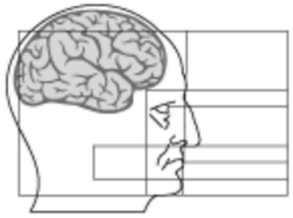
Machine Learning Techniques (Theory & Software)

Maria J. Rosa



MAX PLANCK UCL CENTRE
for Computational Psychiatry and Ageing Research

UCL

TNU Computational Psychiatry Course, 2016

Outline

Theory

Machine Learning framework

Classification

Regression

Linear models, kernels and non-linear models

Prediction error and bias-variance tradeoff

Cross-validation

Software packages

Software

PRoNTo

(Pattern Recognition for

Neuroimaging Toolbox)

Machine Learning

Machine learning is a branch of **artificial intelligence** concerned with **learning information from existing data** and focused on making **predictions to new data**.

Machine learning has an extremely broad application domain, from computer vision to stock market analysis.

Supervised learning – labeled data
Unsupervised learning – unlabeled data
Semi-supervised learning – some labeled and unlabeled data
Reinforcement learning

Machine Learning for Psychiatry

ORIGINAL RESEARCH ARTICLE

Front. Hum. Neurosci., 25 October 2010 | http://dx.doi.org/10.3389/fnhum.2010.00192

A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia

Honghui Yang[1,2]*, Jingyu Liu[2,3], Jing Sui[2,3], Godfrey Pearlson[4,5] and Vince D. Calhoun[2,3,5,6]

ORIGINAL RESEARCH ARTICLE

Front. Psychiatry, 01 June 2012 | http://dx.doi.org/10.3389/fpsyt.2012.00053

Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls

Deanna Greenstein[1]*, James D. Malley[2], Brian Weisinger[1], Liv Clasen[1] and Nitin Gogtay[1]

Archival Report

Predicting the Naturalistic Course of Major Depressive Disorder Using Clinical and Multimodal Neuroimaging Information: A Multivariate Pattern Recognition Study

Lianne Schmaal[a], , , Andre F. Marquand[b, c], Didi Rhebergen[d], Marie-José van Tol[e], Henricus G. Ruhé[e, f], Nic J.A. van der Wee[g], Dick J. Veltman[a], Brenda W.J.H. Penninx[a, d]

Quantitative forecasting of PTSD from early trauma responses: A Machine Learning application

Isaac R. Galatzer-Levy[b], , , Karen-Inge Karstoft[b, c], Alexander Statnikov[d, e], Arieh Y. Shalev[a, b]

Original Article

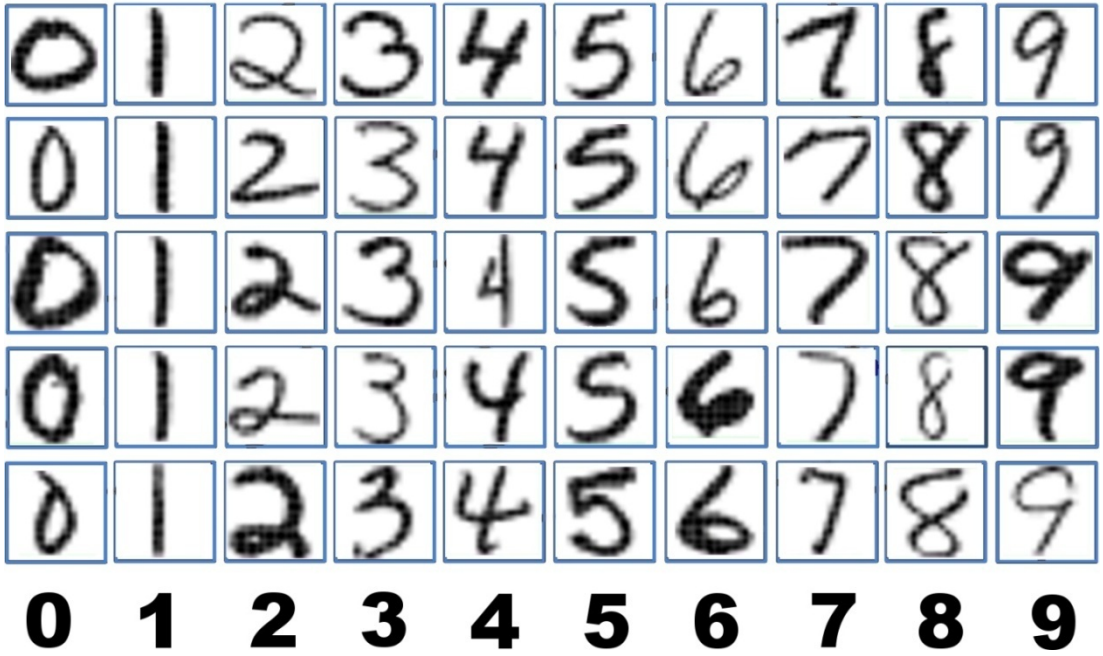
Citation: *Translational Psychiatry* (2012) **2**, e100; doi:10.1038/tp.2012.10
Published online 10 April 2012

Use of machine learning to shorten observation-based screening and diagnosis of autism

Open

D P Wall[1,2], J Kosmicki[1], T F DeLuca[1], E Harstad[3] and V A Fusaro[1]

