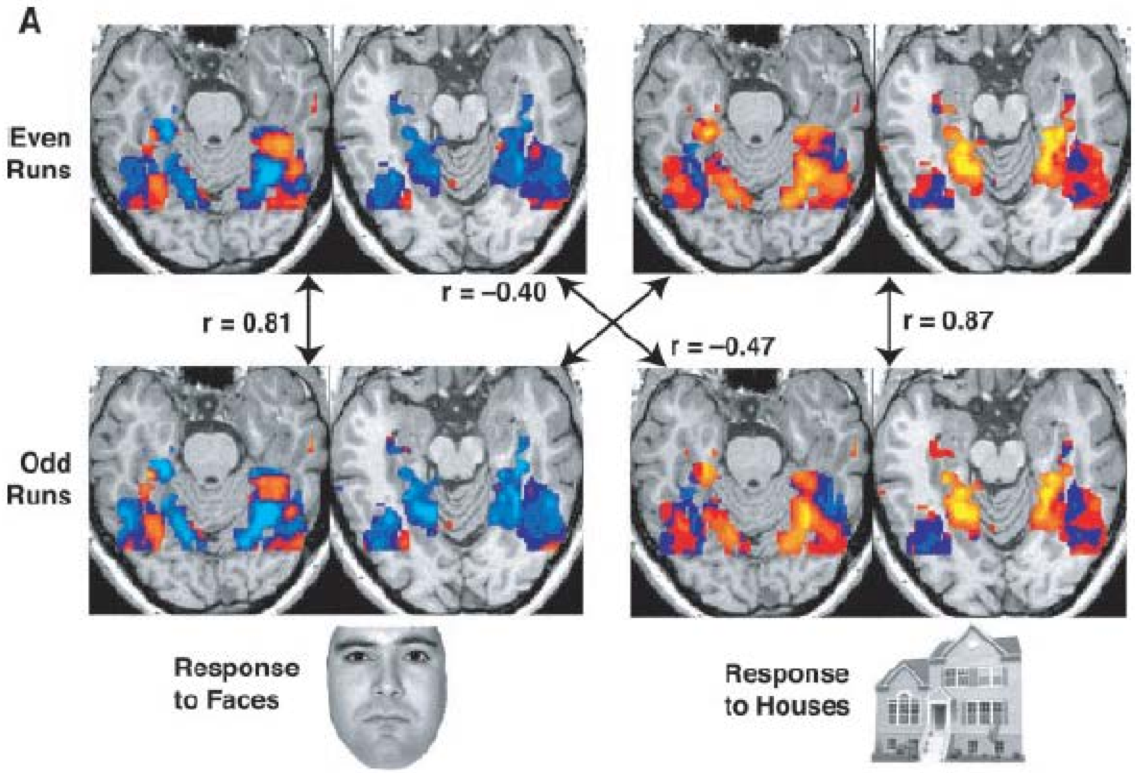
fMRI



Two-dimensional projection of similarity measure for both psychophysical rating and fMRI response.

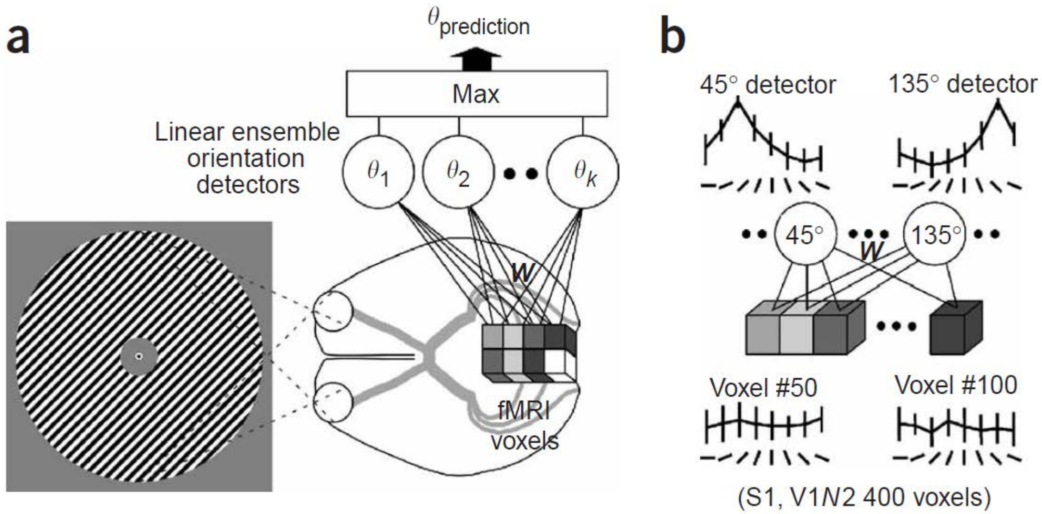
Edelman et al, Psychobiology, 1998

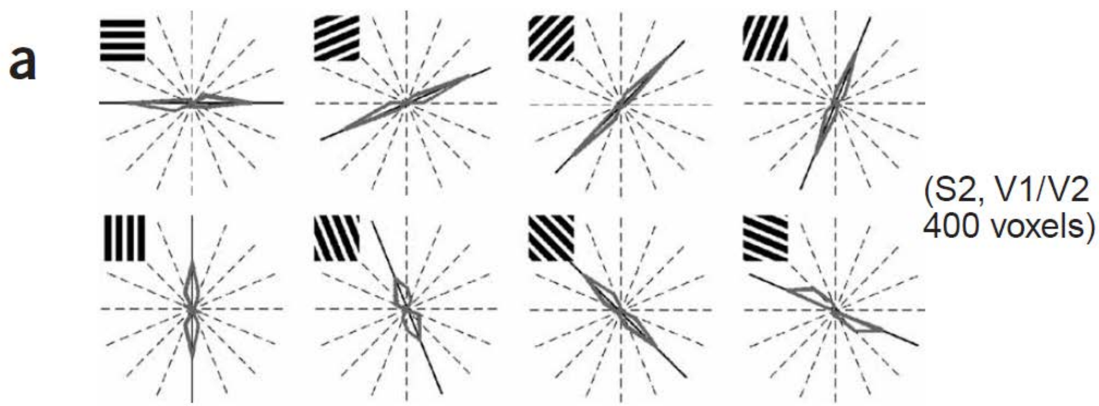
A bit of history – Classification Studies



Haxby et al, Science, 2001

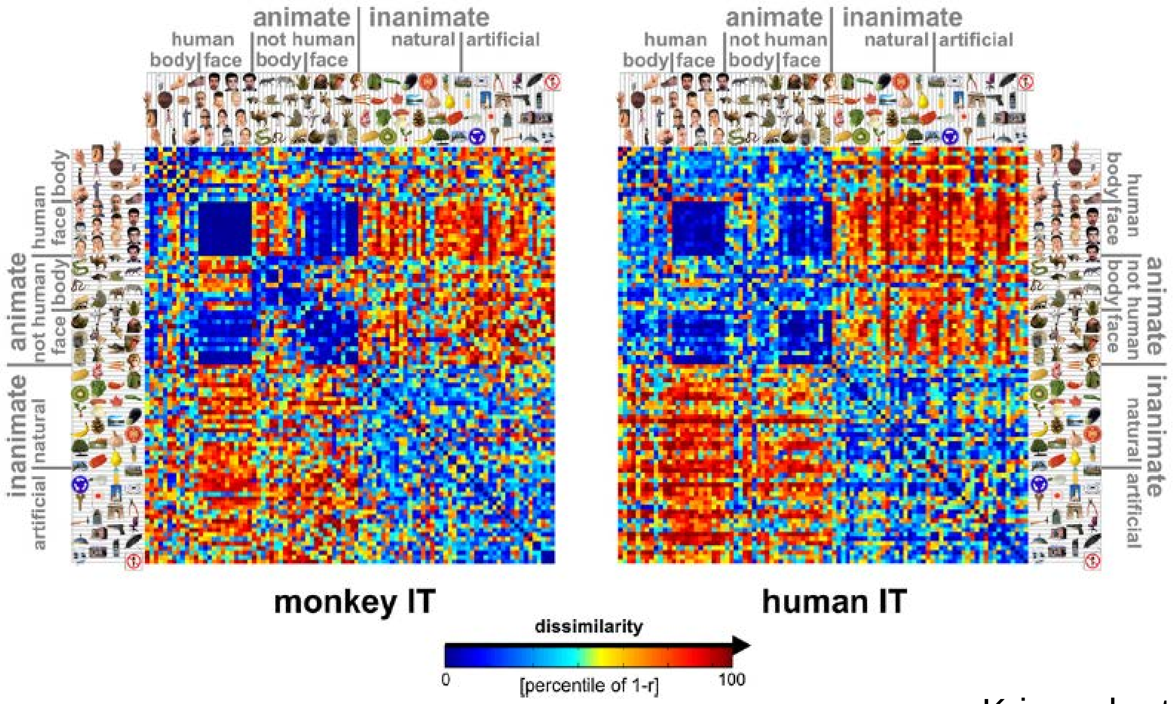
A bit of history – Classification Studies





Kamitani and Tong, Nat Neurosci, 2005

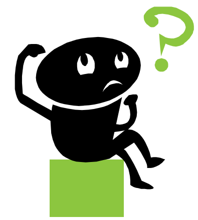
Idea: Compare the similarity of representations (correlation between activation patterns) between different stimuli.
Allows for a comparison between monkey
(neural firing pattern) and human (fMRI activation patterns).



Kriegeskorte et al, Neuron, 2008

Overview

Motivation

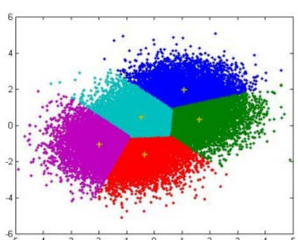


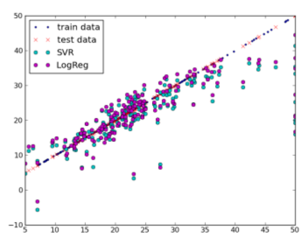
Modelling Terminology

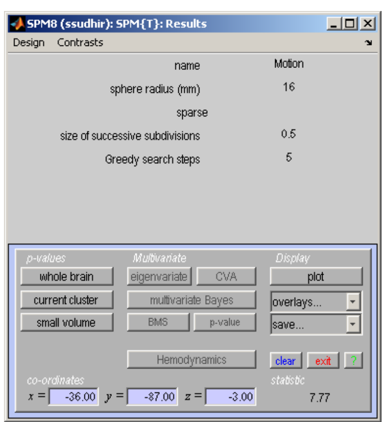
Learning from data

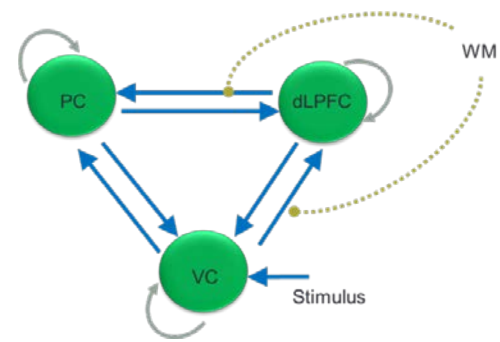
Multivariate Bayes in SPM

Generative Embedding

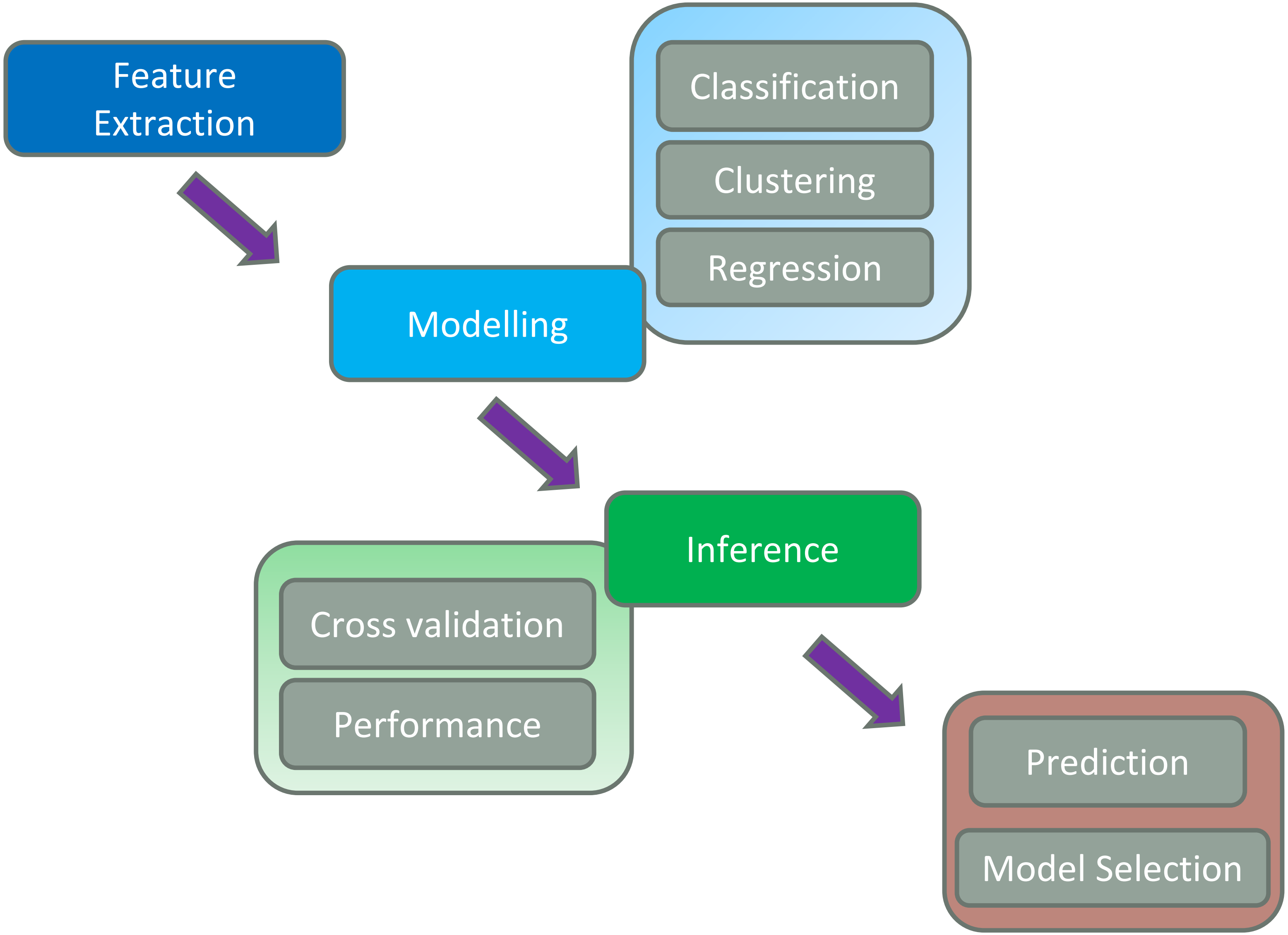








Analysis steps



Feature space

Features

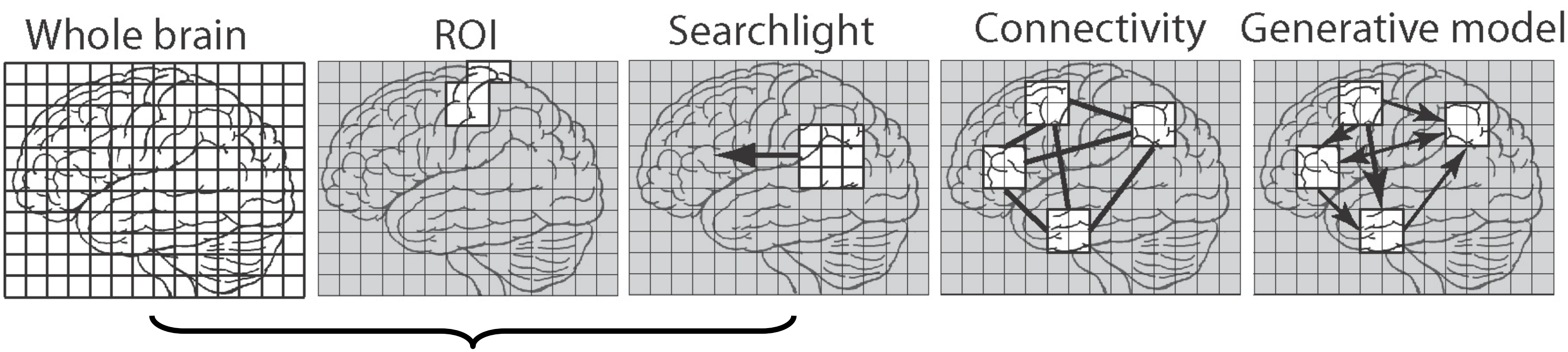
	F_1	F_2	.	.	.	F_P
S_1	1	0.5				
S_2	0	5.7				
.	1	4				
.	1	5.3				
S_N	1	6.6				

Data Points

- Discrete
- Continuous

Feature selection for fMRI multivariate analysis

Different features answer different questions.
Reducing the dimensionality might reduce noise,
but could also reduce relevant information.