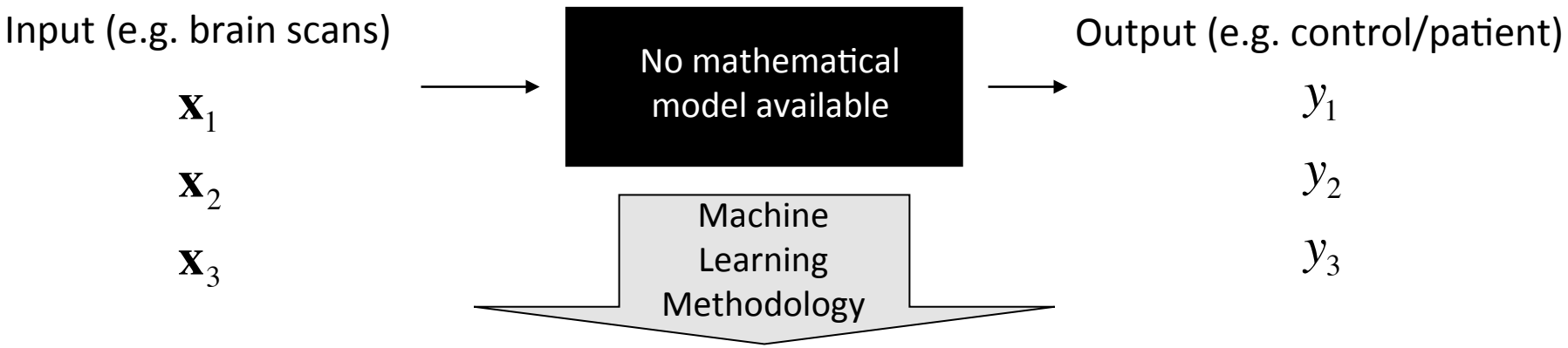
Machine learning framework



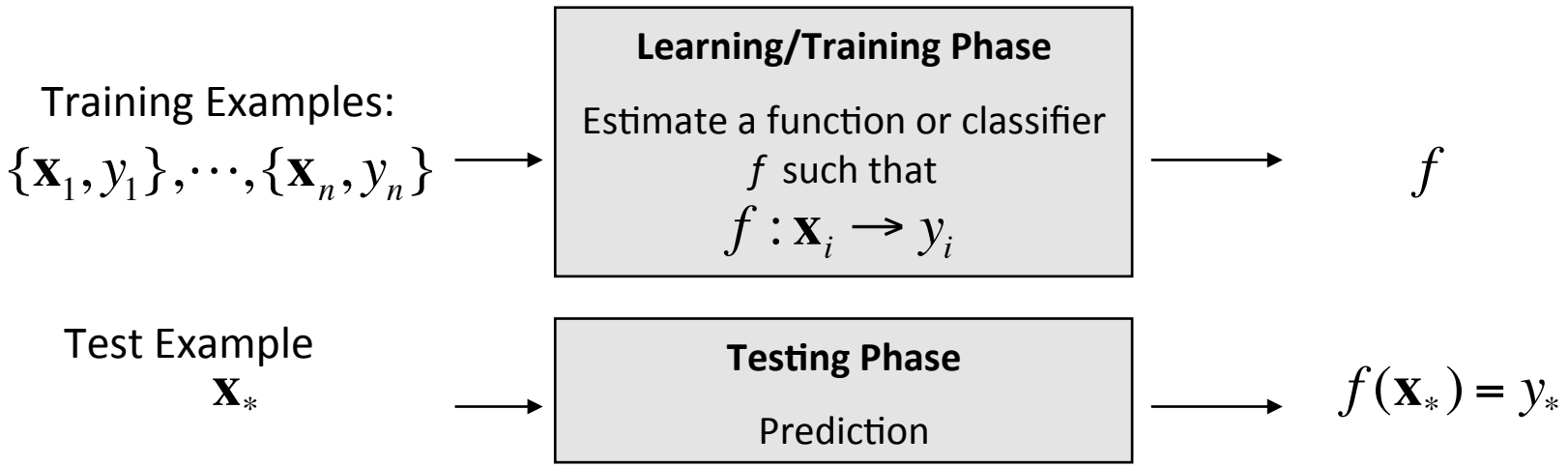
$$f : \mathbf{X} \rightarrow \mathbf{y}$$

$$f : \mathbf{x}_* \rightarrow y_*$$

Machine learning framework

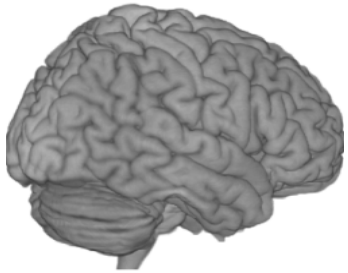


Computer-based procedures that learn a function from a series of examples

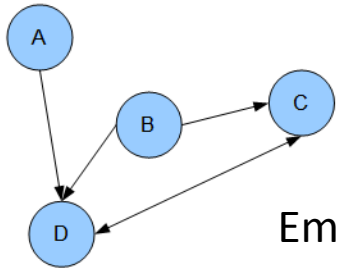


Features, **x**

Whole brain volume



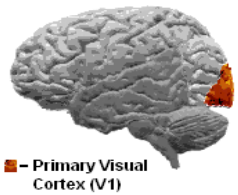
Model

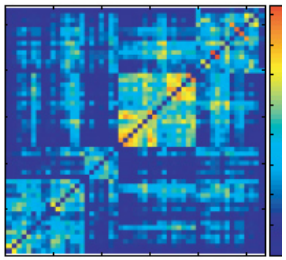


Embedding

Feature Vector, **x**

Region of interest (ROI)

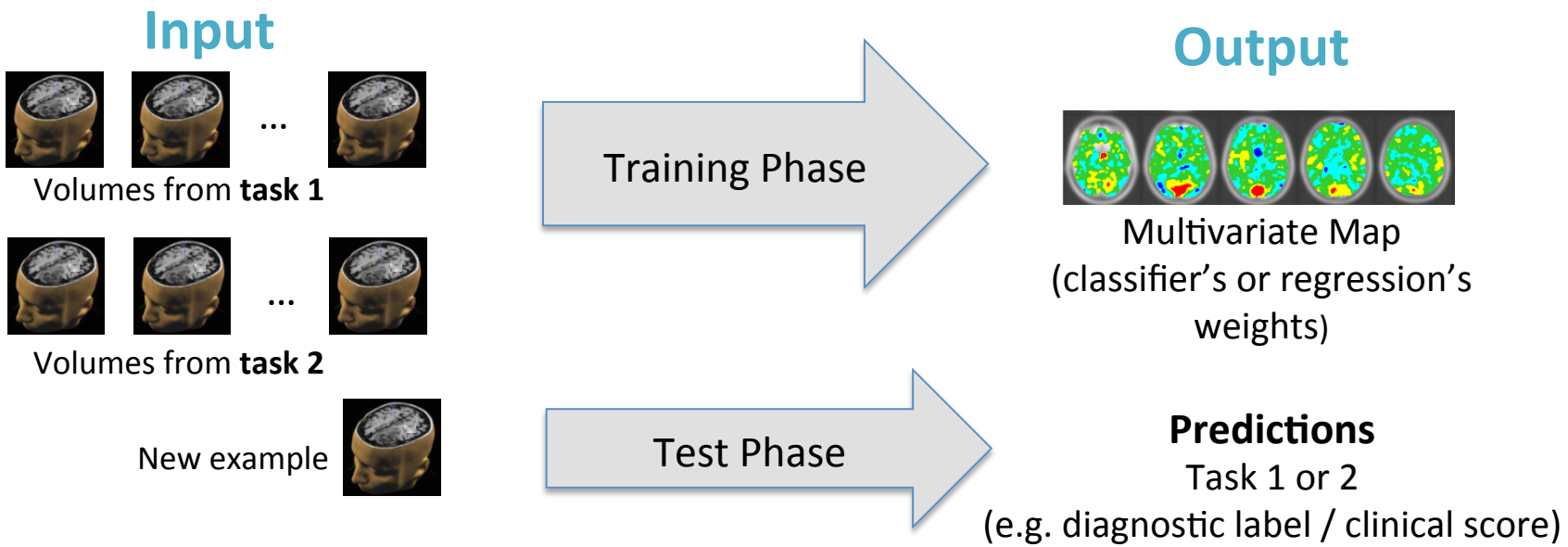




Matrices
(e.g. connectivity)

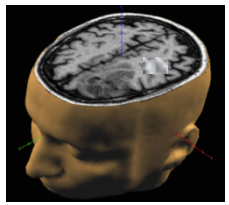
Classification framework

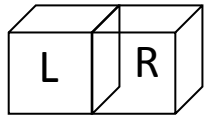
Neuroimaging

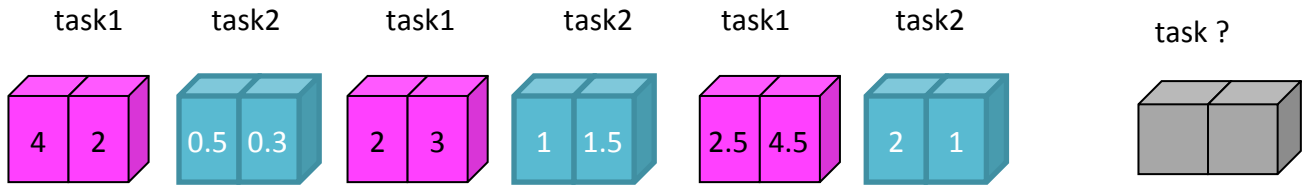


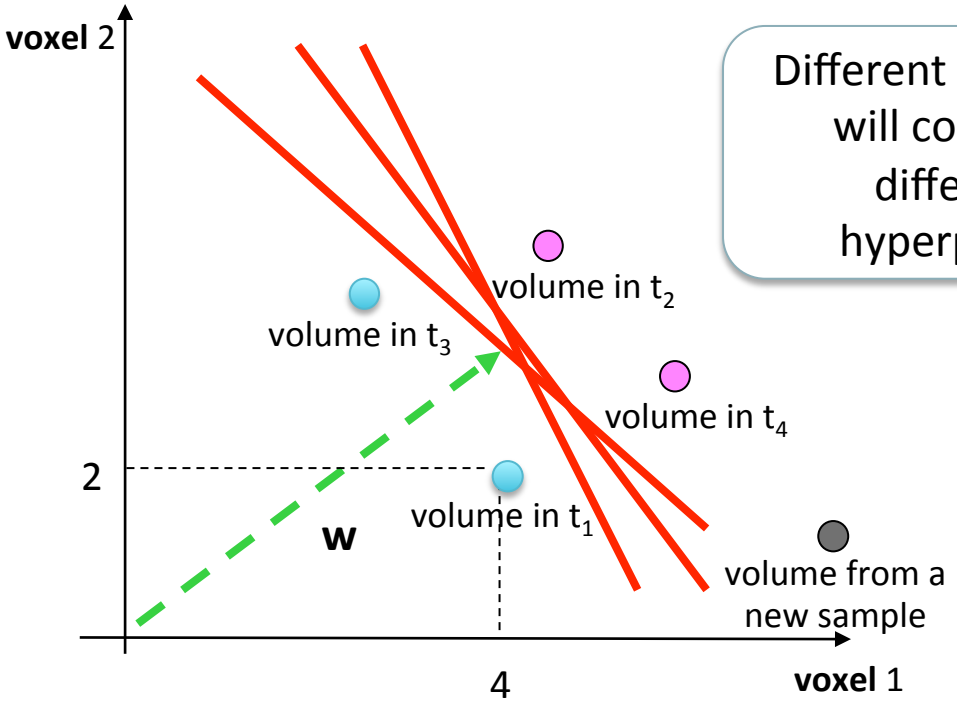
Classification framework

2D example

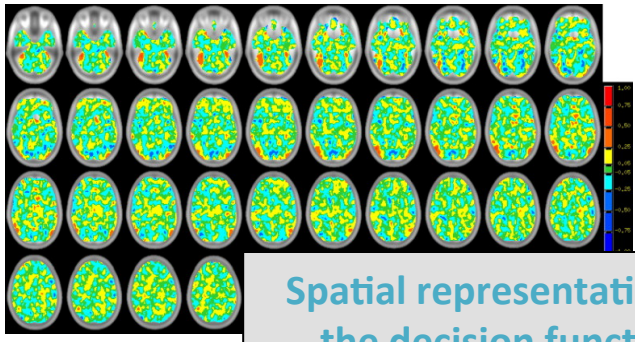








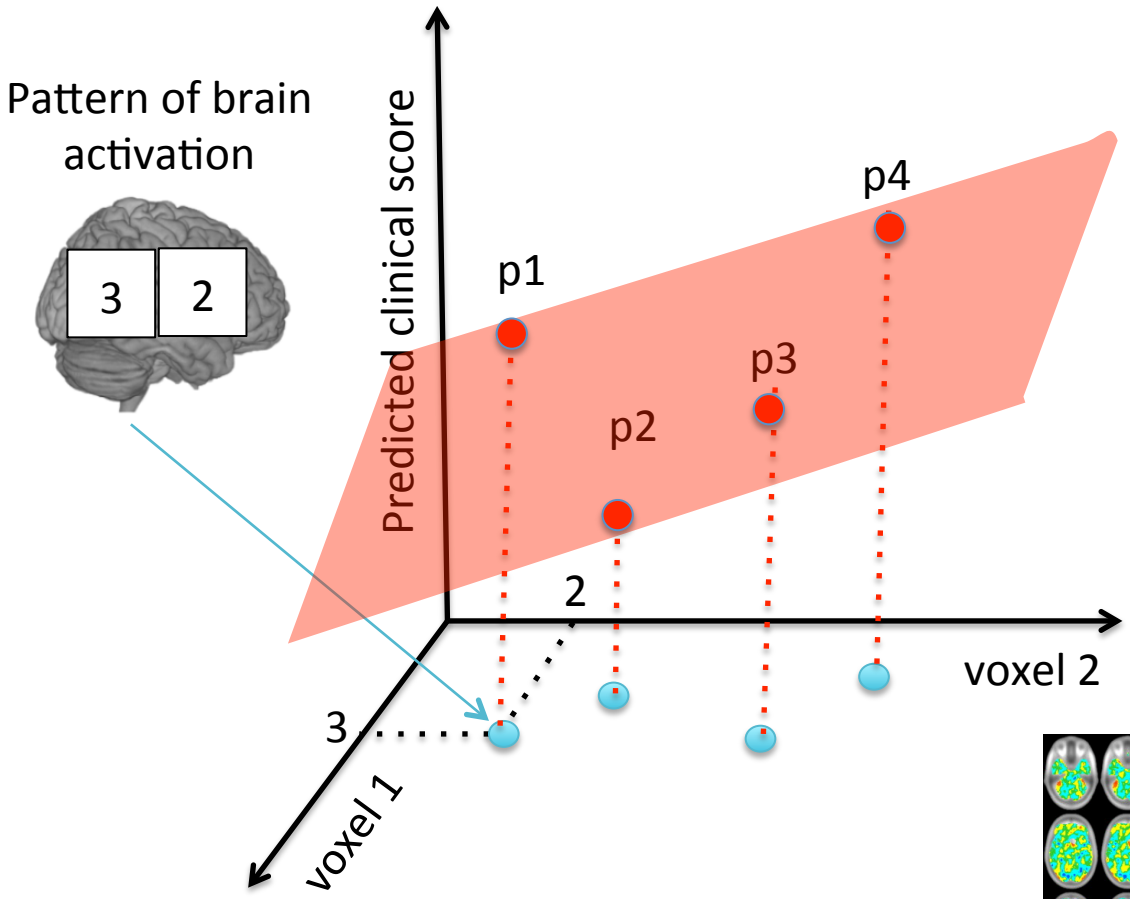
Different classifiers will compute different hyperplanes



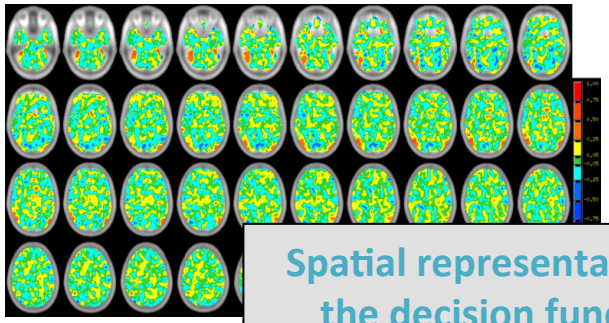
Spatial representation of the decision function

Regression framework

2D example

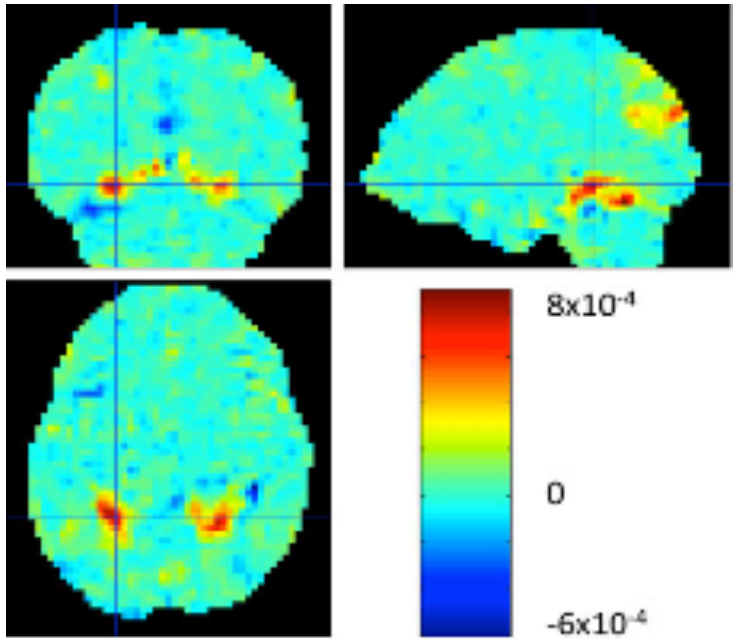


Linear regression models are also parameterized by a weight vector **w** and a bias term b.