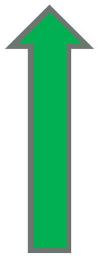


Model parameters
Mean values
Raw data

Correlations between regions

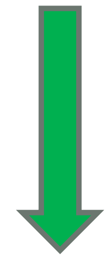
Model Parameters, e.g. DCM

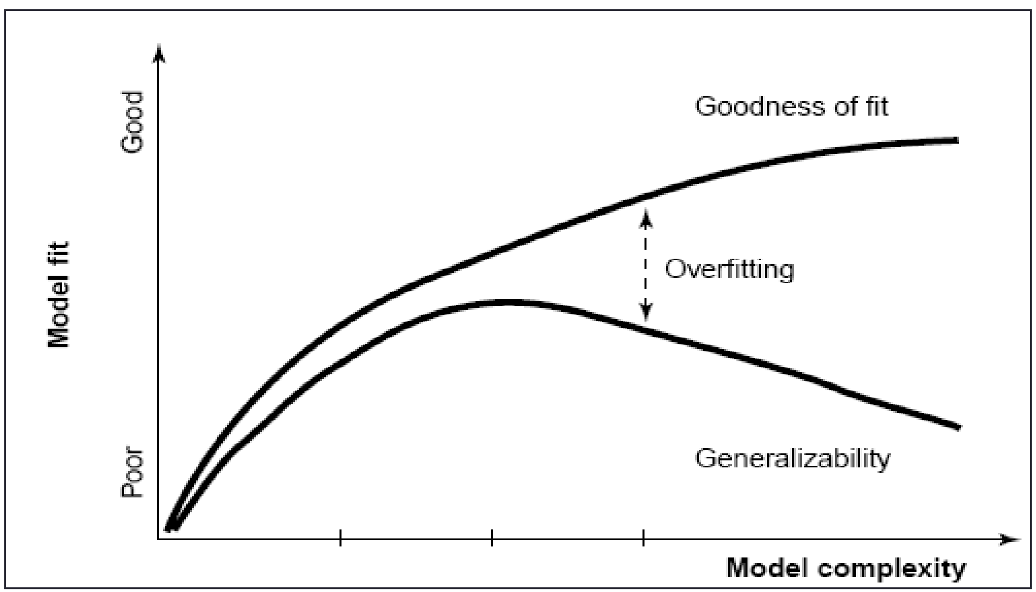
Model selection - Generalizability

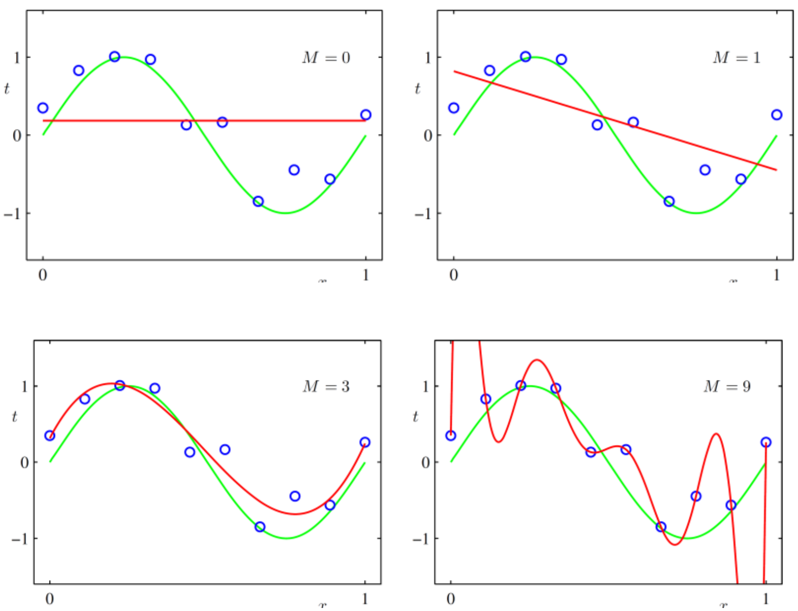


Model
Fit

Model
Complexity







Bishop (2006), Pitt & Miyung (2002), *TICS*

Encoding and decoding models

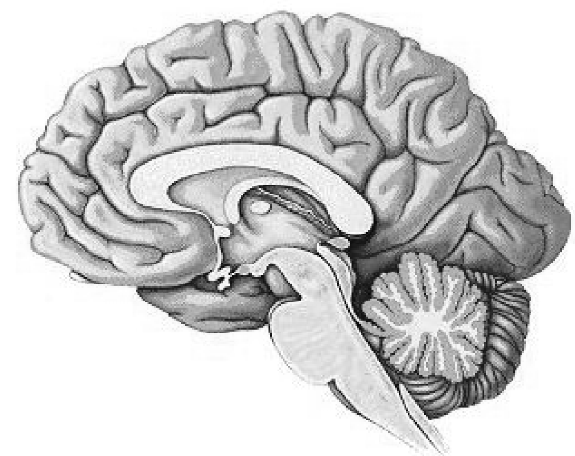
condition
stimulus
response
prediction error

encoding model

$g: X_t \rightarrow Y_t$

decoding model

$h: Y_t \rightarrow X_t$



context (cause or consequence)

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

BOLD signal

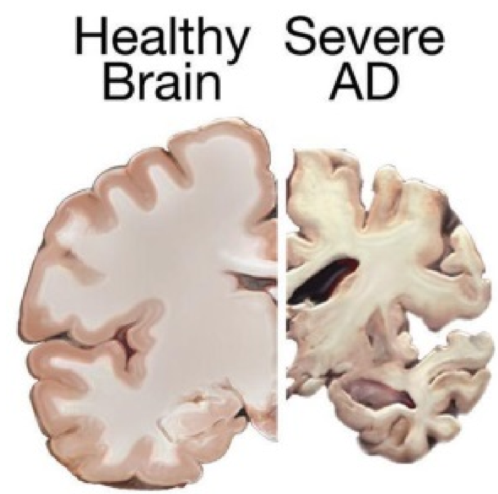
$Y_t \in \mathbb{R}^v$

Modelling goals

- Prediction

Predictive Density

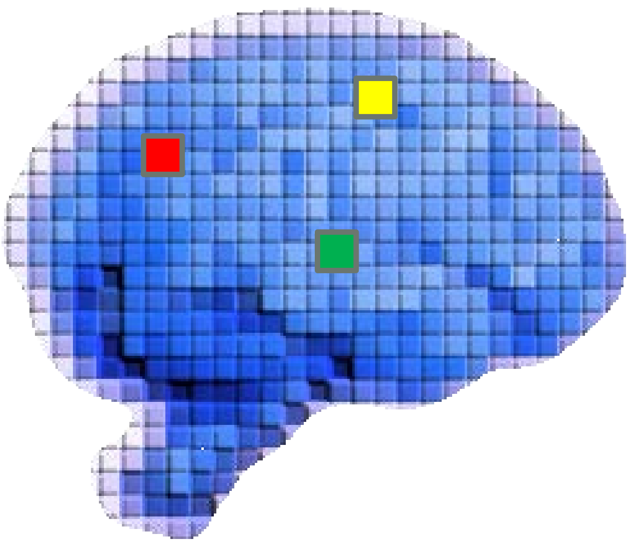
$$p(X_{\text{new}}|Y_{\text{new}}, X, Y) = \int p(X_{\text{new}}|\theta, Y_{\text{new}})q(\theta)d\theta$$



Modelling goals

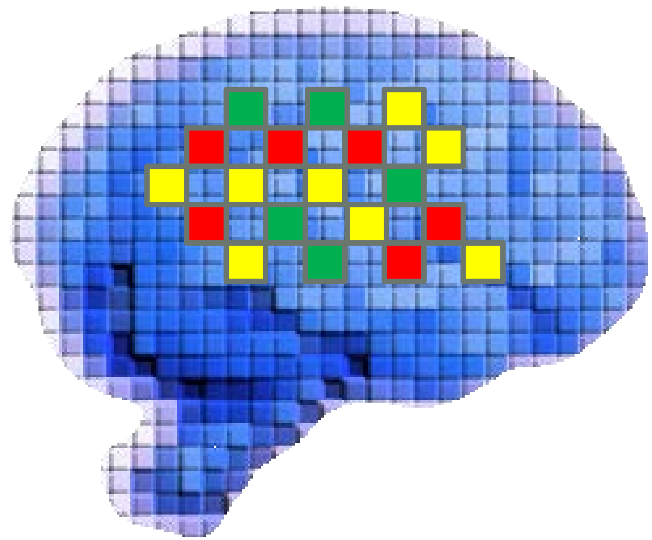
- Model Selection

Sparse Coding



VS

Distributed Coding



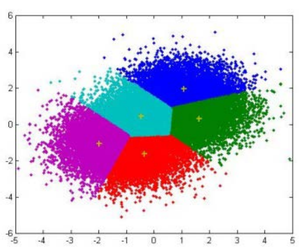
Model Evidence

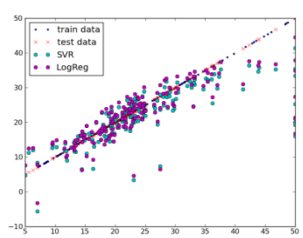
$$p(Y|X) = \int p(Y, \theta|X) d\theta$$

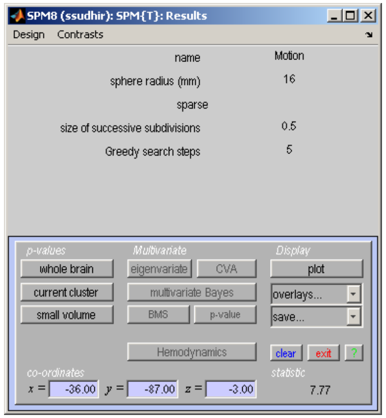
Overview

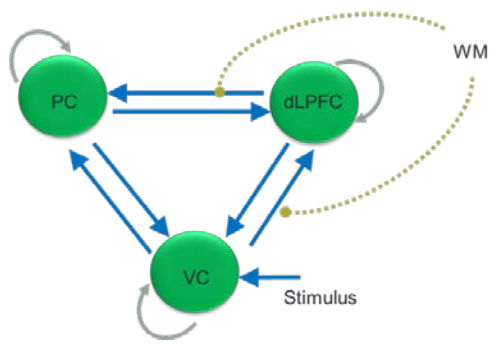
Motivation

Modelling Concepts | Learning From Data | Multivariate Bayes in SPM | Generative Embedding









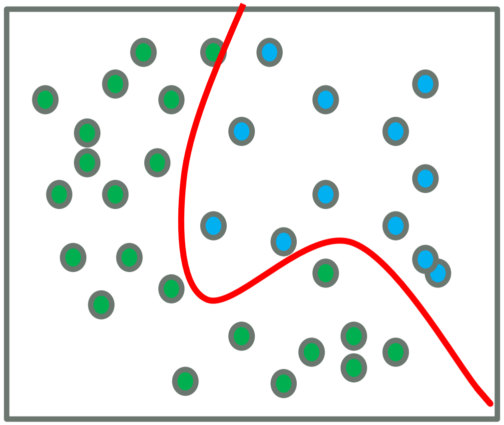
Learning from data

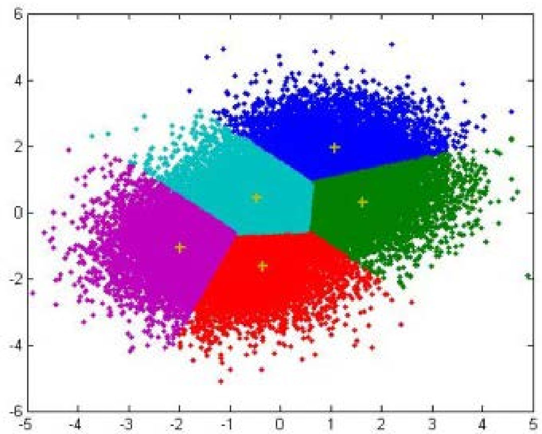
Supervised Learning

Unsupervised Learning

Reinforcement Learning

Semi-supervised Learning





Labels for training data are known!

Labels for training data are NOT known!

Supervised learning

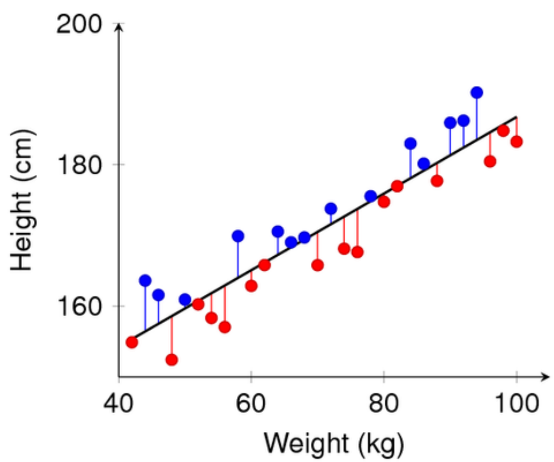
Independent variables
X

Function - f

Regression

dependent variable
Y

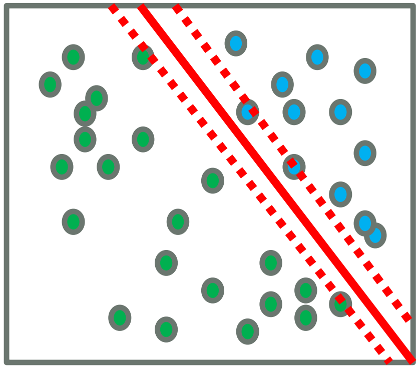
Continuous



Classification

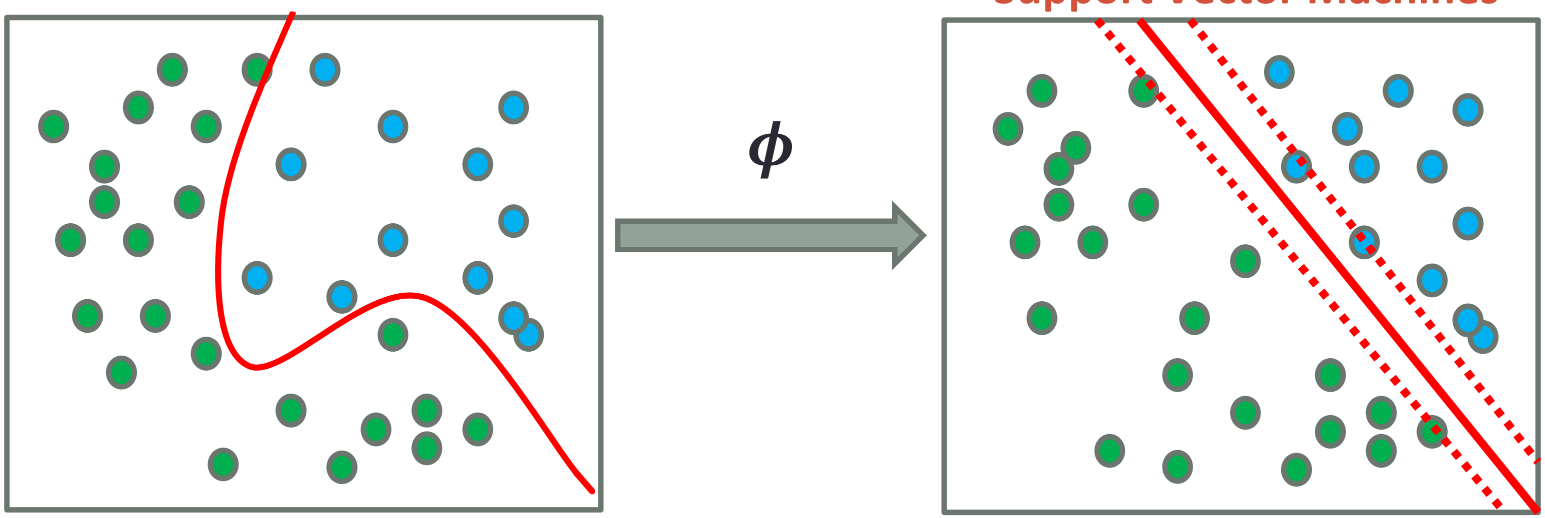
Categorical

Classification



Kernel Methods

Support Vector Machines

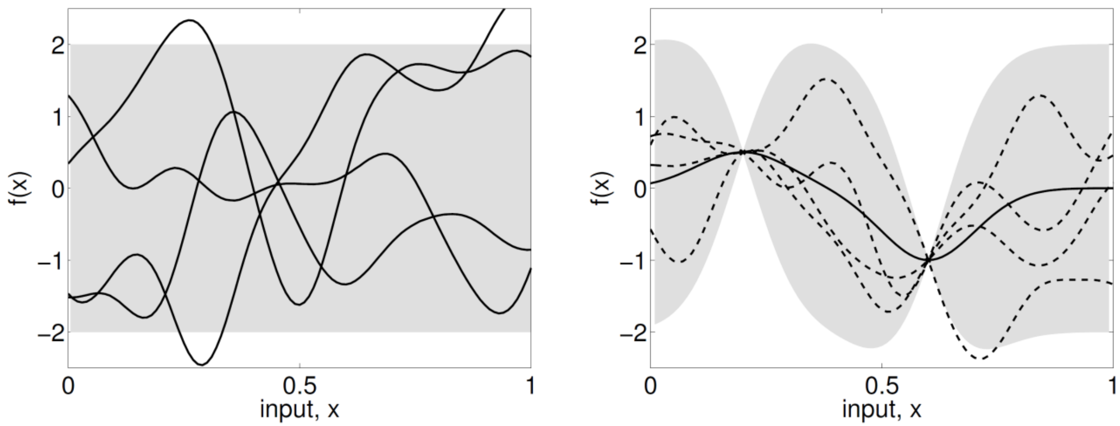


- Kernel Function – $\mathrm{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\phi}(\boldsymbol{x_i}).\boldsymbol{\phi}(\boldsymbol{x_j})$

Kernel methods for pattern analysis, Taylor , Cristianini, 2004

Other popular classifiers

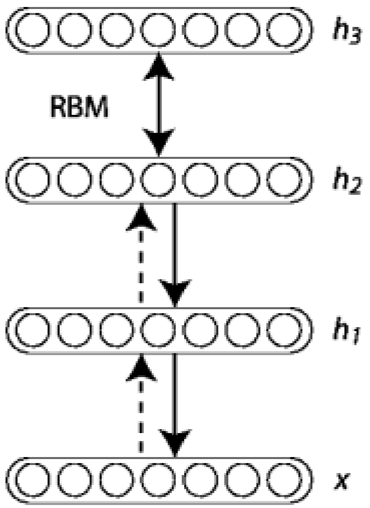
- Gaussian Processes



(a), prior (b), posterior

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006,

- Deep Belief networks



$$P(x, h^1, \ldots, h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \right) P(h^{\ell-1}, h^\ell)$$

http://deeplearning.net/tutorial/DBN.html

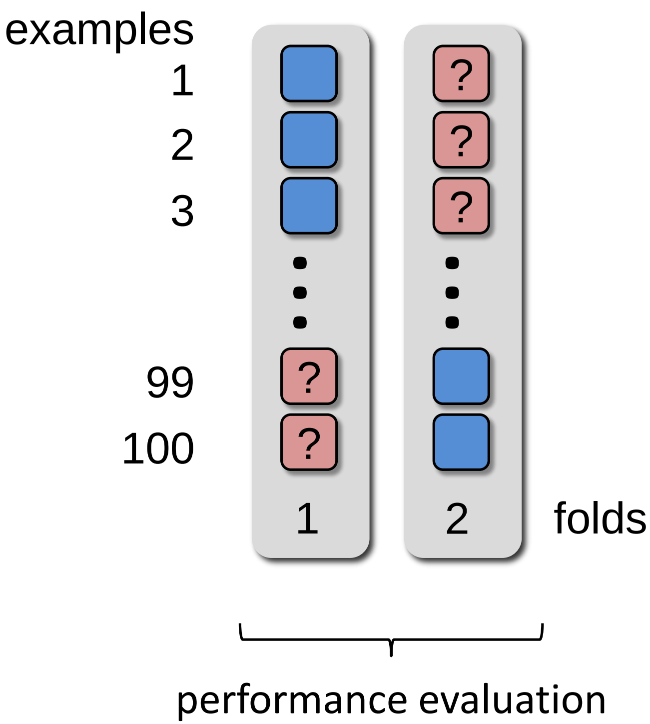
G.E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets”, Neural Computation, vol 18, 2006

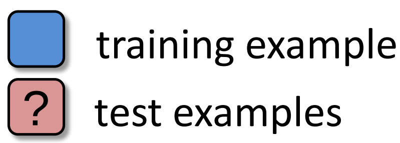
Generative and Discriminative classifiers

- Generative classifiers
  - Learn the parameters for the functions p(Y) and p(X|Y), e.g. Naïve Bayes Classifier
- Discriminative classifiers
  - Learn the parameters for p(Y|X), e.g. logistic regression, SVM

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:

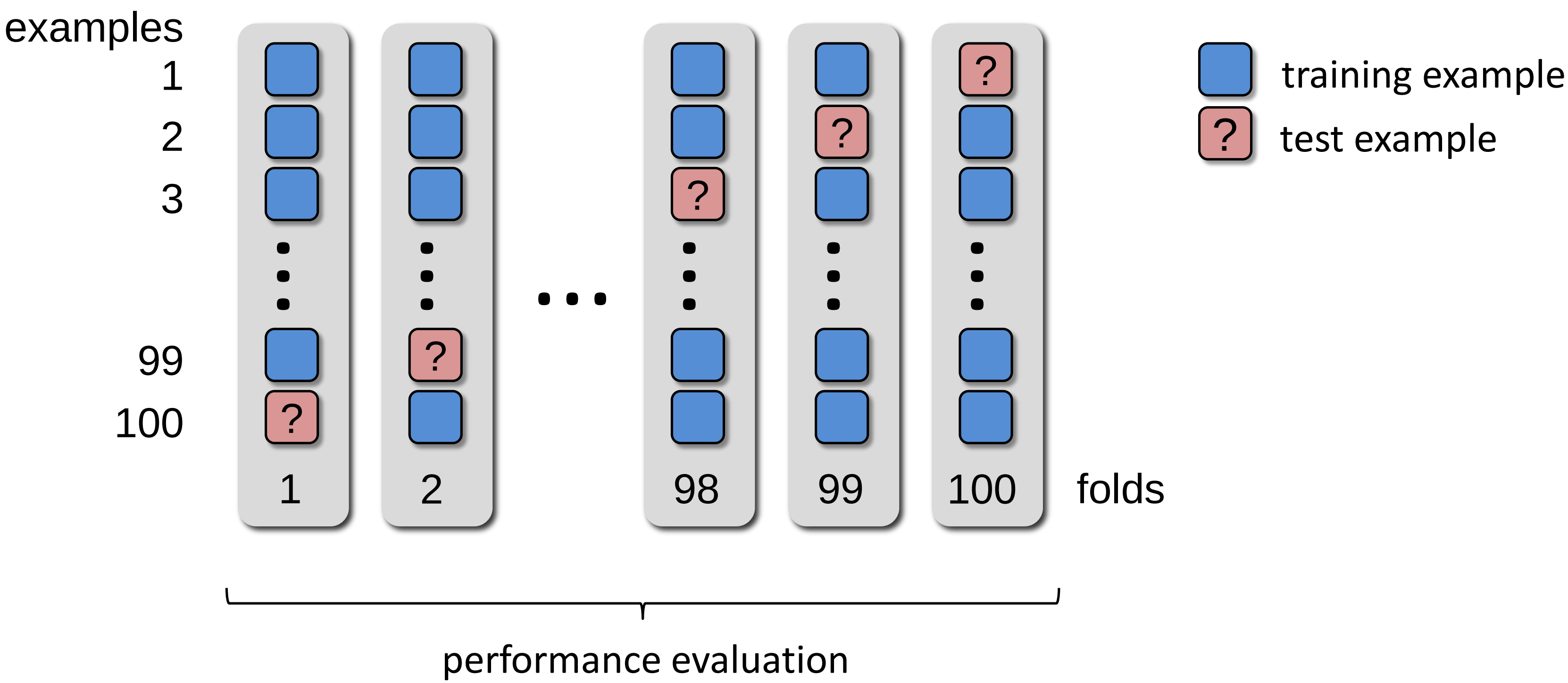




- Model Selection
- Performance evaluation
  - Balanced Accuracy
  - F1 Score

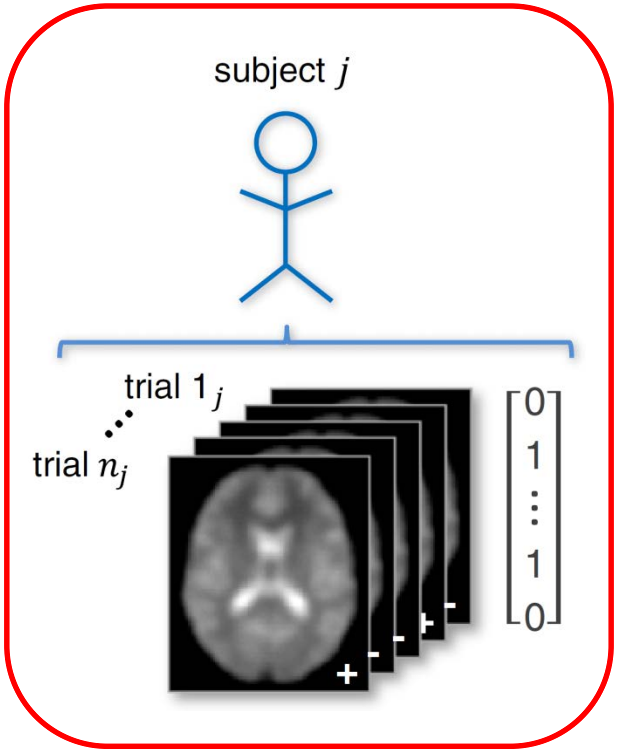
Cross-validation

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



In fMRI often leave one-run-out

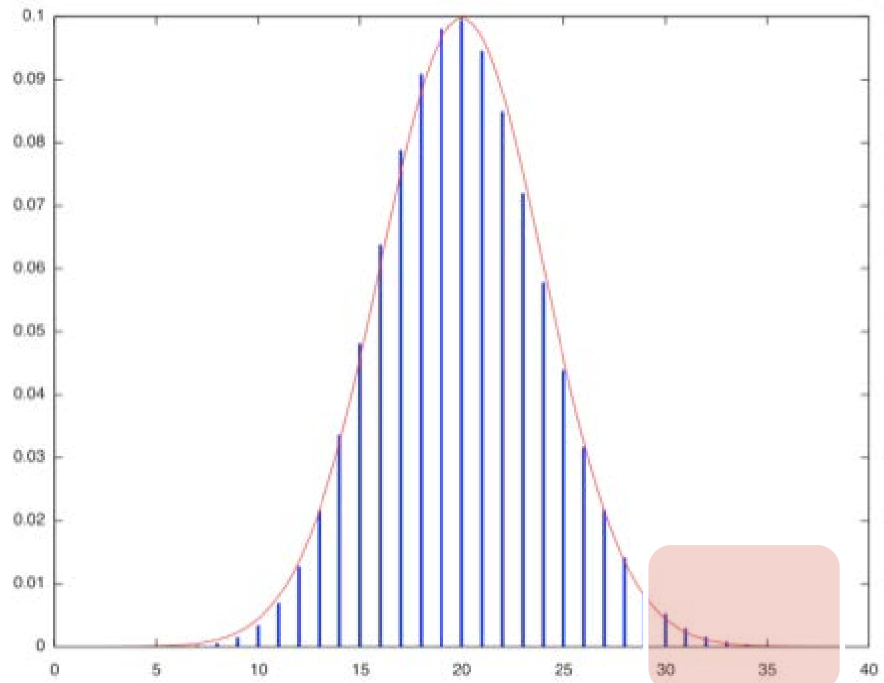
Performance – Single Subject



Brodersen et al. 2013, *NeuroImage*

Binomial Test

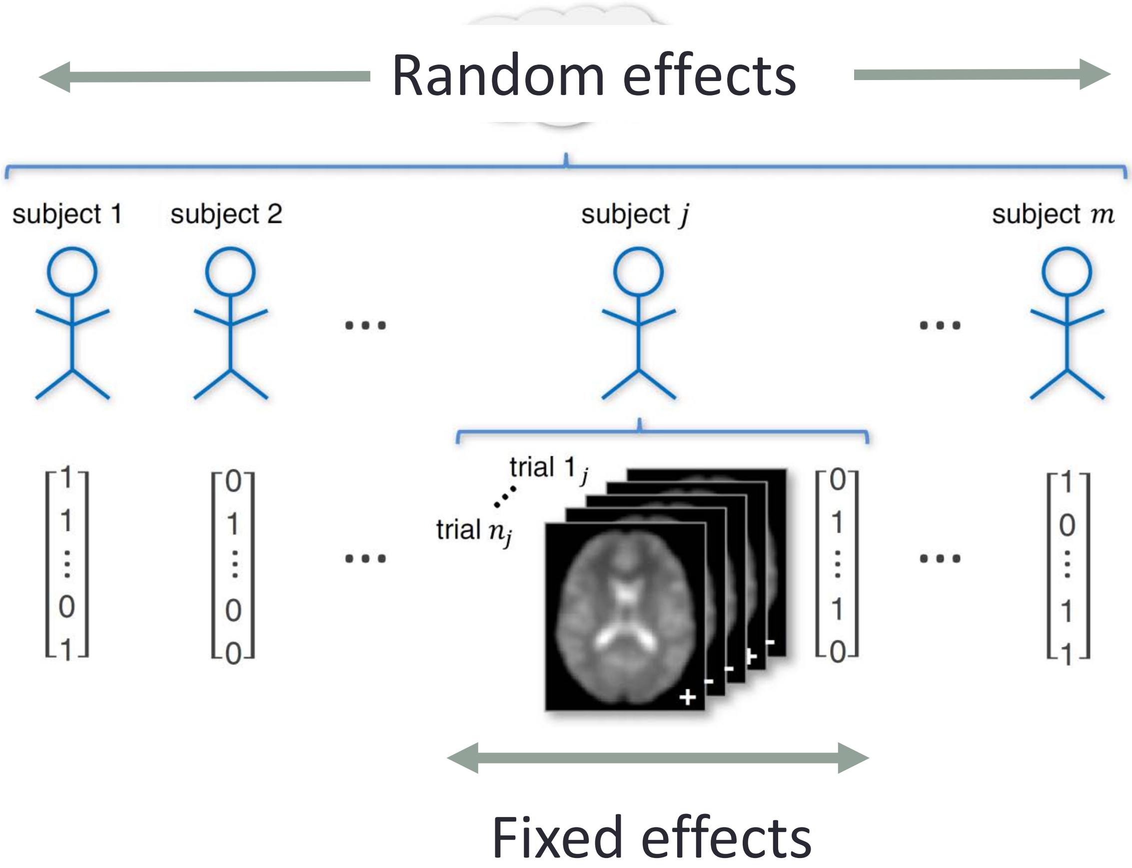
$$p = P(X \geq k|H_0) = 1 - B(k|n, \pi_0)$$

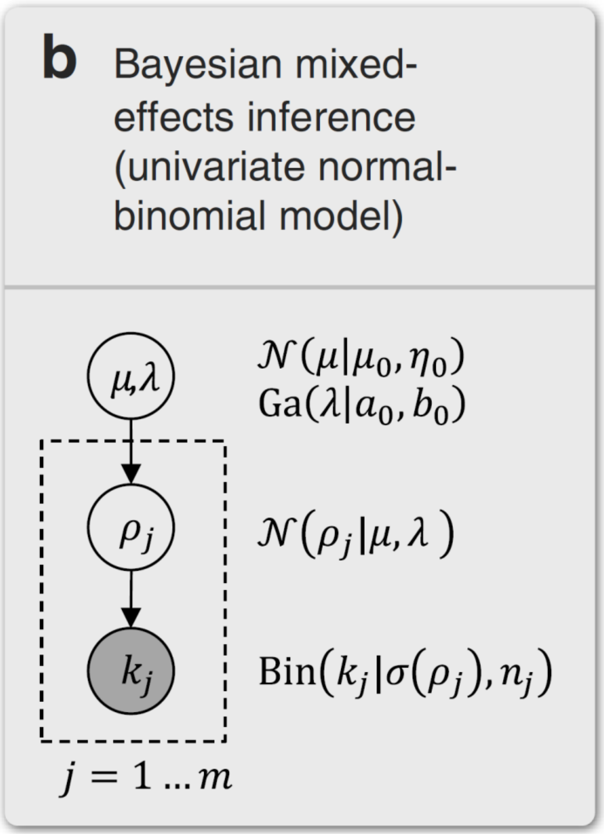


k=30

!!! Cross-validated data are not necessarily binomially distributed → Permutation tests are better!!!

Performance – Mulitple subjects

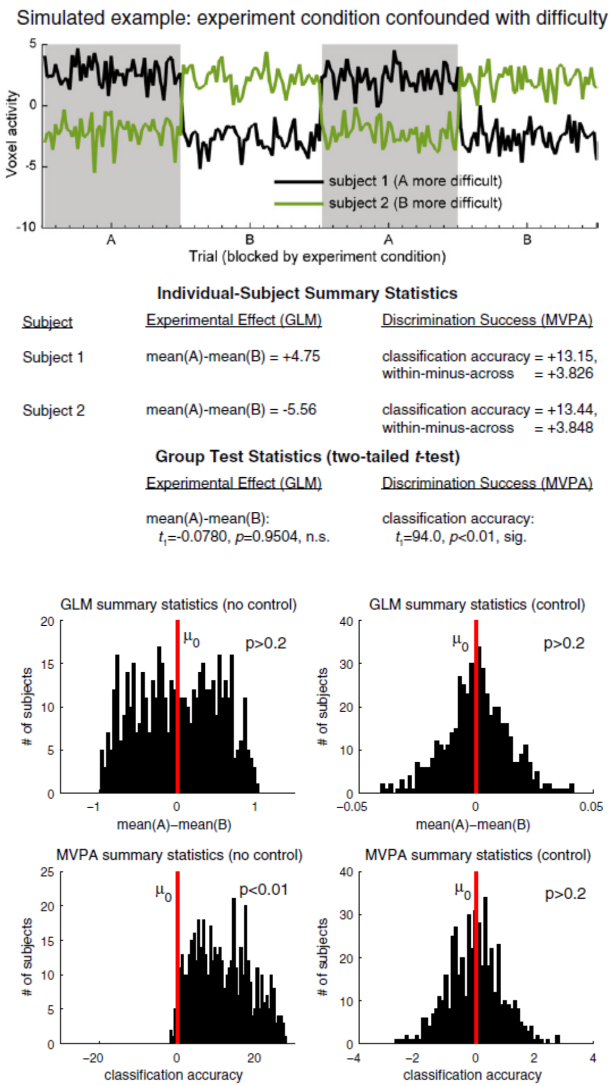


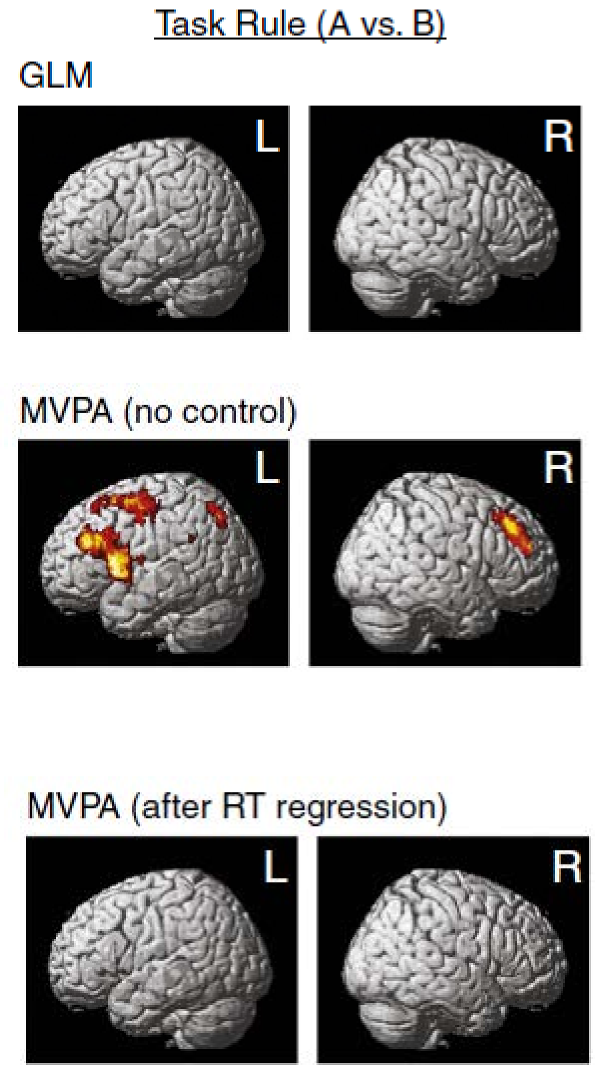


http://www.translationalneuromodeling.org/tapas/

Brodersen et al. 2013, *NeuroImage*

Confounds – GLM vs. MVPA





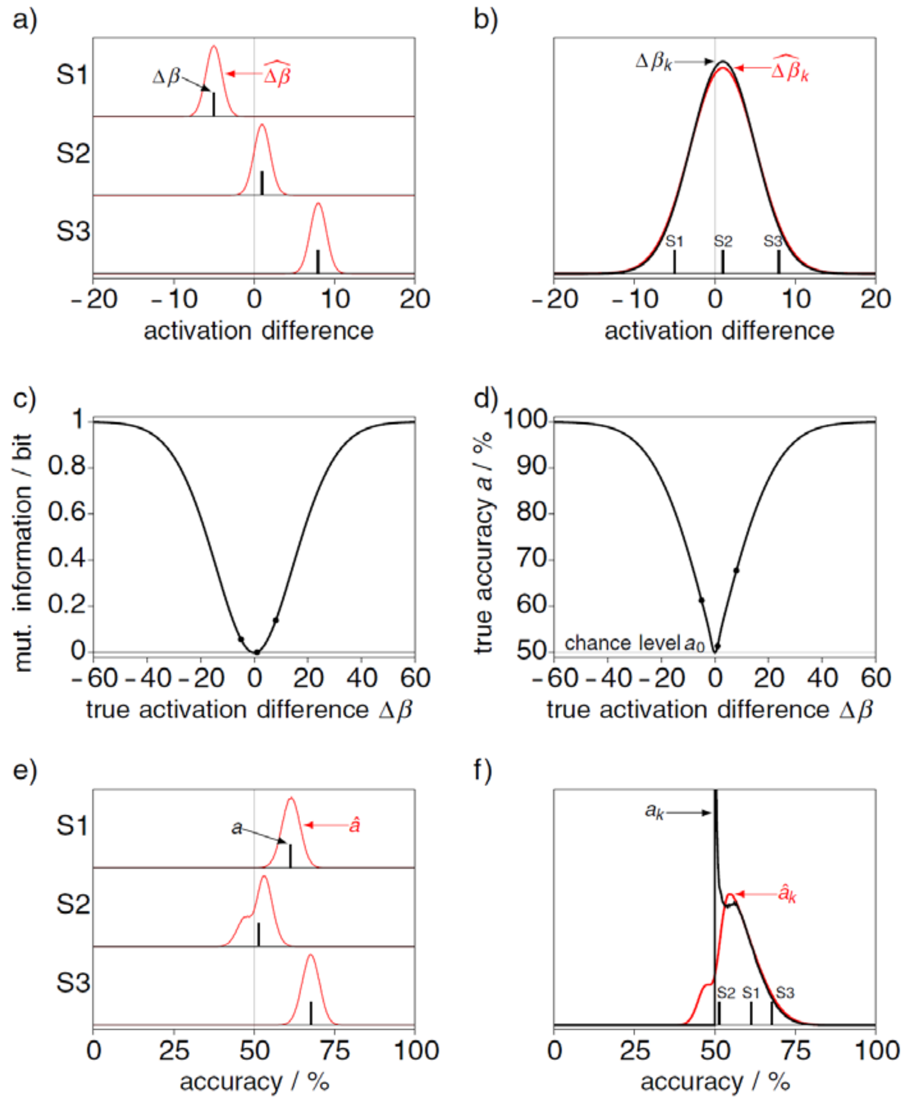
Todd et al. 2013, *NeuroImage*

Second level t-tests for accuracies?