Spatial representation of the decision function

Weight vectors



- ✓ It is a spatial representation of the predictive function.
- ✓ Shows the contribution of each feature/ voxel for the prediction.
- ✓ Multivariate pattern -> All voxels with weights different from zero contribute to the final prediction (no arbitrary threshold should be applied).
- ✓ **No local inference (like SPM)!**

Linear models

- Linear models (e.g. hyperplanes in case of classification) are parameterized by a weight vector **w** and a bias term *b*.
- The weight vector can be expressed as a linear combination of training examples $\mathbf{x}_i$

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

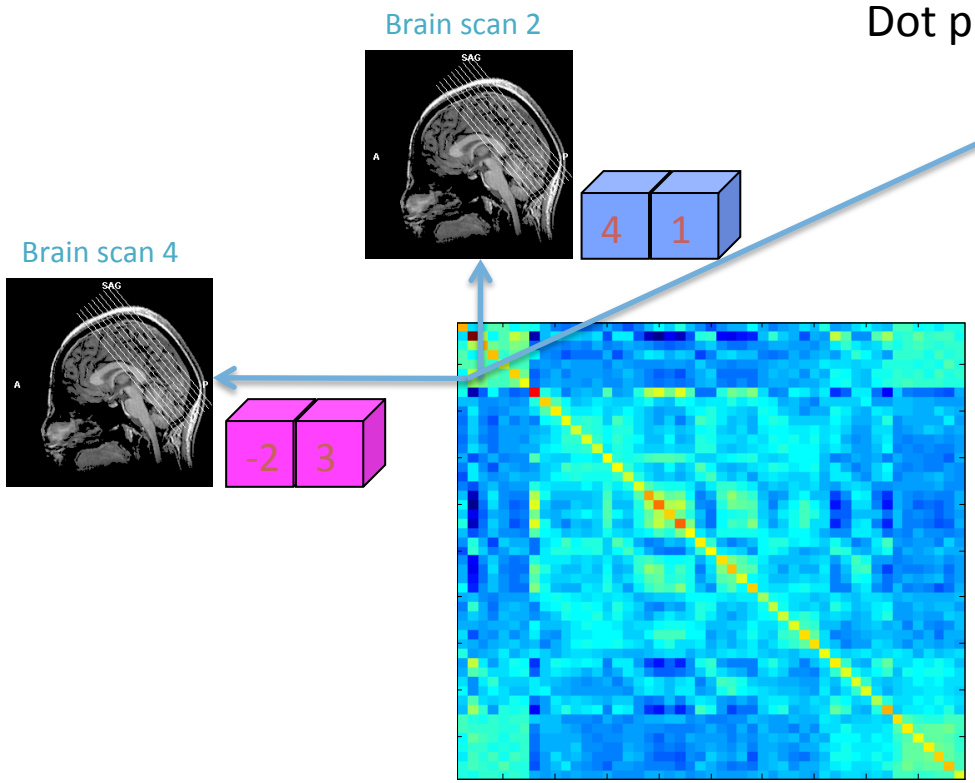
- The general equation for making predictions for a test example $\mathbf{x}_*$ is:

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

- $f(x_*) > 0$ -> Class 1 (e.g. patients or task 1)
- $f(x_*) < 0$ -> Class 2 (e.g. controls or task 2)
- $f(x_*)$ value (regression)

Probabilistic approaches:
p(class 1) = 1 – p(class 2)

Kernel Matrix (similarity measure)



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

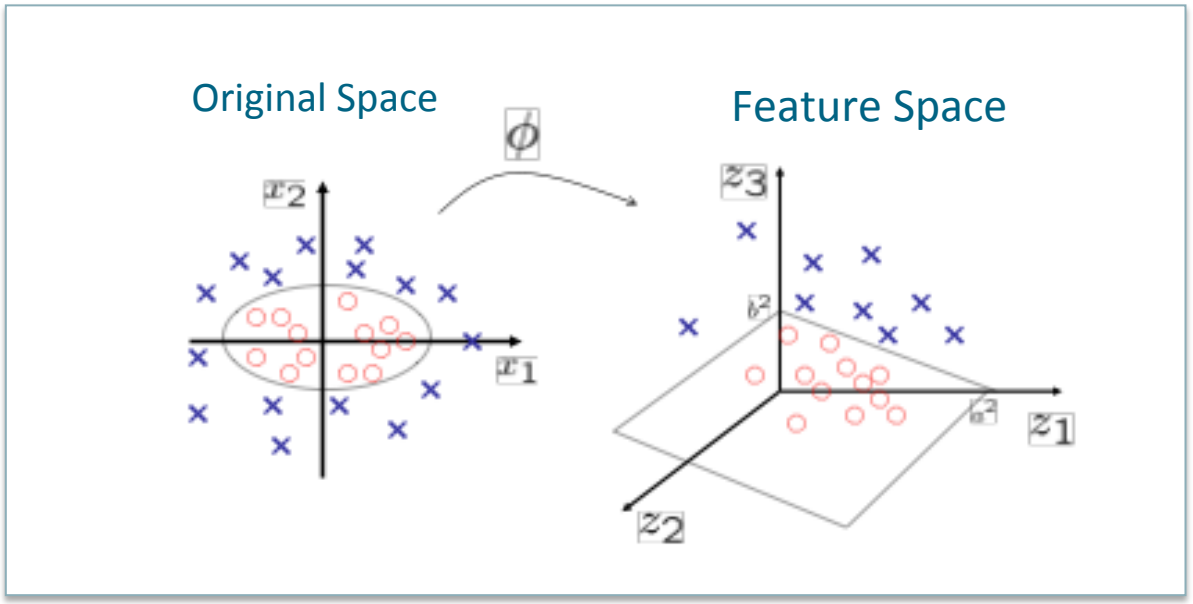
$$f(x_*) = \sum_{i=1}^{N} \alpha_i \phi(x_i) \cdot \phi(x_*) + b$$

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

- A simple type of similarity measure between two vectors is a dot product (linear kernel).
- Kernel is a function that, for given two pattern **x** and **x***, returns a real number characterizing their similarity.

Nonlinear models

$$K(x,z) = \langle \phi(x), \phi(z) \rangle$$



Nonlinear kernels are used to map the data to a higher dimensional space as an attempt to make it linearly separable.

Kernel trick:

Enable the computation of similarities in the feature space without having to compute the mapping explicitly.

Examples of Kernel Methods

- Support Vector Machines (SVM)
- Gaussian Processes (GPs)
- Kernel Ridge Regression
- Kernel Fisher Discriminant
- Relevance Vector Regression
- Multi-kernel SVM

Measures of performance for classification

- Accuracy statistics can be shown in a confusion matrix :

  Class 1 accuracy (p_0) = A/(A+B)

  Class 2 accuracy (p_1) = D/(C+D)

  Accuracy (p) = (A+D)/(A+B+C+D)

		$\hat{\omega}$	
		ω_0	ω_1
truth	ω_0	A	B
	ω_1	C	D

- Medical research (ω_0 =healthy, ω_1 =disease):

Sensitivity: "What is the probability that the disease is PRESENT and that I detect it?"): D/(C+D) = TP/(FN+TP)

Specificity: "disease is ABSENT and I report no disease": A/(A+B) = TN/(FP+TN)

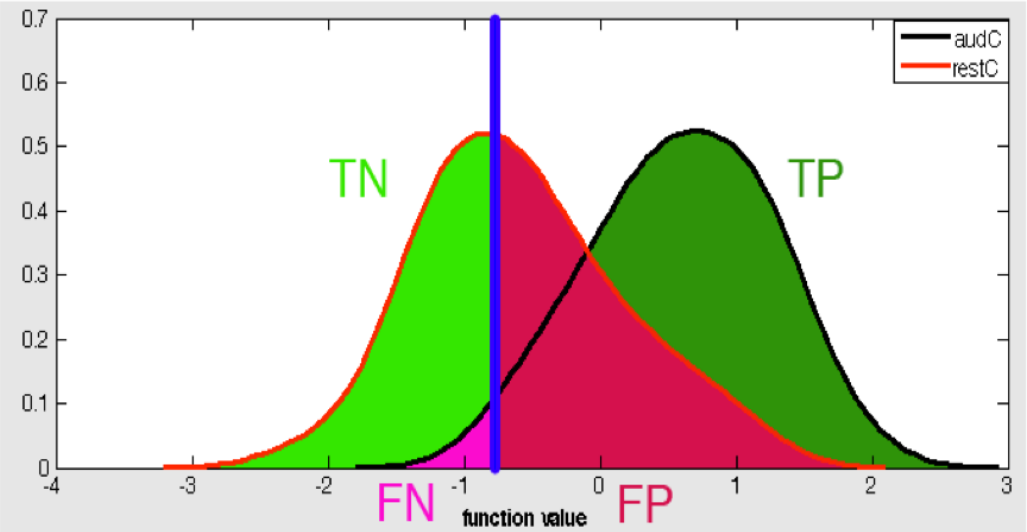
Perfect: B=C=0. Be suspicious if this happens!