True β-Values are normally distributed.

True accuracies are not normal and truncated at chance.

A possible solution is given by Allefeld et al.



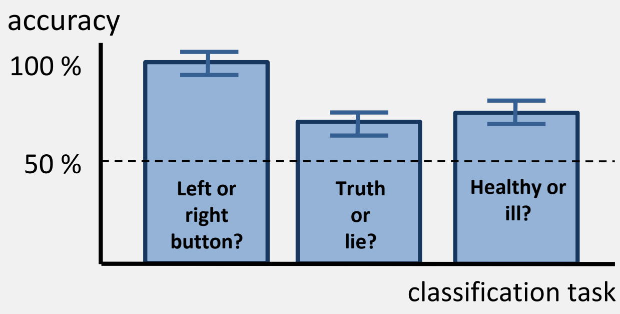
Allefeld et al. Neuroimage, 2016

Statistical testing with classification

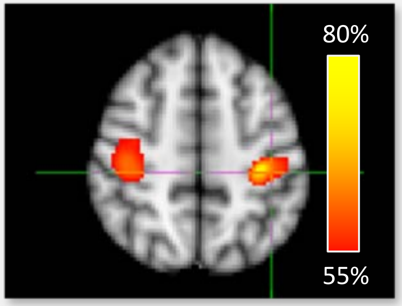
- Within subjects:
  - Permutation statistics
  - Parametric tests ar not valid (assumptions not met), e.g. Biomial- or t-test (c.f. Schreiber and Krekelberg, 2013).
- Across subjects:
  - Assumptions for t-tests are not met
  - Full Bayesian model (Bordersen et al. 2013, but assumptions are not met for CV)
  - Use prevalence statistic proposed in Allefeld et al., 2016

Research questions for classification

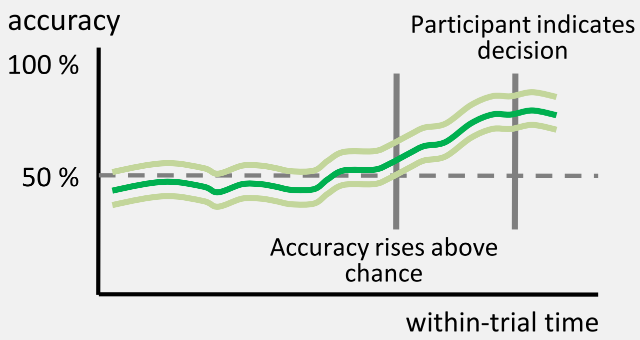
Overall classification accuracy



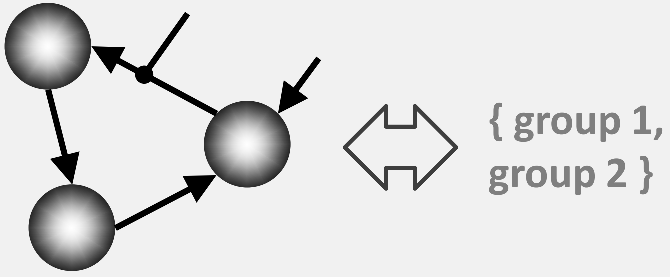
Spatial deployment of discriminative regions



Temporal evolution of discriminability

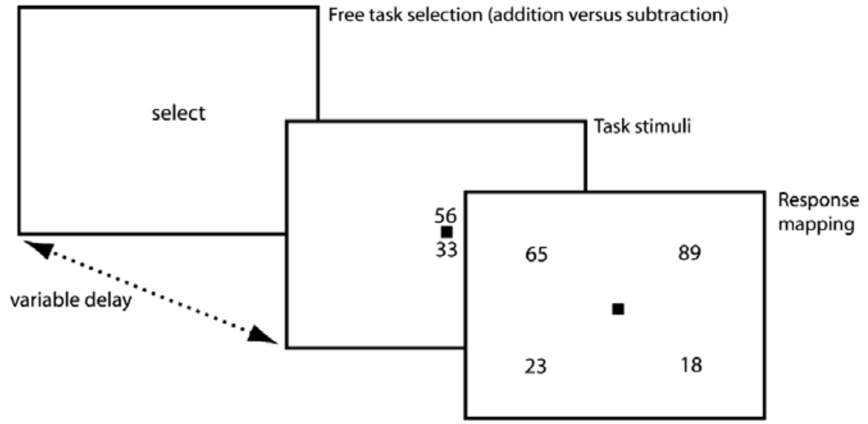


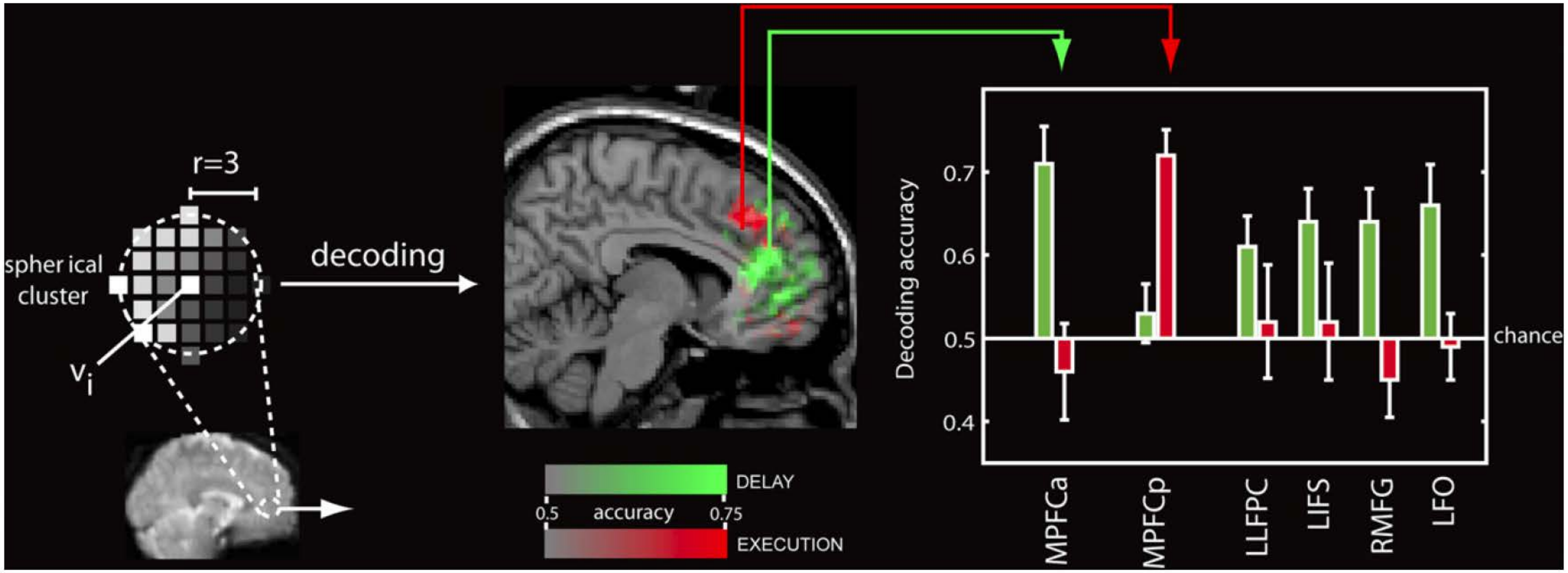
Model-based classification



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

Decoding «hidden» intentions – searchlight approach

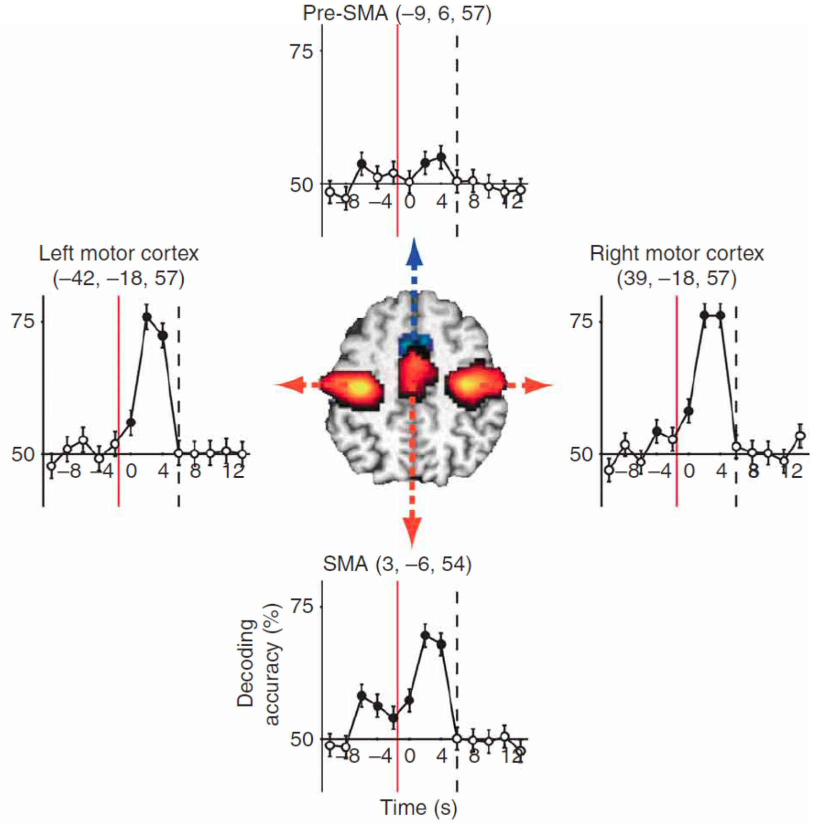


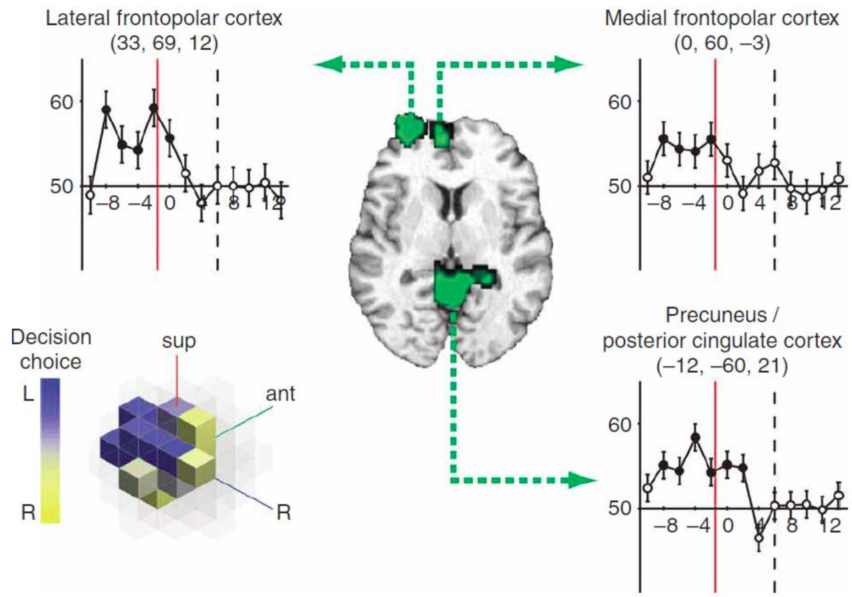


Haynes et al., Current Biology, 2007

Decoding of free decisions

Decoding of fingerpresses (red line). Participants freely choose timing and hand.

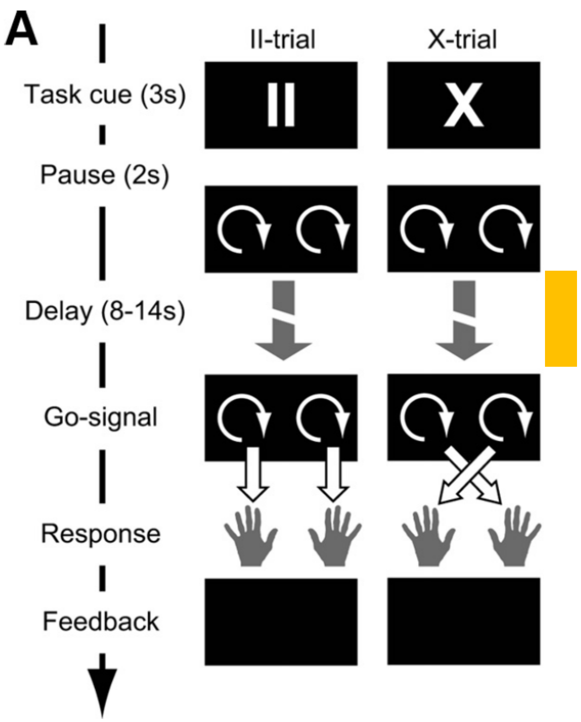




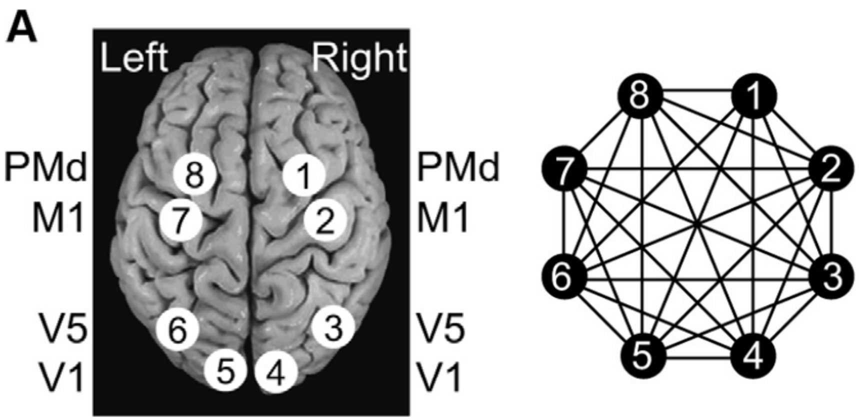
Earliest information about left-right long before execution – free will?

Soon et al., Nat Neurosci, 2008

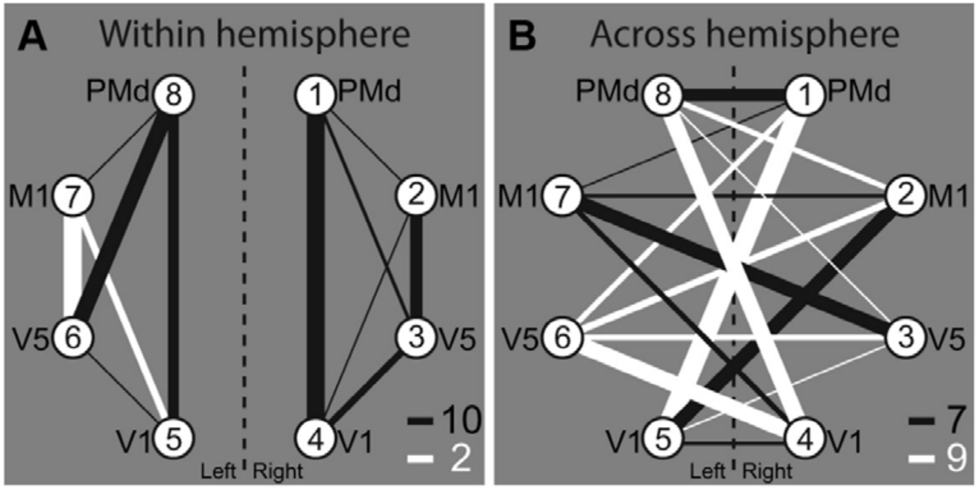
Decoding task preparation – connectitivy based decoding



SV-Classifier on connectivity graph (correlation)



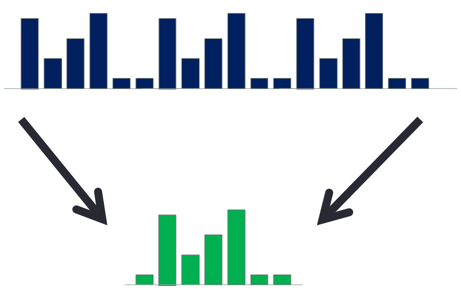
Discriminative maps

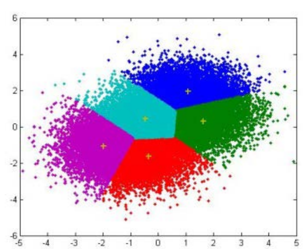


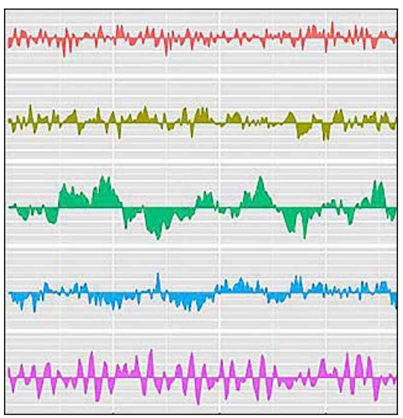
Heinzle et al., J Neurosci, 2012

Unsupervised learning

- Building a representation of data
  - Dimensionality Reduction
  - **Clustering**
    - K-means
    - Mixture models
  - Time series





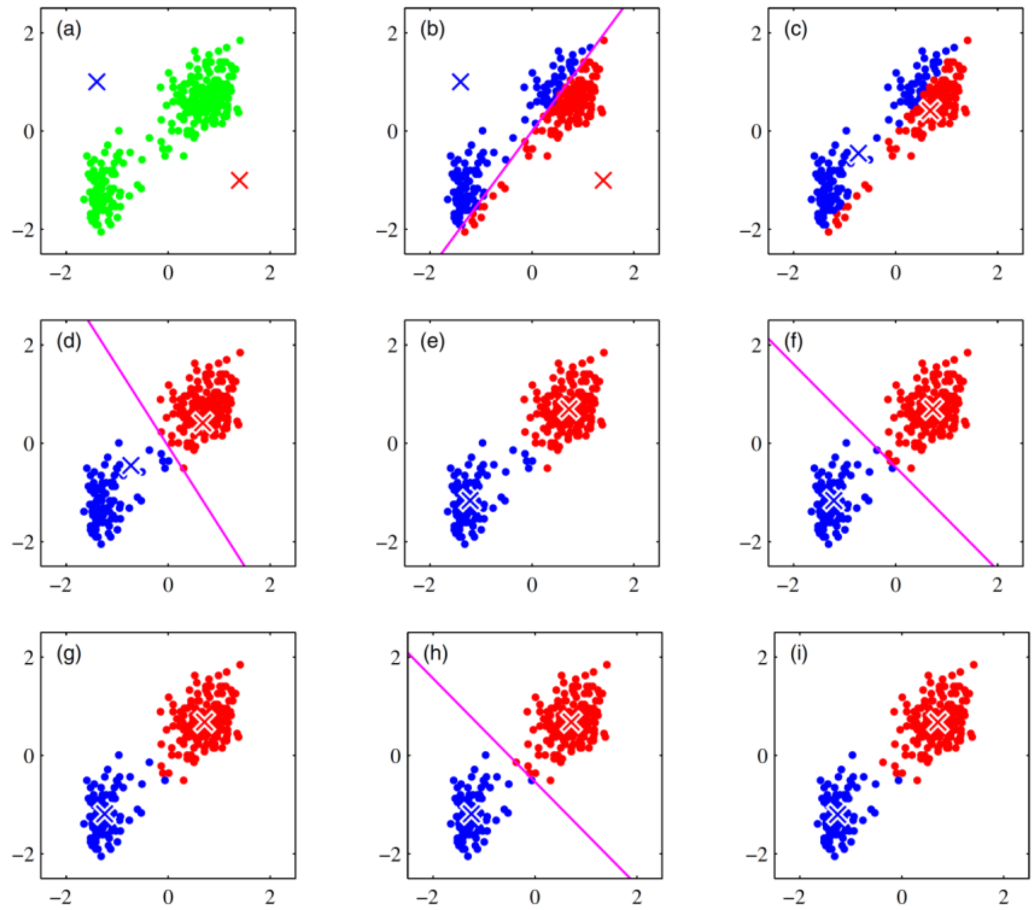


K-means clustering

- Cost function

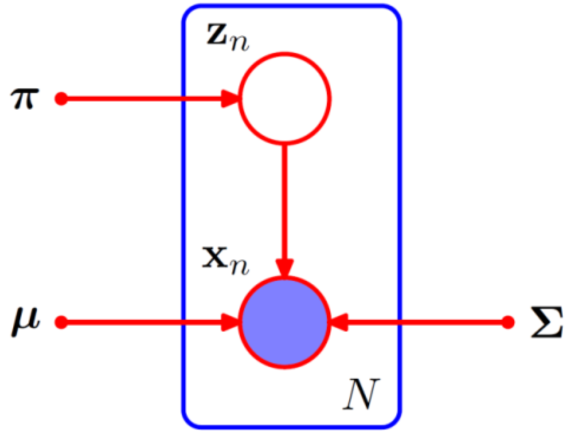
$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

- Algorithm
  1. Initialize
  2. Estimate assignments
  3. Estimate cluster centroids
  4. Repeat 2,3 until convergence

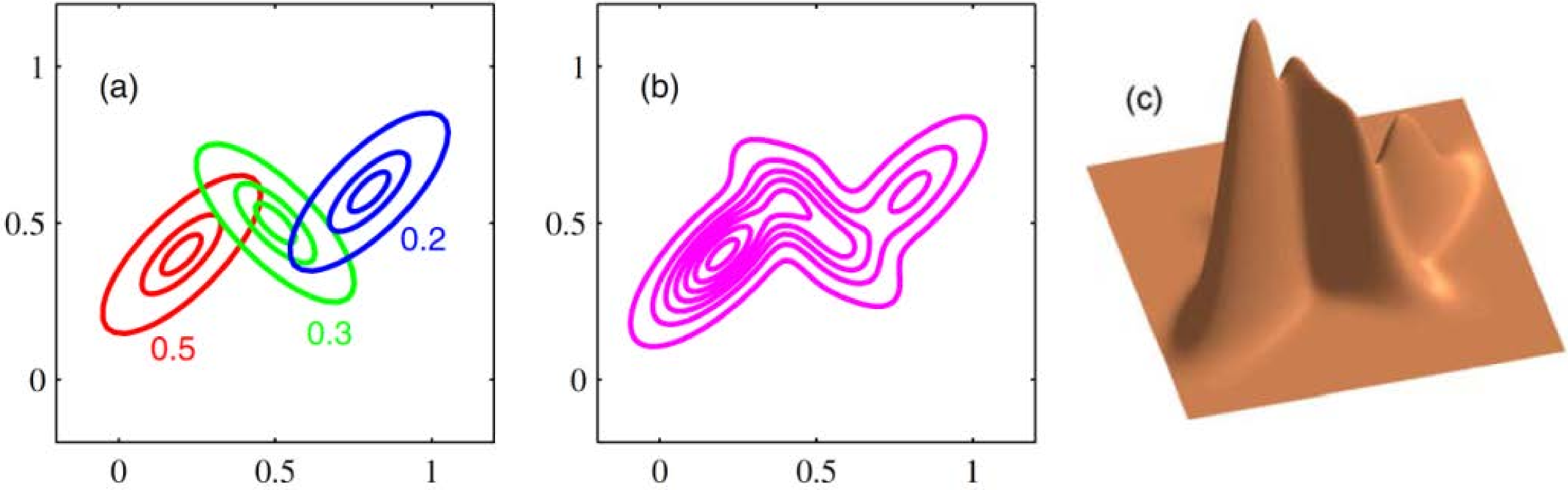


Bishop PRML (2006)

Clustering – Mixture of Gaussians



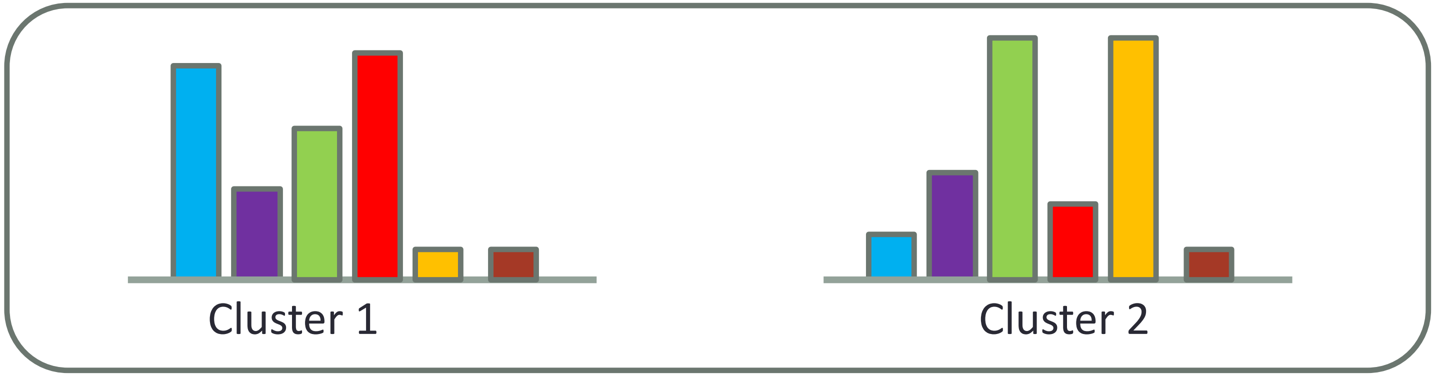
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



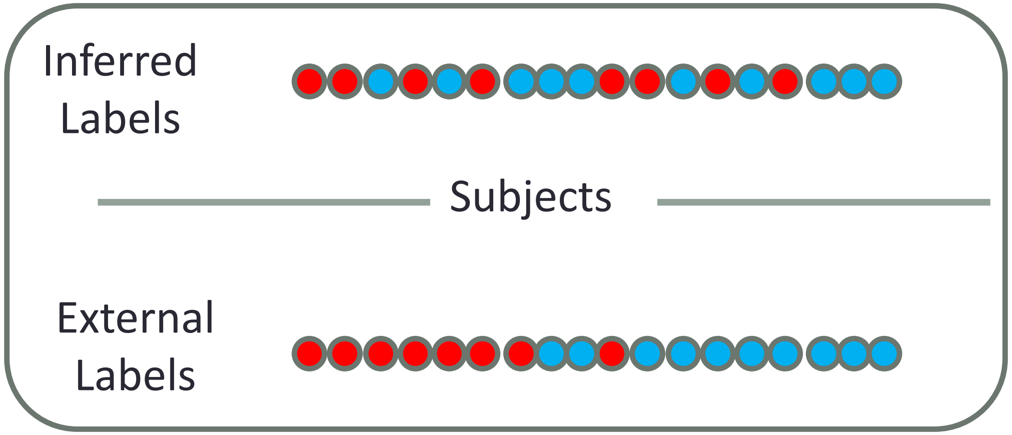
Bishop PRML (2006)

Interpretation

- Cluster parameters



- Internal Criterion – Model Evidence
- External Criterion - Purity



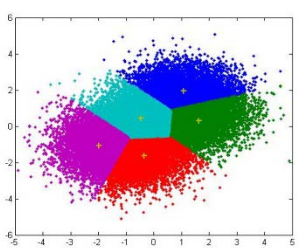
Normalized Mutual Information (NMI)

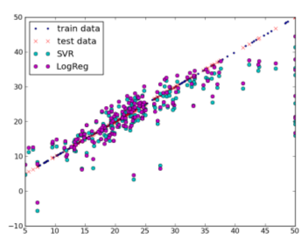
Balanced purity

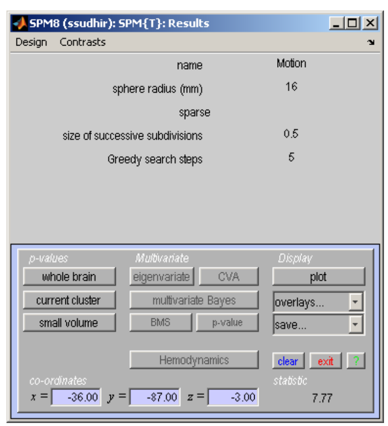
Rand Index

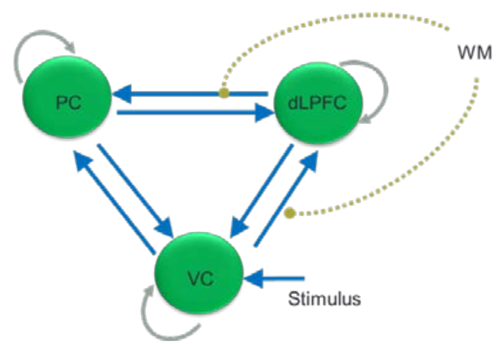
Motivation

- Modelling
- Learning from Data
- Multivariate Bayes in SPM
- Generative Embedding