Random: A=B=C=D. Same as flipping a coin.

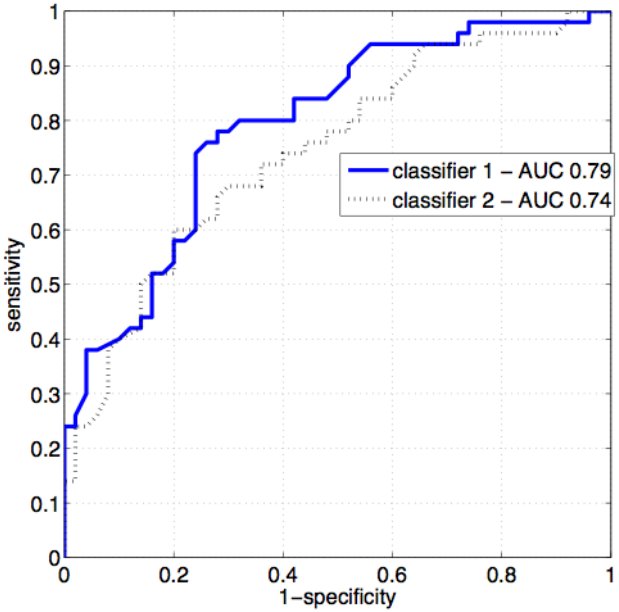
Measures of performance for classification

Sensitivity / specificity

Increasing sensitivity can only come at the cost of decreasing specificity, and vice-versa



ROC Curve



The **Receiver Operating Characteristic (ROC) curve** is a good way of seeing the sens/spec tradeoff over the operating range of a classifier.

Measures of performance for regression

Correlation

$$CORR = \frac{\sum_n (y_n - \mu_y)(f(\boldsymbol{x}_n) - \mu_f)}{\{\sum_n (y_n - \mu_y)^2 \sum_n (f(\boldsymbol{x}_n) - \mu_f)^2\}^{\frac{1}{2}}}$$

Coefficient of determination

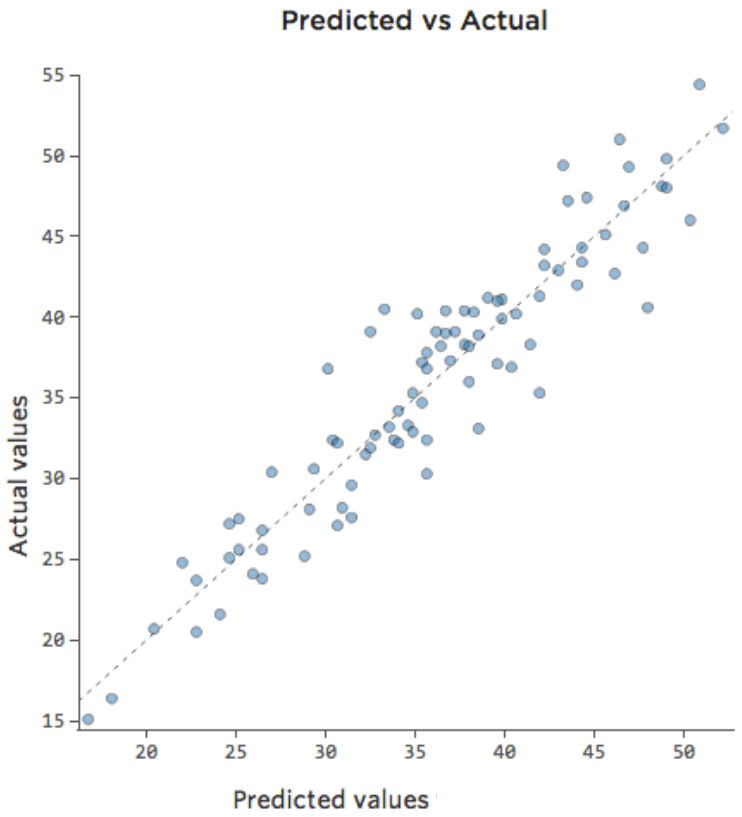
$$R^2 = CORR^2$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{N} \sum_n (y_n - f(\mathbf{x}_n))^2$$

Normalised MSE

$$Norm.MSE = \frac{MSE}{(y_{max} - y_{min})}$$



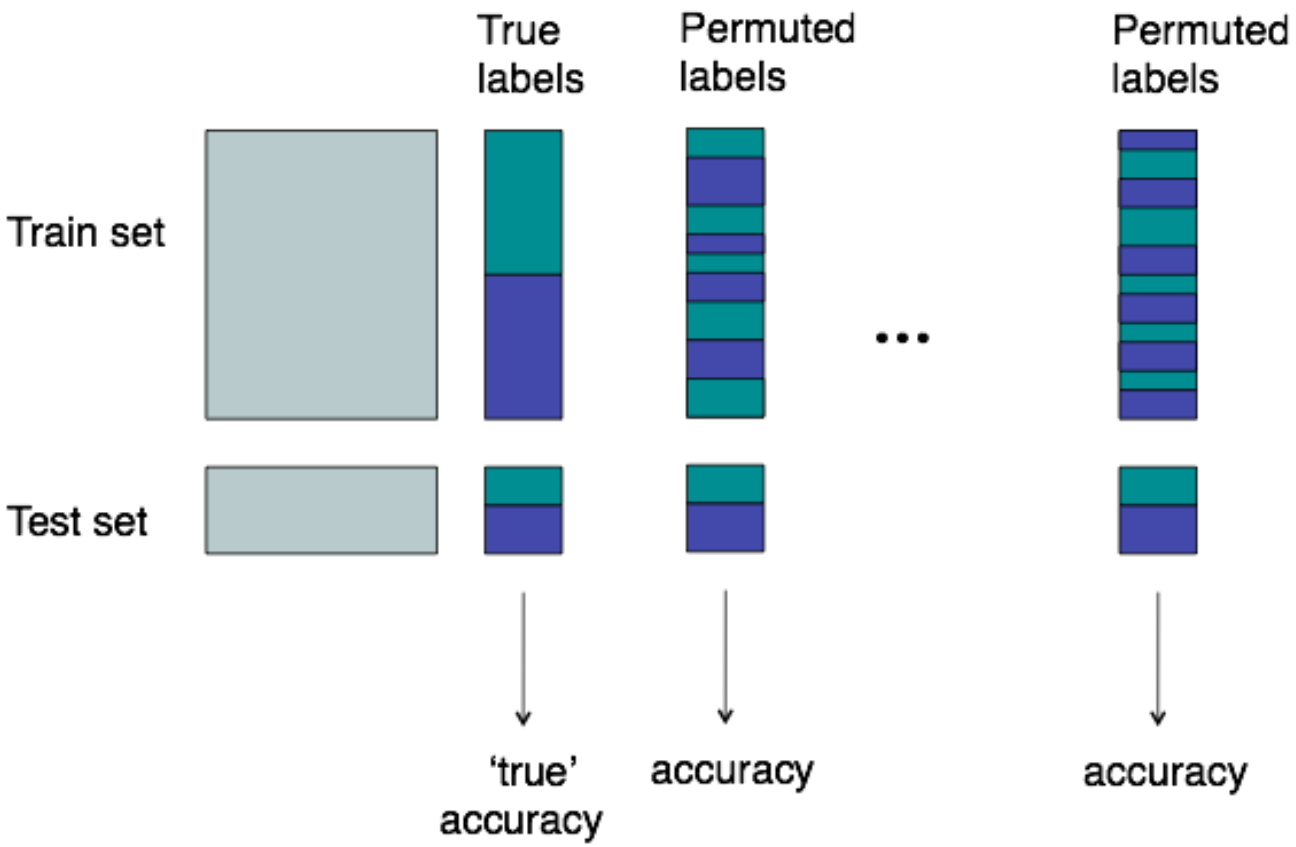
Model evaluation

57% accuracy: significant?
0.57 correlation: significant?

- **Parametric approaches**:
  - Binomial distribution (k successes / n attempts)
  - T-tests (correlation != 0)
  - Assumption that the test examples are drawn independently (not true for many fMRI designs)
  - Cross-validation can change underlying distribution

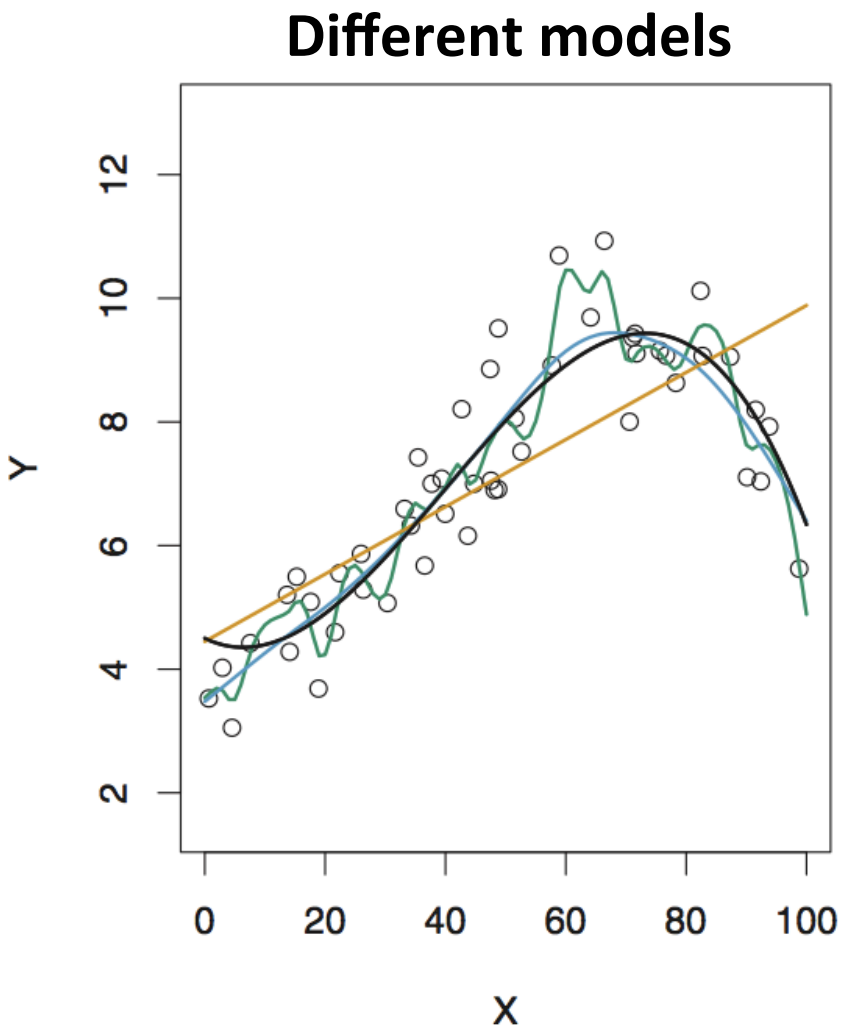
- **Non-Parametric approaches**:
  - Permutation tests
  - Do not assume distribution for the data
  - Computational cost

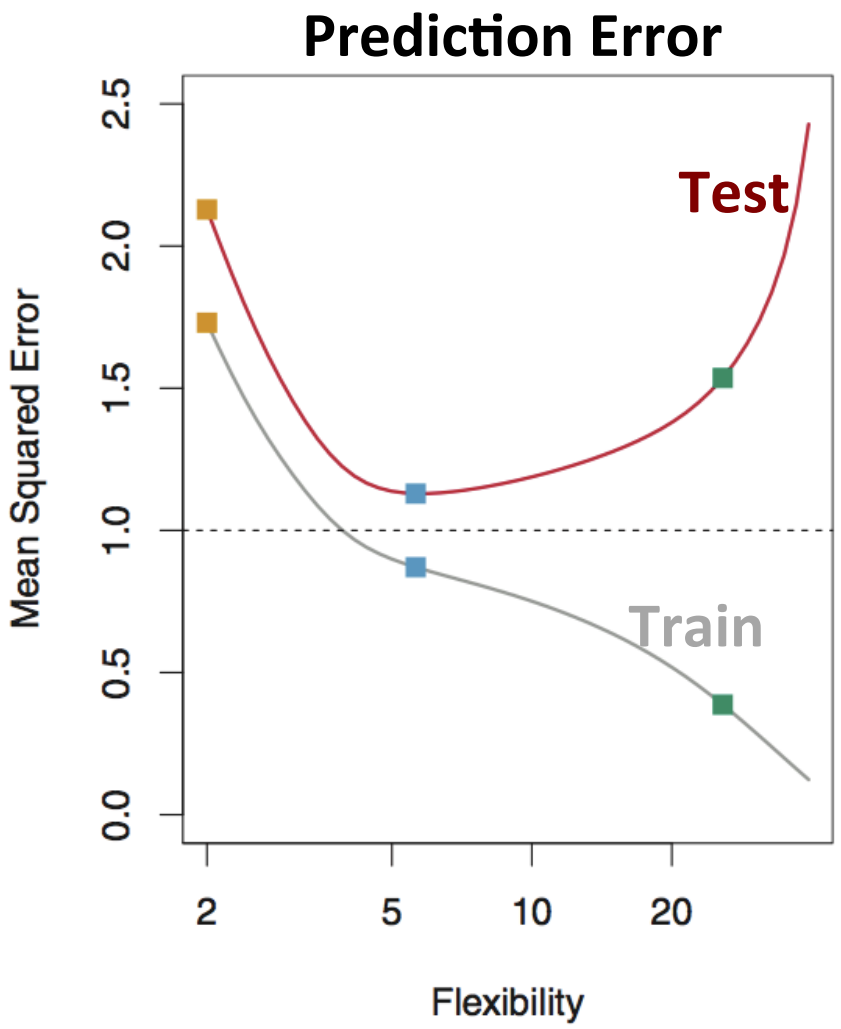
Permutation tests



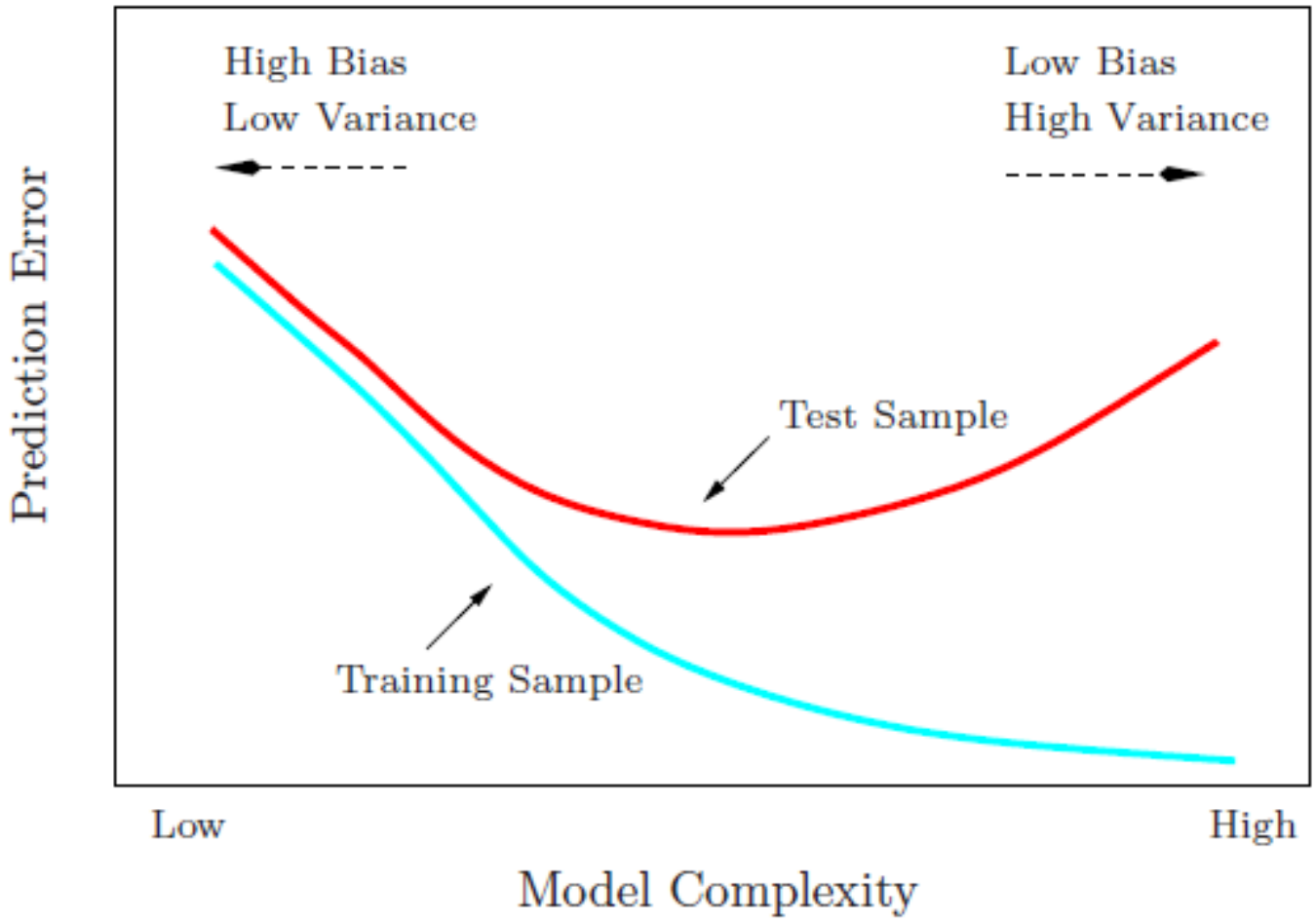
P-value: $$\frac{1}{M}\sum_{m}^{M}(p_m^{perm} \geq p^{real})$$