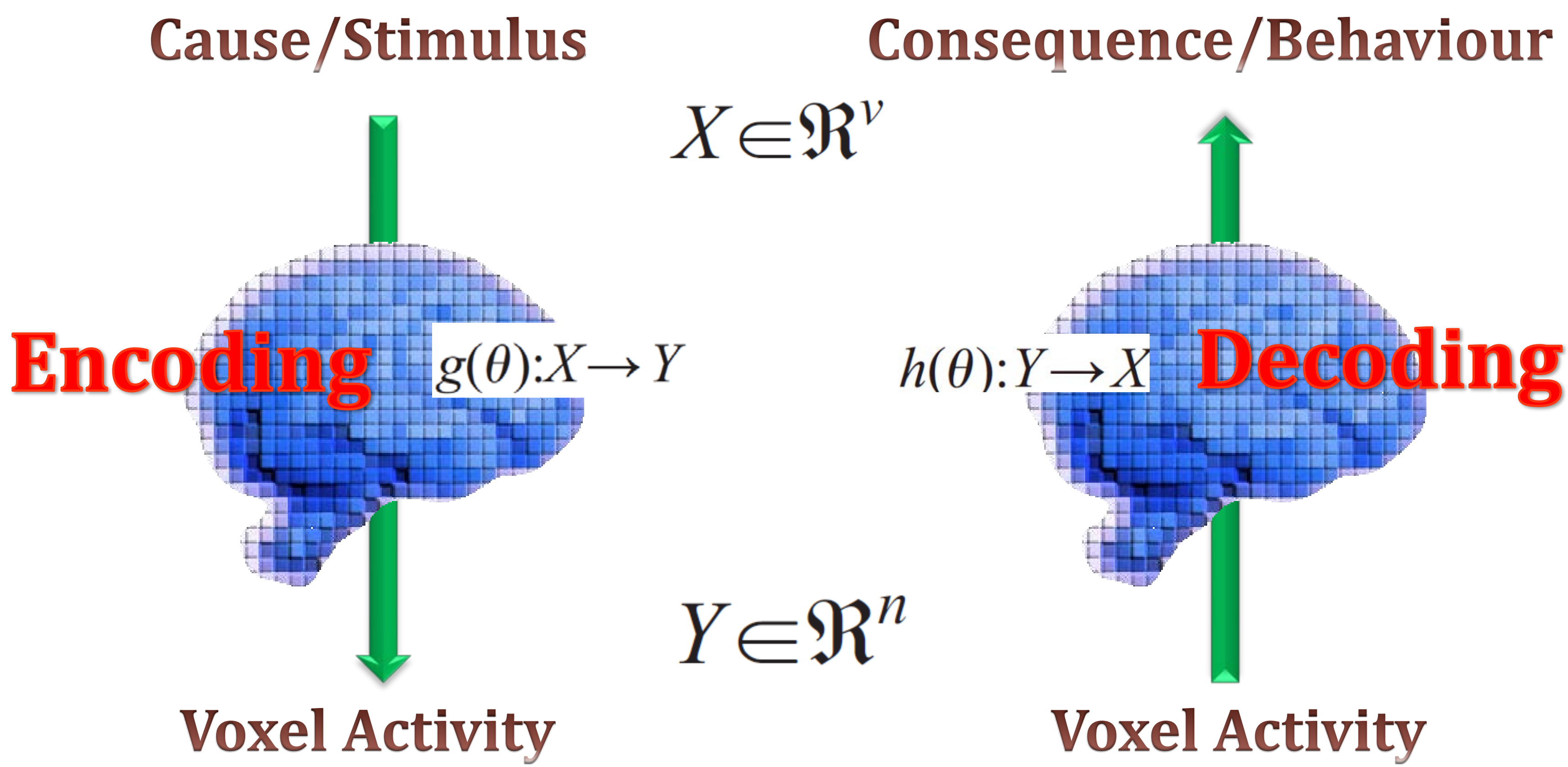




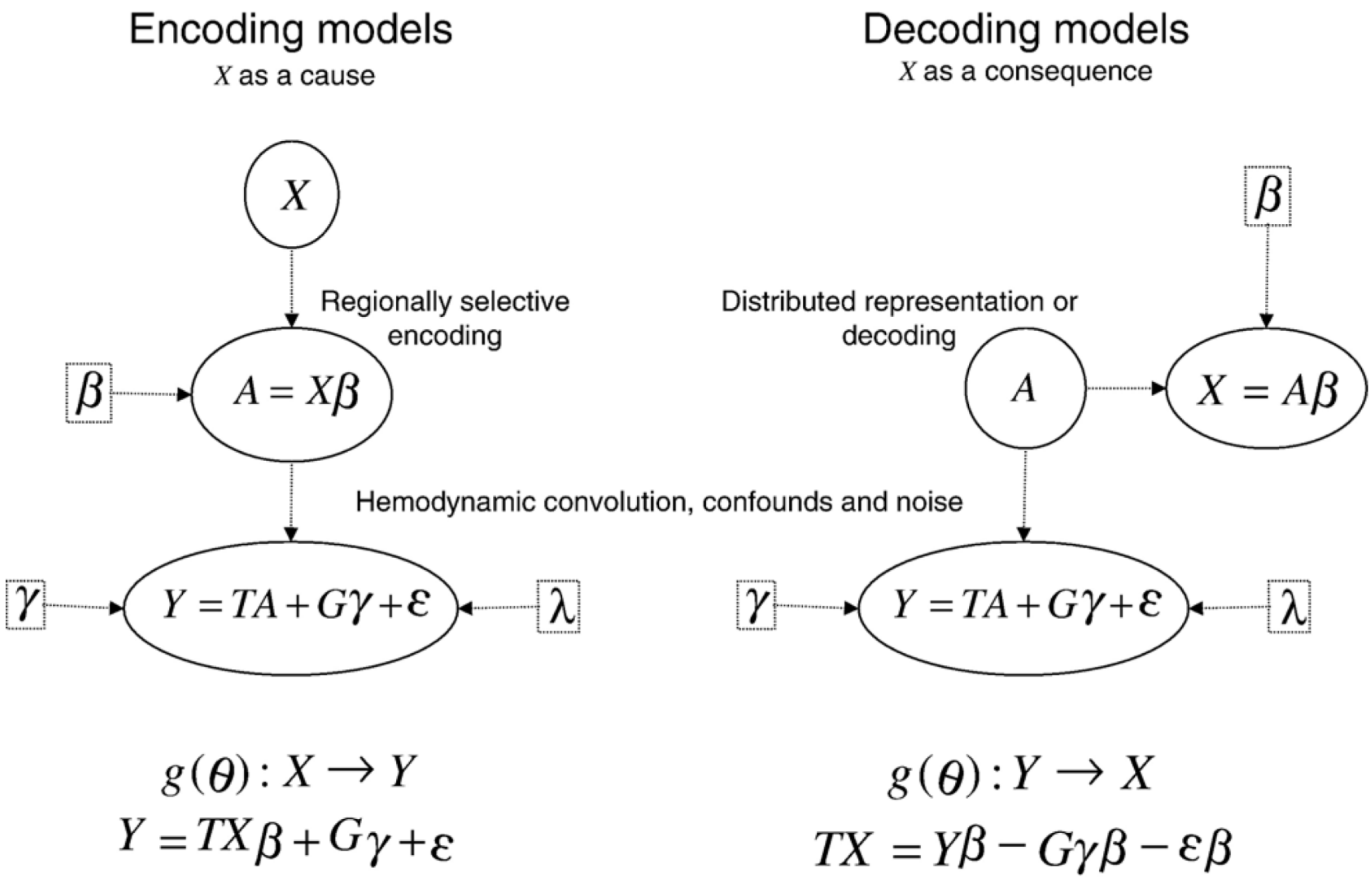




Encoding vs. Decoding models

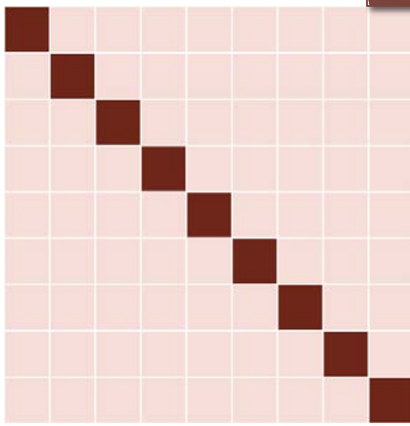


Encoding vs. Decoding models

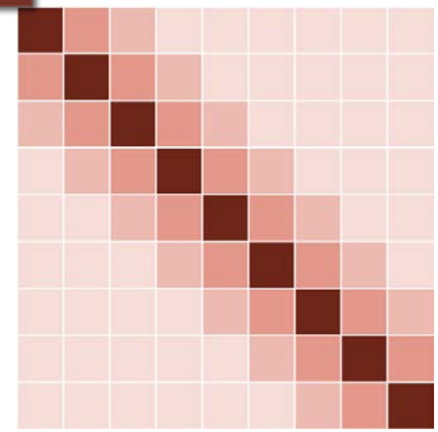


Coding Hypotheses

Sparse vectors



Spatial vectors

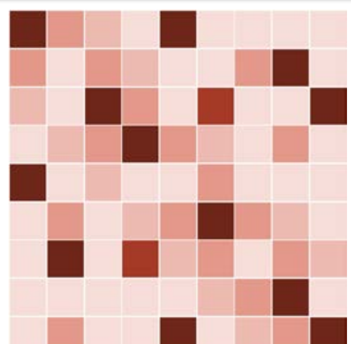


Smooth vectors

Distributed vectors

Singular vectors
of data

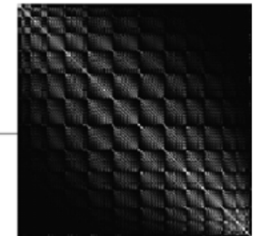
$$UDV^T = RY^T$$



Support vectors

$$U = RY^T$$

Coding Hypotheses



Null: $U = \varnothing$

Spatial vectors: $U = I$

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

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

Support vectors: $U = RY^T$

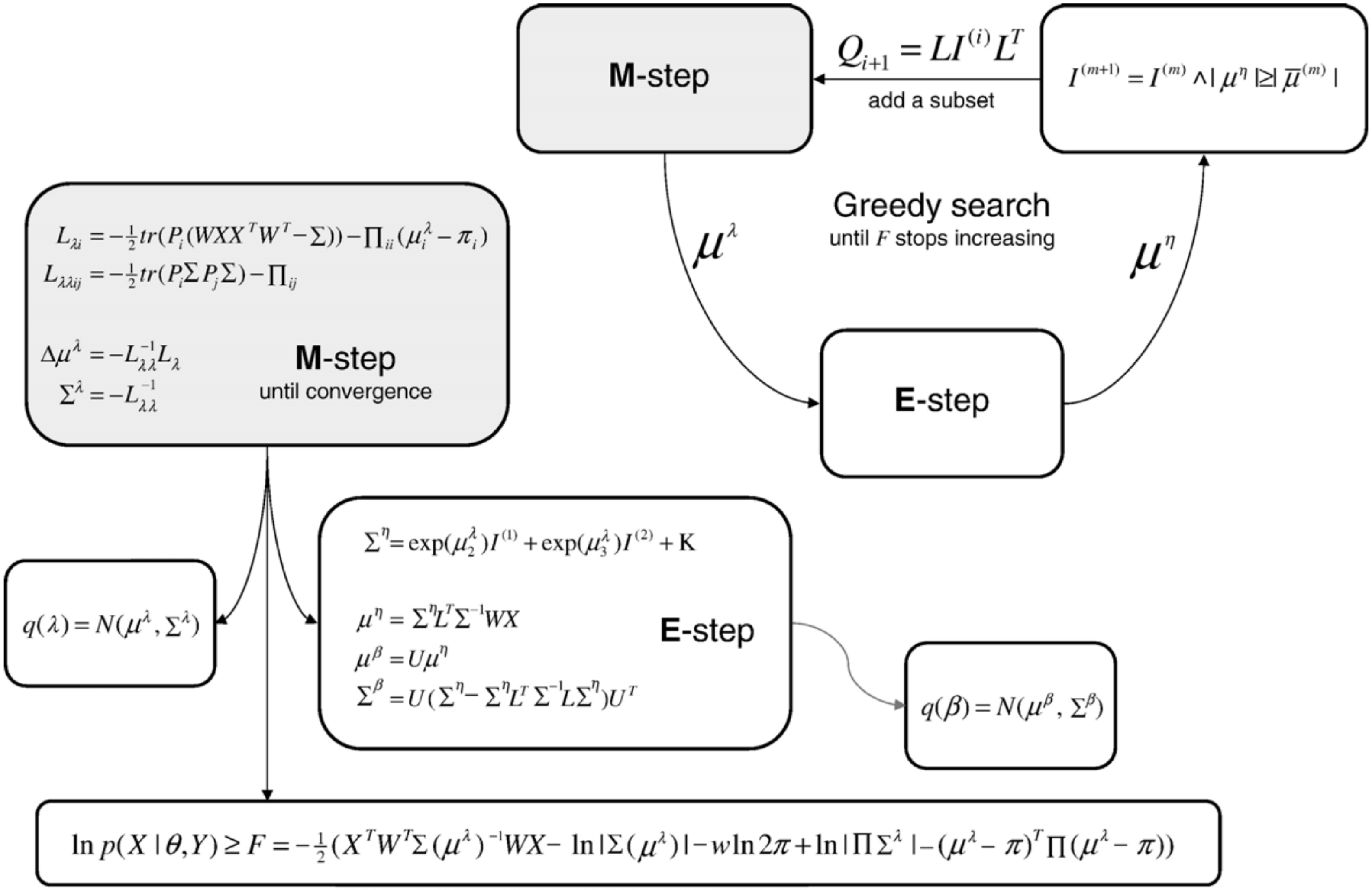
$$
\begin{aligned}
WX &= RY\beta + \varsigma \\
\beta &= U\eta \\
\operatorname{cov}(\varsigma) &= \Sigma^{\varsigma}(\lambda) = \exp(\lambda^{\varsigma})RVR^T \\
\operatorname{cov}(\eta) &= \Sigma^{\eta}(\lambda) = \exp(\lambda_1^{\eta})I^{(1)} + \ldots + \exp(\lambda_m^{\eta})I^{(m)}
\end{aligned}
$$

$$p(\beta) = N(0, U\Sigma^{\eta}\, U^T)$$

$$
\begin{aligned}
W &= RT \\
R &= orth(I - GG^{-})^T
\end{aligned}
$$

Friston et al. 2008 NeuroImage

Solved with variational Bayes



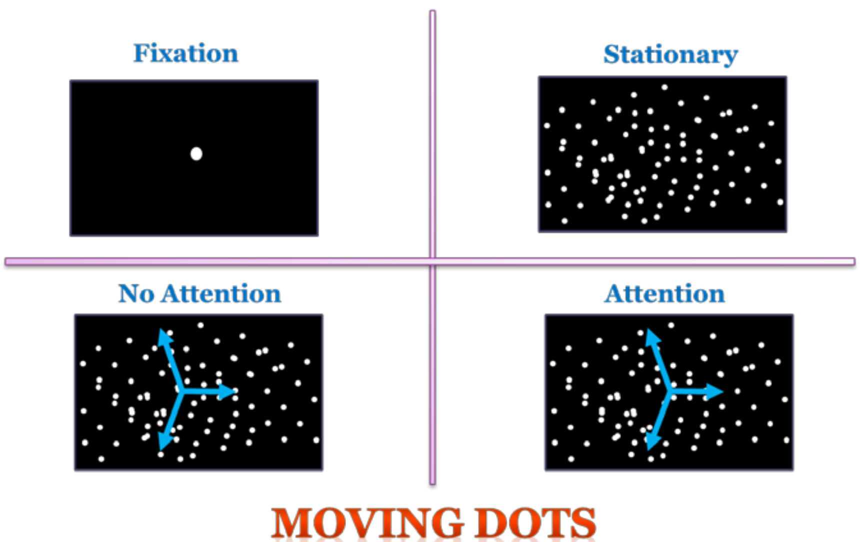
Friston et al. 2008 NeuroImage

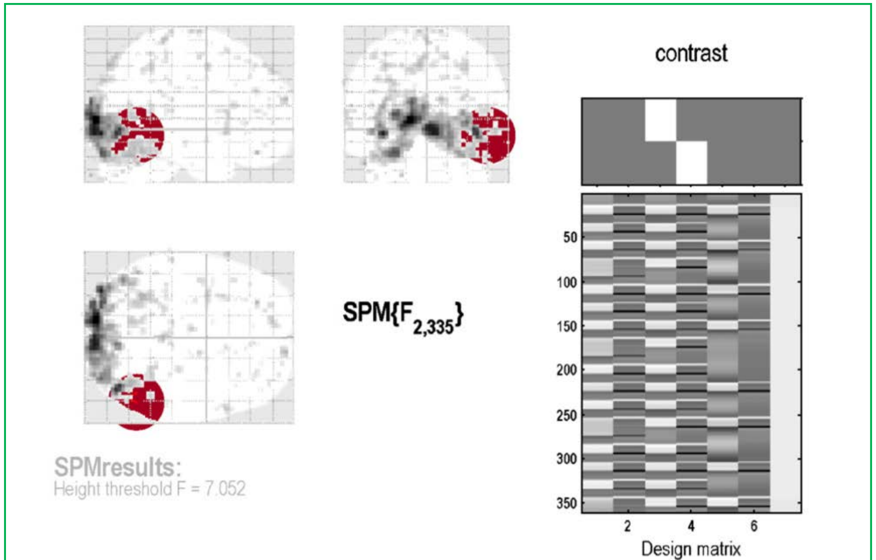
Example – Decoding of motion.

Attention to motion dataset - Büchel & Friston 1999 Cerebral Cortex

Experimental factors:

1. Photic
2. Motion
3. Attention





Statistics: *search volume: 16.0mm sphere at [48,-63,0]*

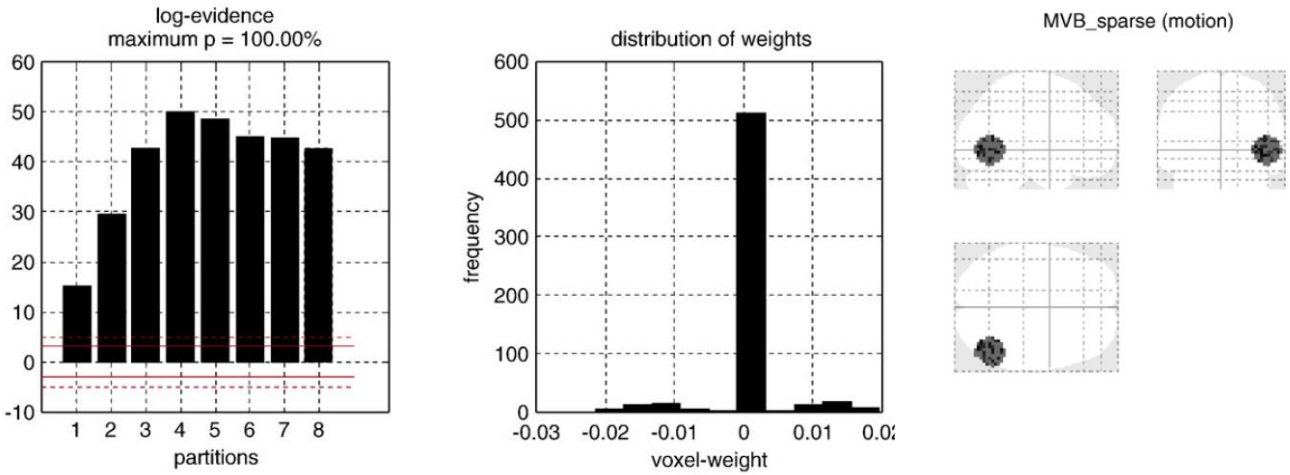
set-level p	set-level c	cluster-level $p_{corrected}$	cluster-level k_E	cluster-level $p_{uncorrected}$	voxel-level $p_{FWE-corr}$	voxel-level $p_{FDR-corr}$	voxel-level F	voxel-level $(Z_{\equiv})$	voxel-level $p_{uncorrected}$	mm mm mm
0.006	4		39		0.000	0.000	40.66	Inf	0.000	39 -72 0
					0.000	0.000	40.09	Inf	0.000	42 -75 -3
			3		0.002	0.000	13.35	4.55	0.000	39 -63 -12
			1		0.124	0.003	8.67	3.52	0.000	51 -66 -3
			1		0.436	0.010	7.36	3.18	0.001	48 -63 0

table shows 16 local maxima more than 4.0mm apart

Height threshold: F = 7.05, p = 0.001 (0.513) {p<0.001 (unc.)}
Extent threshold: k = 0 voxels, p = 1.000 (0.513)
Expected voxels per cluster, <k> = 1.212
Expected number of clusters, <c> = 0.72
Expected false discovery rate, <= 0.01

Degrees of freedom = [2.0, 335.0]
FWHM = 7.5 7.4 7.1 mm mm mm; 2.5 2.5 2.4 {voxels};
Volume: 15741; 583 voxels; 43.3 resels
Voxel size: 3.0 3.0 3.0 mm mm mm; (resel = 14.69 voxels)

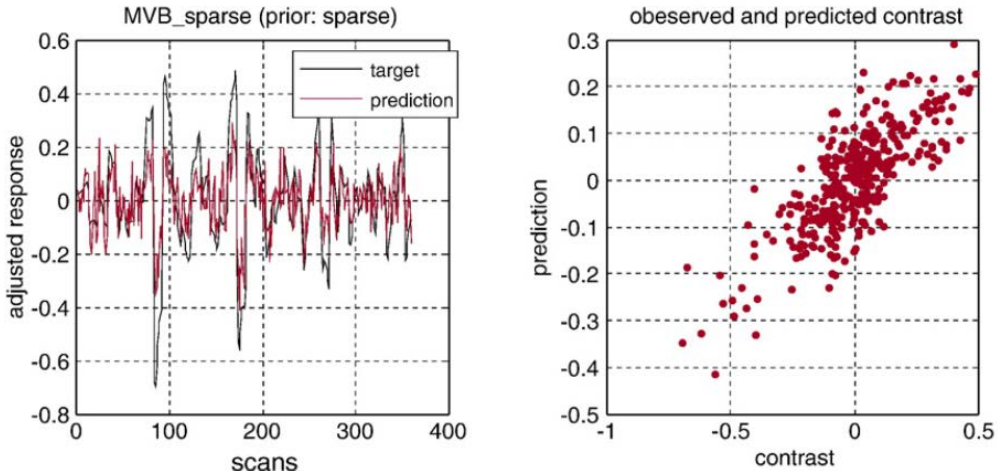
Friston et al. 2008 NeuroImage



Posterior probabilities at maxmia

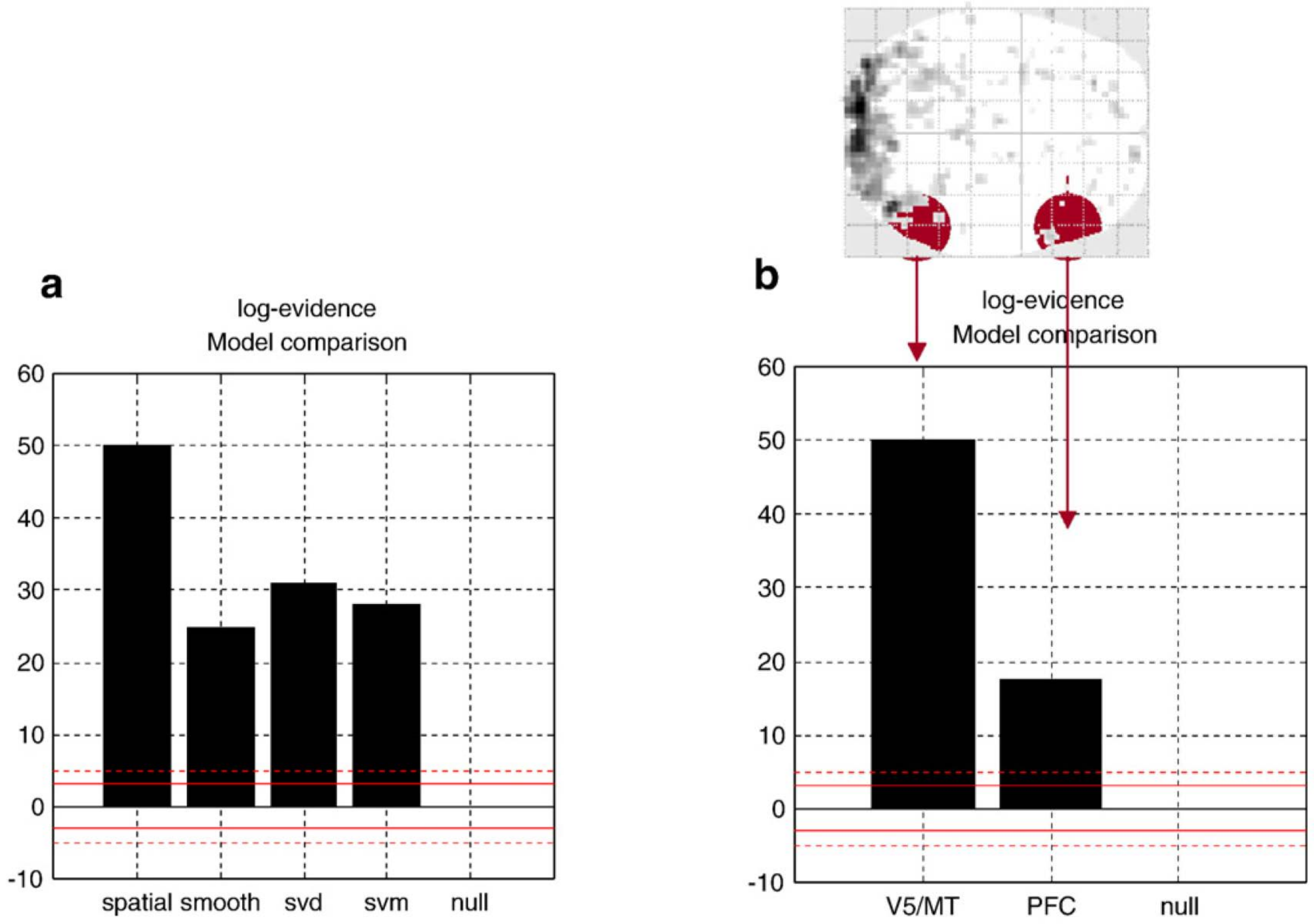
p(lwl > 0)	location (x,y,z)	weight (w)
p = 0.991	48,-78,0mm	q = -0.0208;
p = 0.977	48,-72,-3mm	q = -0.0215;
p = 0.973	36,-72,3mm	q = 0.0185;
p = 0.972	45,-51,9mm	q = 0.0188;
p = 0.968	39,-66,-6mm	q = -0.0180;
p = 0.966	42,-54,-3mm	q = -0.0168;
p = 0.963	45,-75,-6mm	q = 0.0196;
p = 0.954	54,-54,9mm	q = 0.0154;
p = 0.947	63,-60,3mm	q = -0.0161;
p = 0.945	42,-63,0mm	q = 0.0150;
p = 0.942	60,-60,-9mm	q = -0.0136;
p = 0.942	36,-57,6mm	q = -0.0167;