Prediction error





Prediction error / Bias-Variance tradeoff



Cross-validation (CV)

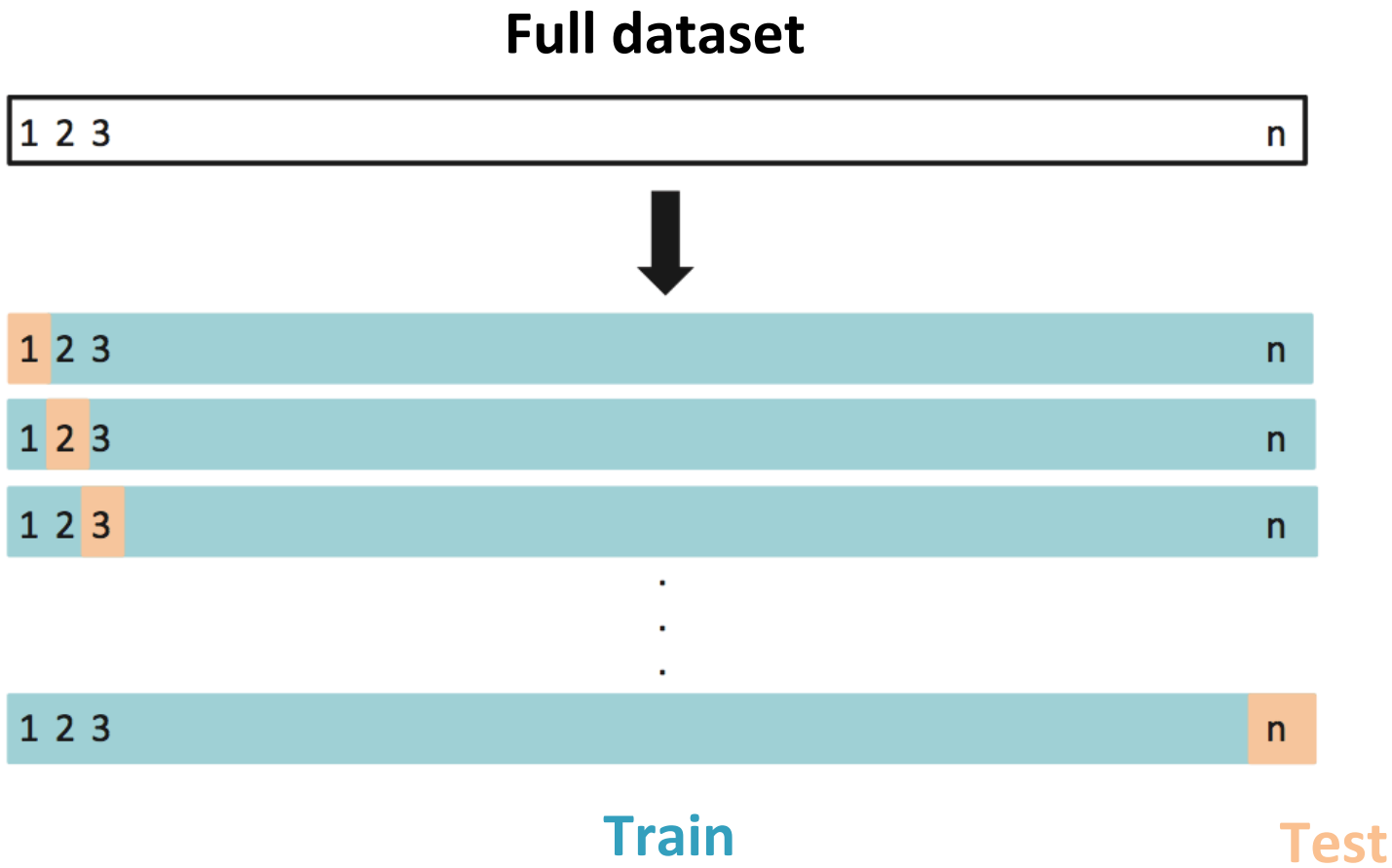
- Validation set approach

Drawbacks:

- Uses few observations and tends to overestimate the test error
- Test error estimates are highly variable

Cross-validation

Leave-one-out (LOO)



Leave each sample out for testing and use the rest for training. Repeat n times.

Cross-validation (CV)

LOO-CV

Main advantages:

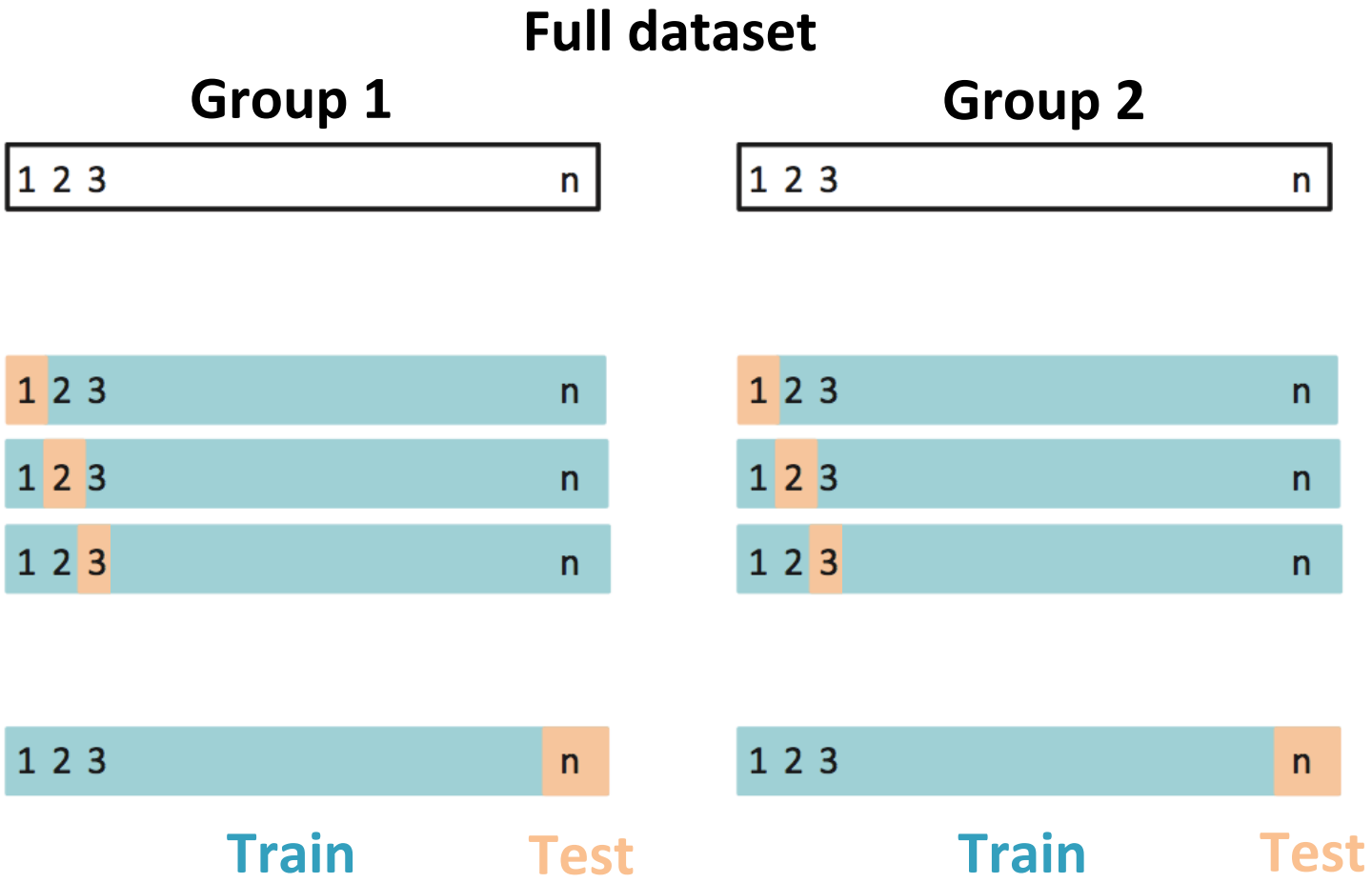
- Better use of data than half-split approach for small sample-sized data
- Almost unbiased test error estimate