583 voxels; 360 scans



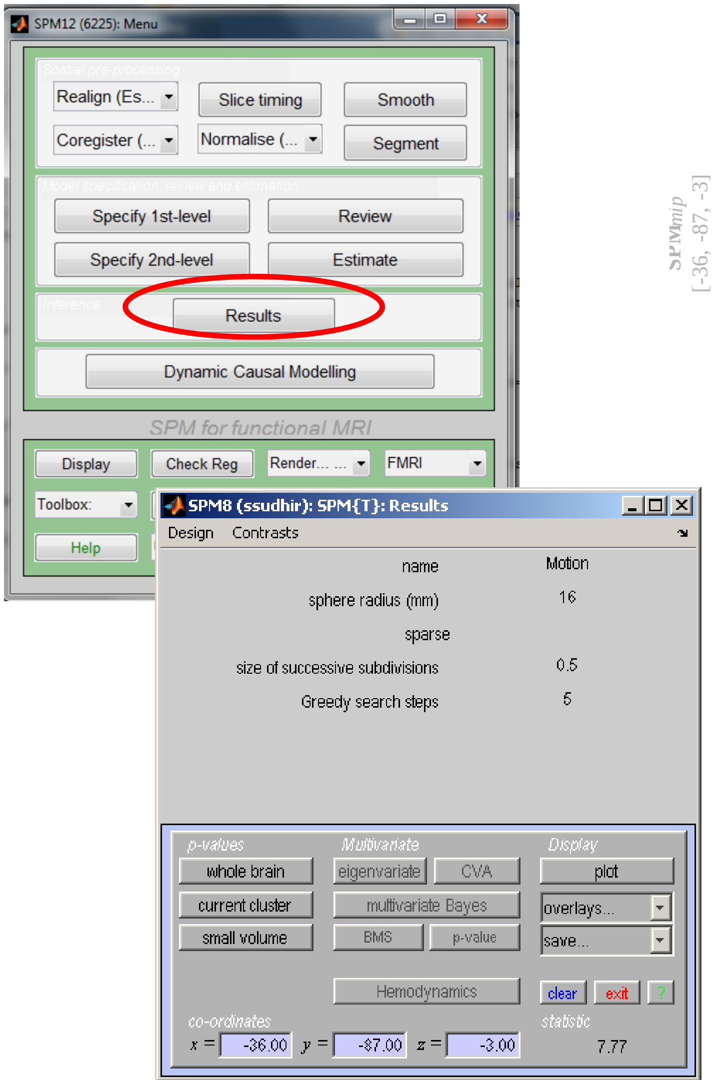
Friston et al. 2008 NeuroImage

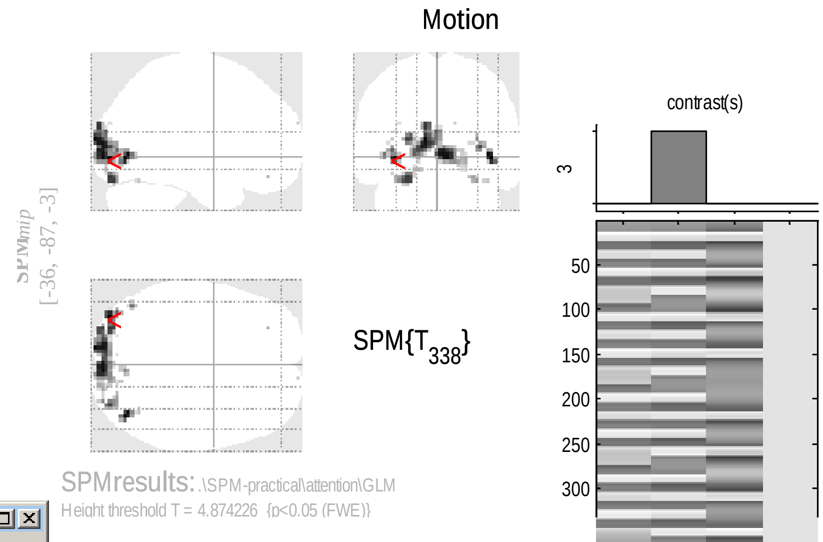
Results

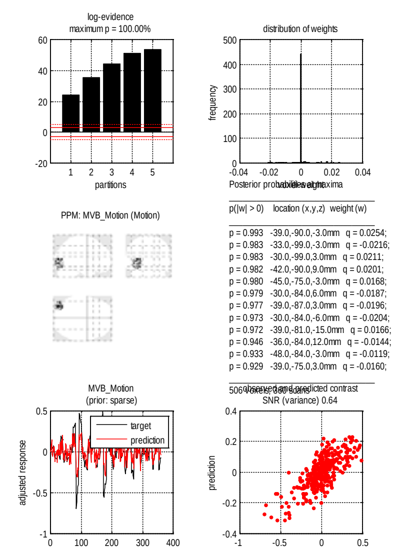


Friston et al. 2008 NeuroImage

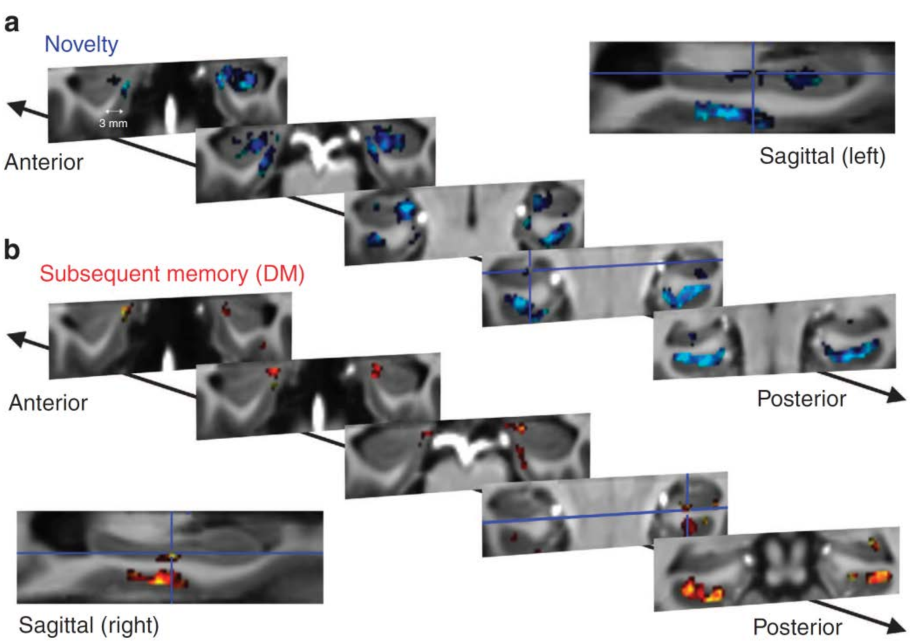
Multivariate Bayes in SPM

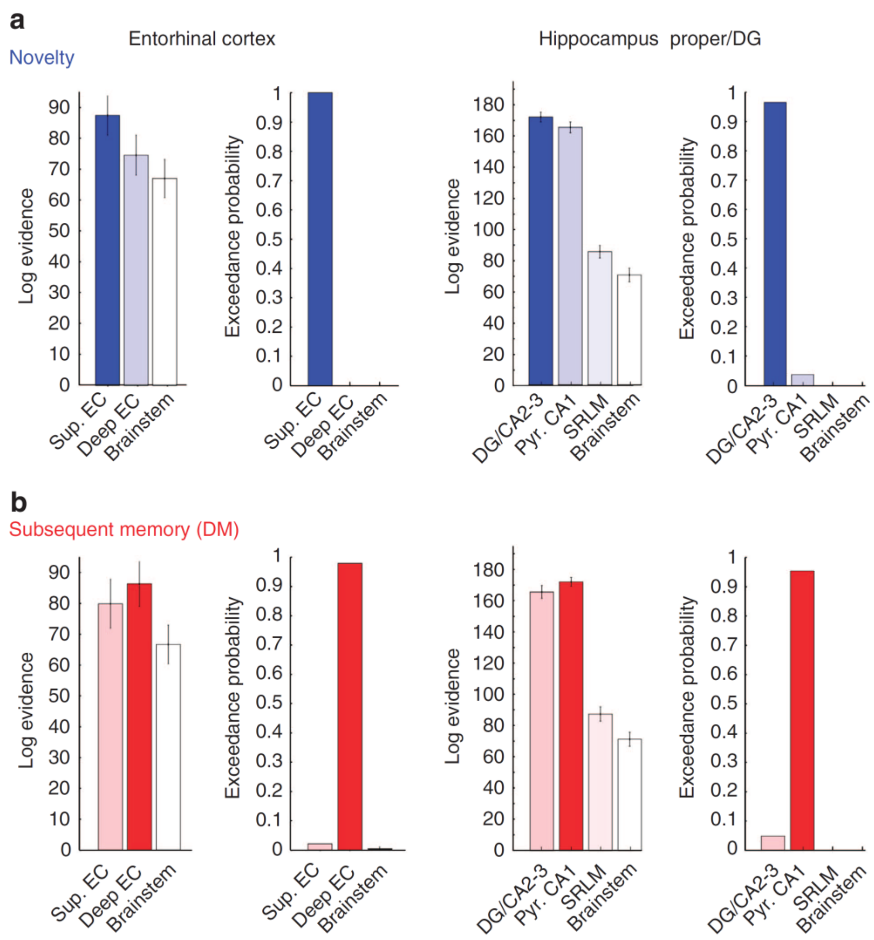






Laminar activity related to novelty and episodic memory





Maas et al. 2014 Nature Communications

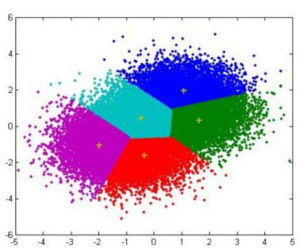
Motivation

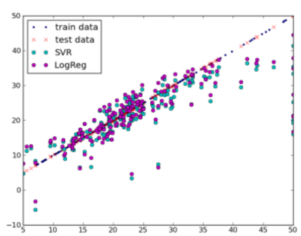
Modelling Principles

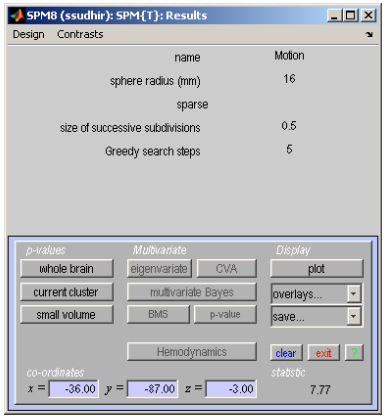
Learning from Data

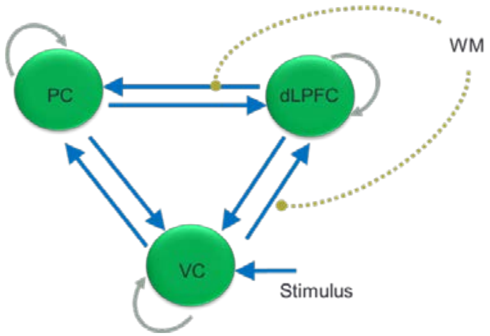
Multivariate Bayes in SPM

Generative Embedding

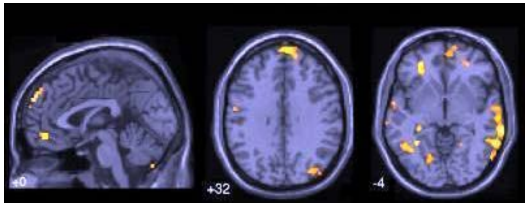






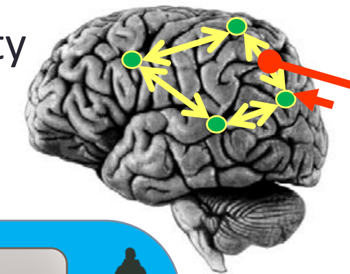


Classifying Groups of Subjects



Voxel activity

Connectivity



Subject 1
Subject 2
⋮
Subject N

Dynamic causal model (DCM)

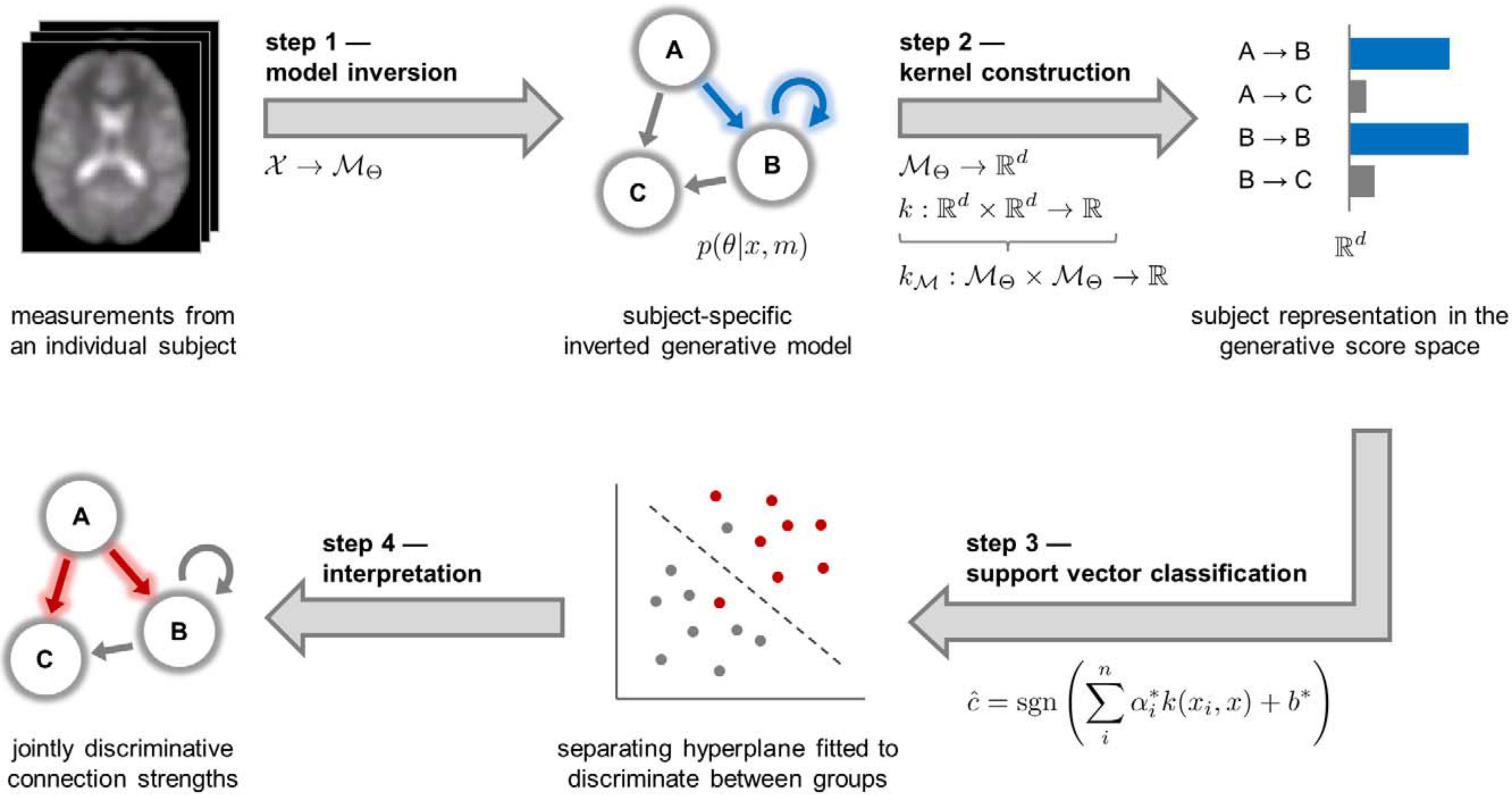
- High dimensionality
- Unusual cluster distributions
- Lack of interpretation

Subject 1
Subject 2
⋮
Subject N

Classification
Clustering

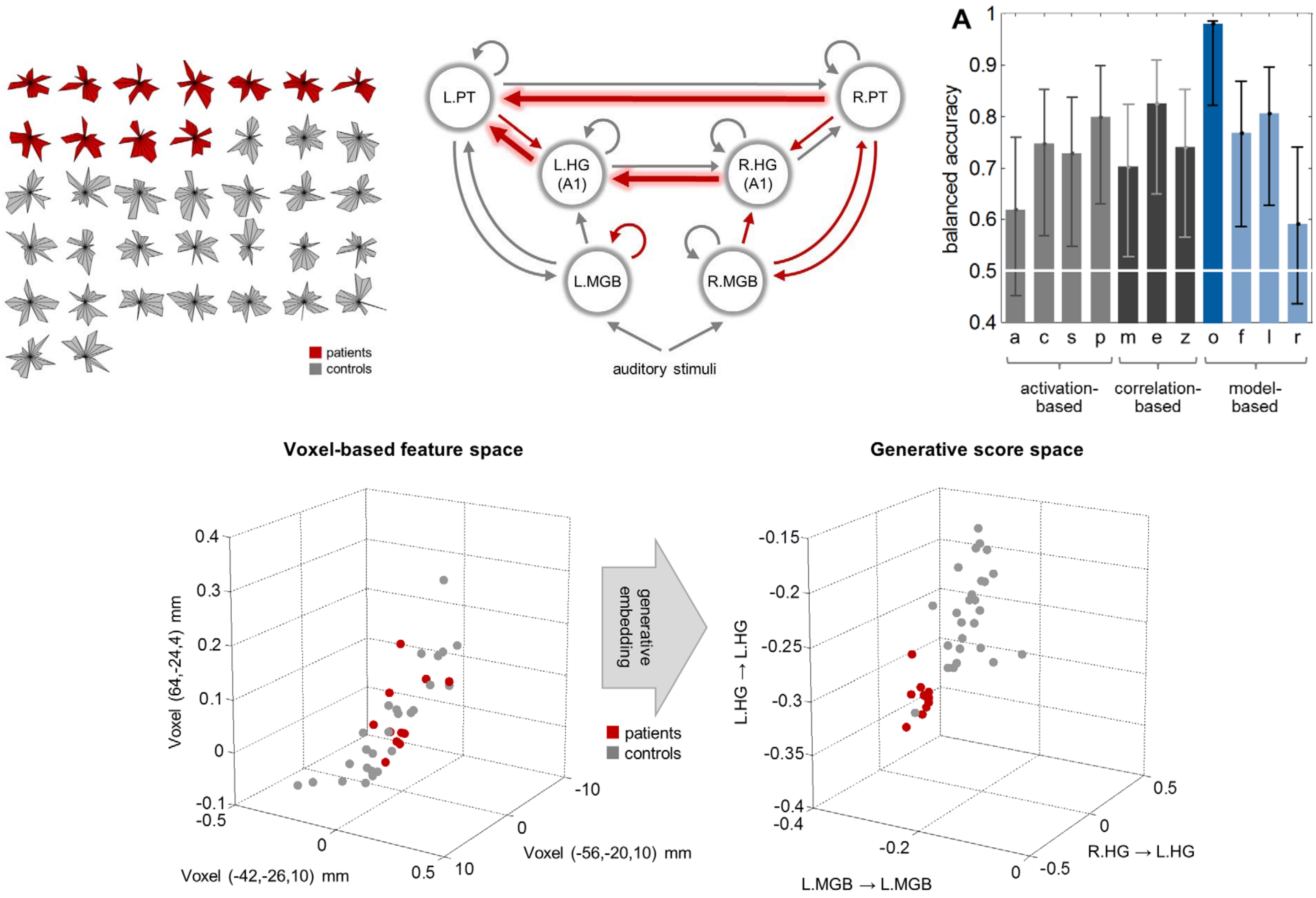
Group 1
Group 2

Generative Embedding



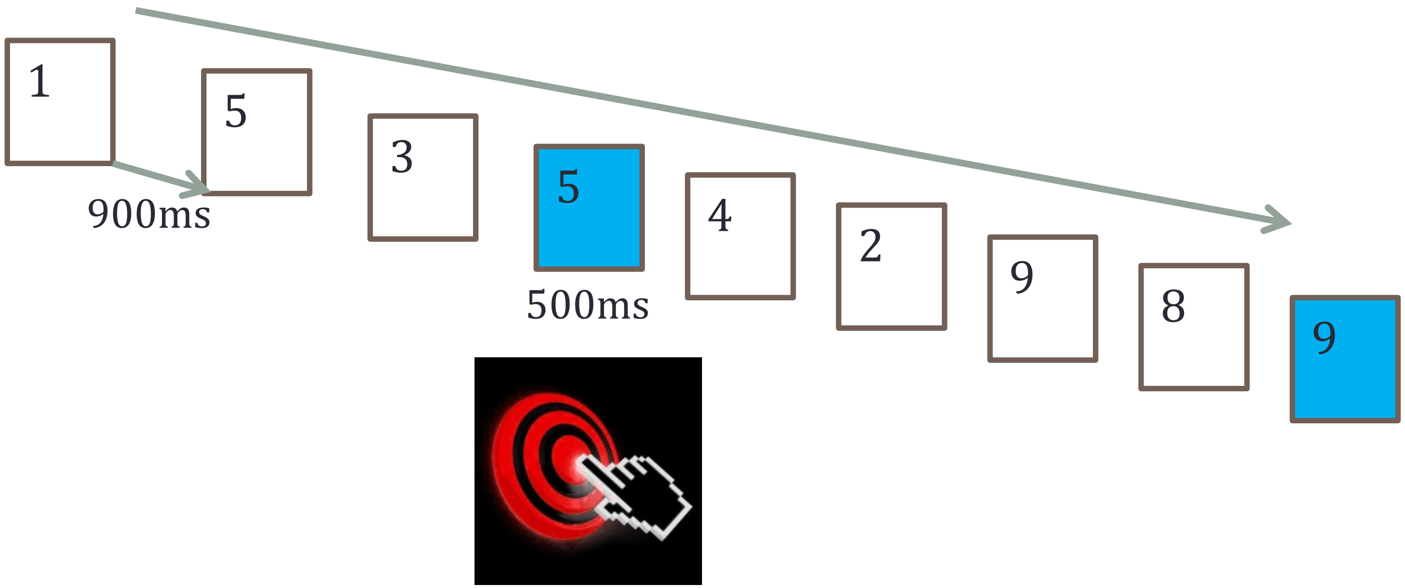
Brodersen et al. PLOS computation biology 2011.

DCM for speech processing



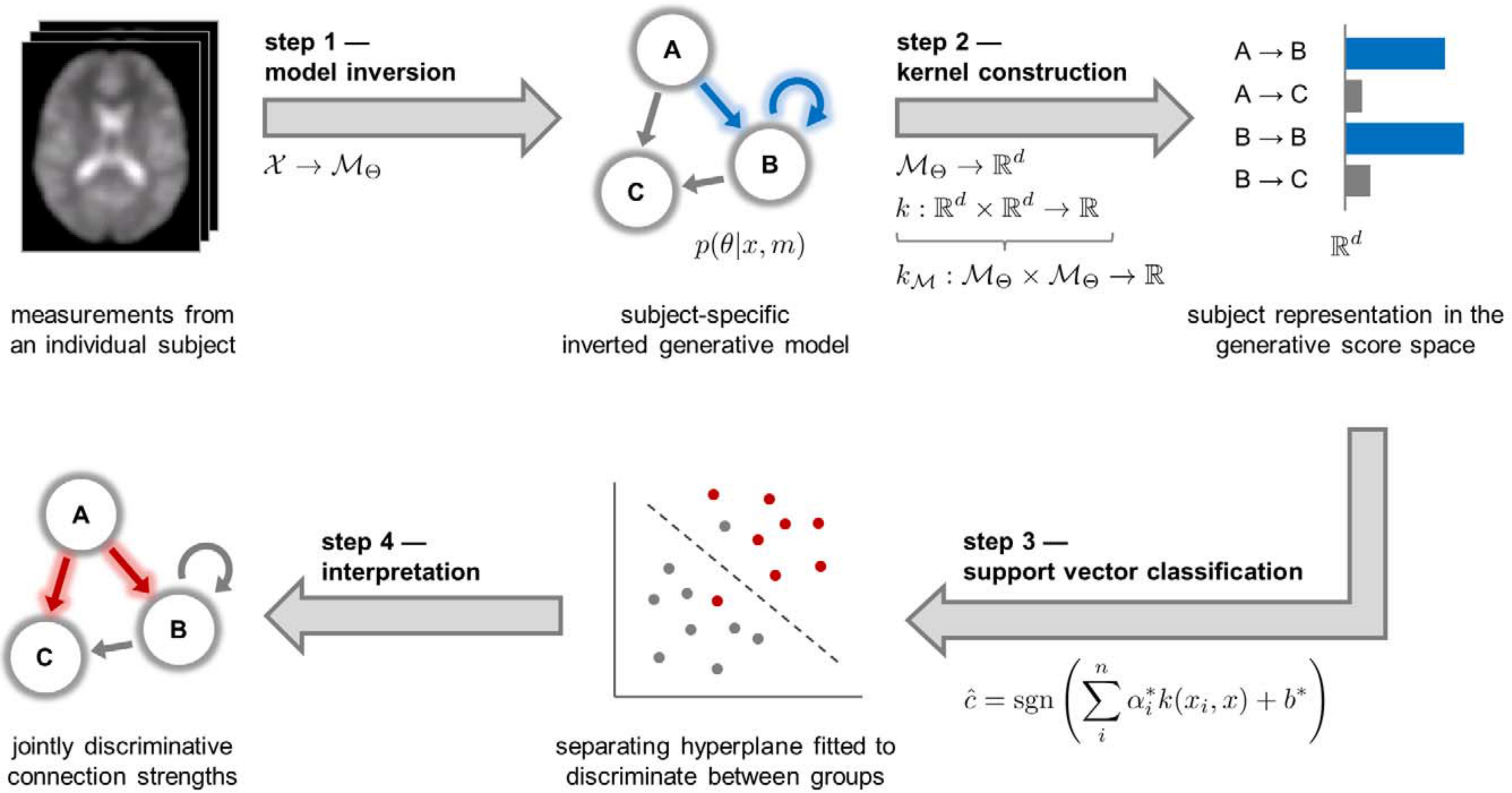
Working memory in Schizophrenia

- 41 Schizophrenia patients (DSM IV,ICD 10), 42 controls
- Visual numeric n-back working memory task



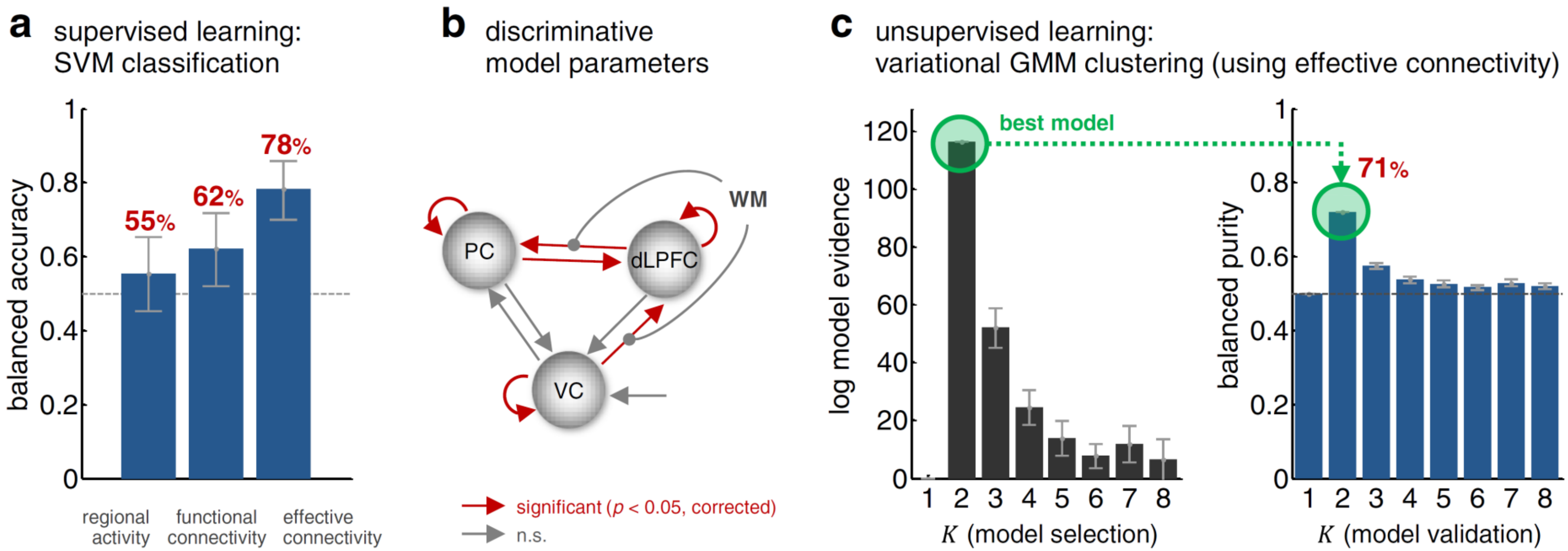
Deserno et al (2012) *The Journal of Neuroscience*

Model based clustering



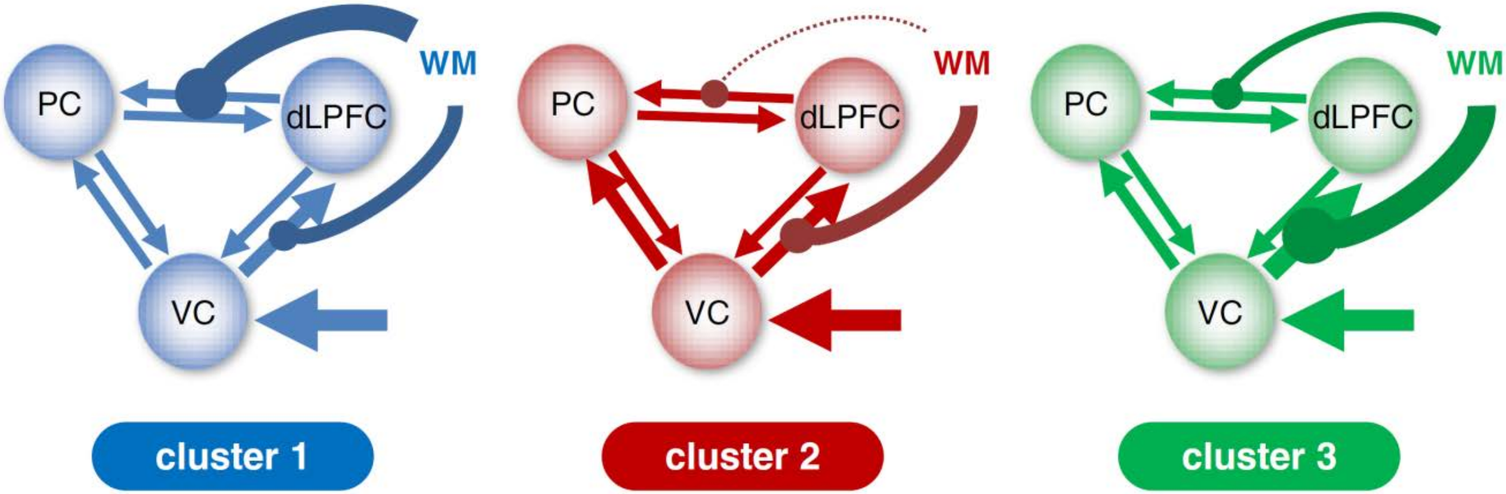
Brodersen et al 2014 Neuroimage

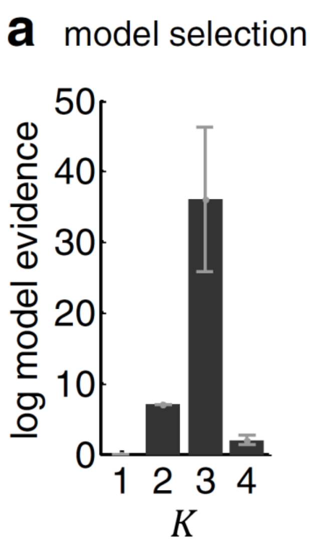
Results healthy vs. schizophrenia patients

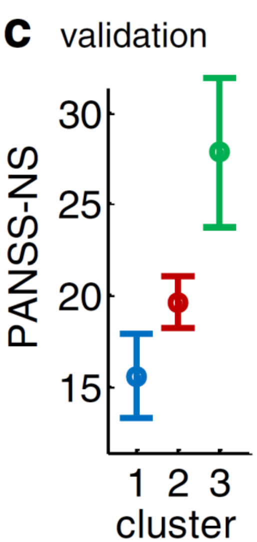


Brodersen et al 2014 Neuroimage

Within patients clustering







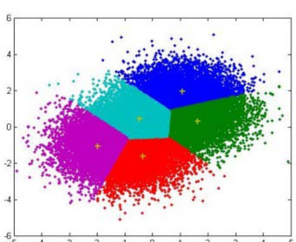
Brodersen et al 2014 Neuroimage

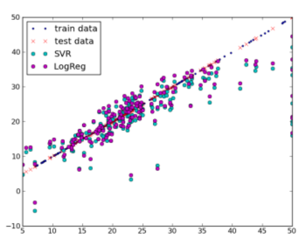
Be aware

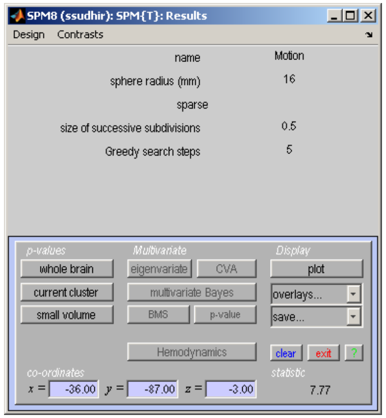
- Interpretation of decoding or classification results is difficult.
- The decoded information must be in the data, but in what features exactly is often hard to find out …

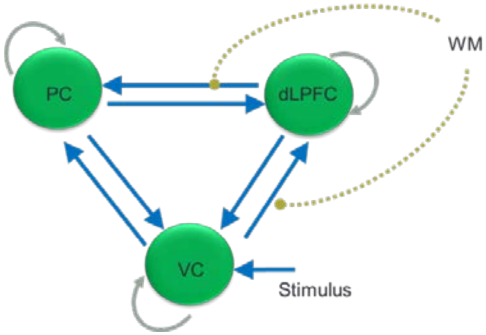
Summary

- Summary
  - Modelling Principles
  - Learning from Data
  - Multivariate Bayes in SPM
  - Generative Embedding









Acknowledgments

Many thanks to K.E. Stephan, Sudhir S. Raman and K. Brodersen for sharing their teaching material.