Main disadvantages:

- Computationally intensive
- Test error estimate has high variance

Cross-validation

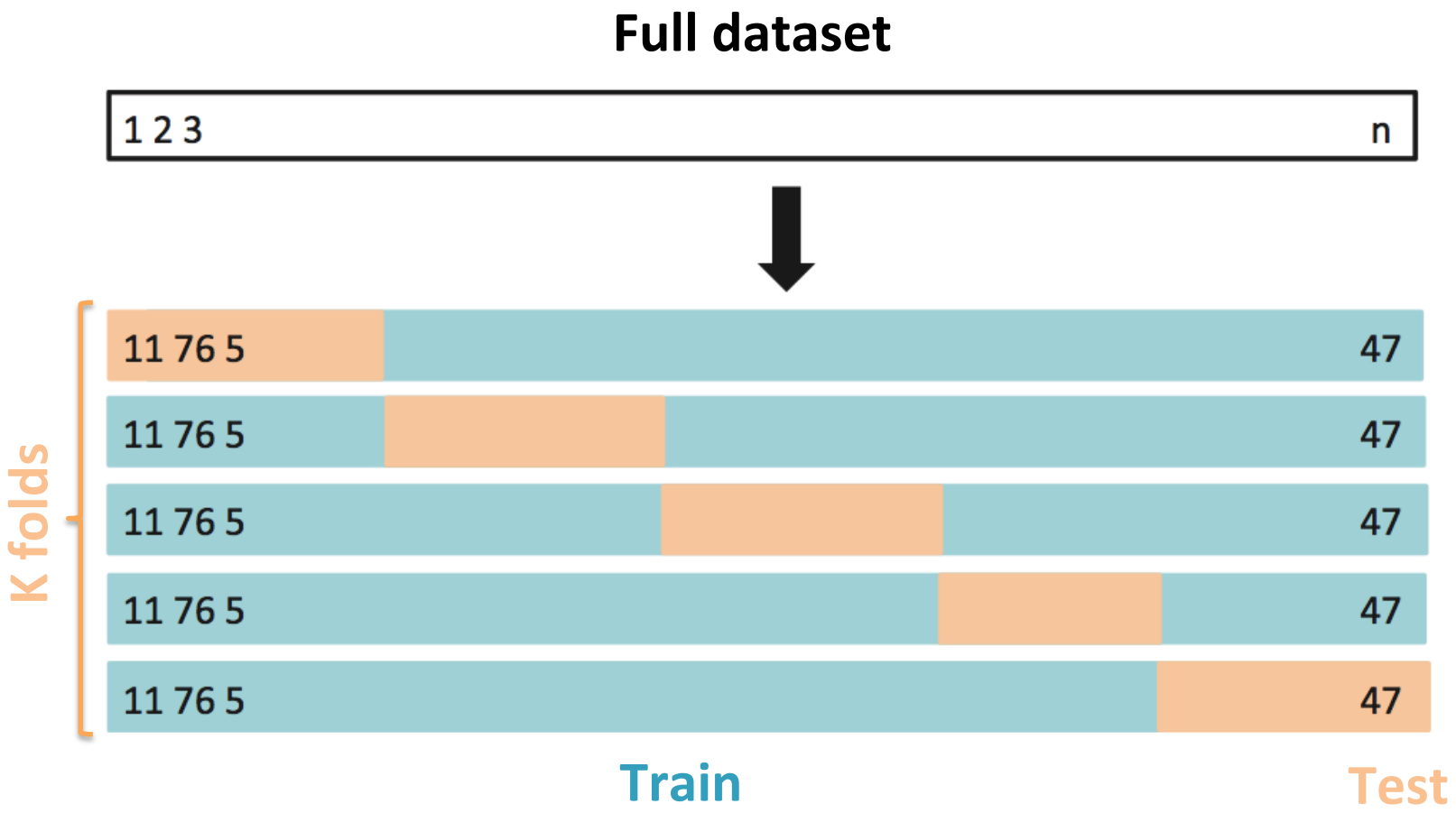
Leave-one-sample-per-group-out (LOSGO)



If subjects/samples in each group are paired (e.g. repeated measures)

Cross-validation

K-fold cross validation



Cross-validation (CV)

- K-fold CV

Main advantages:

- Test error estimate has less variance than LOO-CV
- Computationally less intensive

Main disadvantages:

- Higher bias of test error estimate than LOO-CV

Common *k* choices: 5 and 10

Nested cross-validation

- Problem: use CV to select best model and assess model performance (test error)
- Solution:

  Run CV inside CV for model or feature selection / Bayesian Models

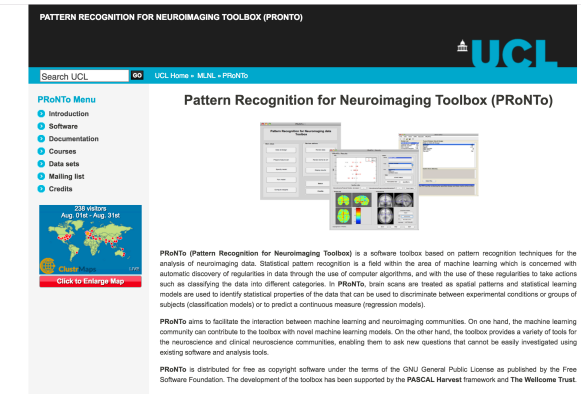
Take-home messages

- Always separate data intro training and testing sets
- Use cross-validation
- Be careful with correlated data (e.g. fMRI)
- Use nested cross-validation for model or feature selection
- Use permutation tests to assess significance of performance measure

Machine learning packages (neuroimaging)

- Princeton's MVPA
- The Decoding Toolbox
- KCL's PROBID
- PyMVPA
- Scikit-learn and NiLearn
- CoSMoMVPA
- **PRoNTo**

www.mlnl.cs.ucl.ac.uk/pronto/



Recommended reading

- James et al., *Introduction to Statistical Learning*, Springer, 2014.
- Duda et al., *Pattern Recognition*, Wiley, 2001.
- Hastie et al., *The elements of statistical learning*, Springer, 2009.
- Pereira et al., *Machine learning classifiers and fMRI: A tutorial overview*, NeuroImage 45, 2009.
- Kriegeskorte et al., *Circular analysis in systems neuroscience: the dangers of double dipping*, Nature Neuroscience 12, 2009.
- Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, IJCAI, 1995.

Thank You

QUESTIONS?

m.rosa@ucl.ac.uk