Pattern Recognition: validation & inference

Christophe Phillips

slides from **Jessica Schrouff**

Course 2021

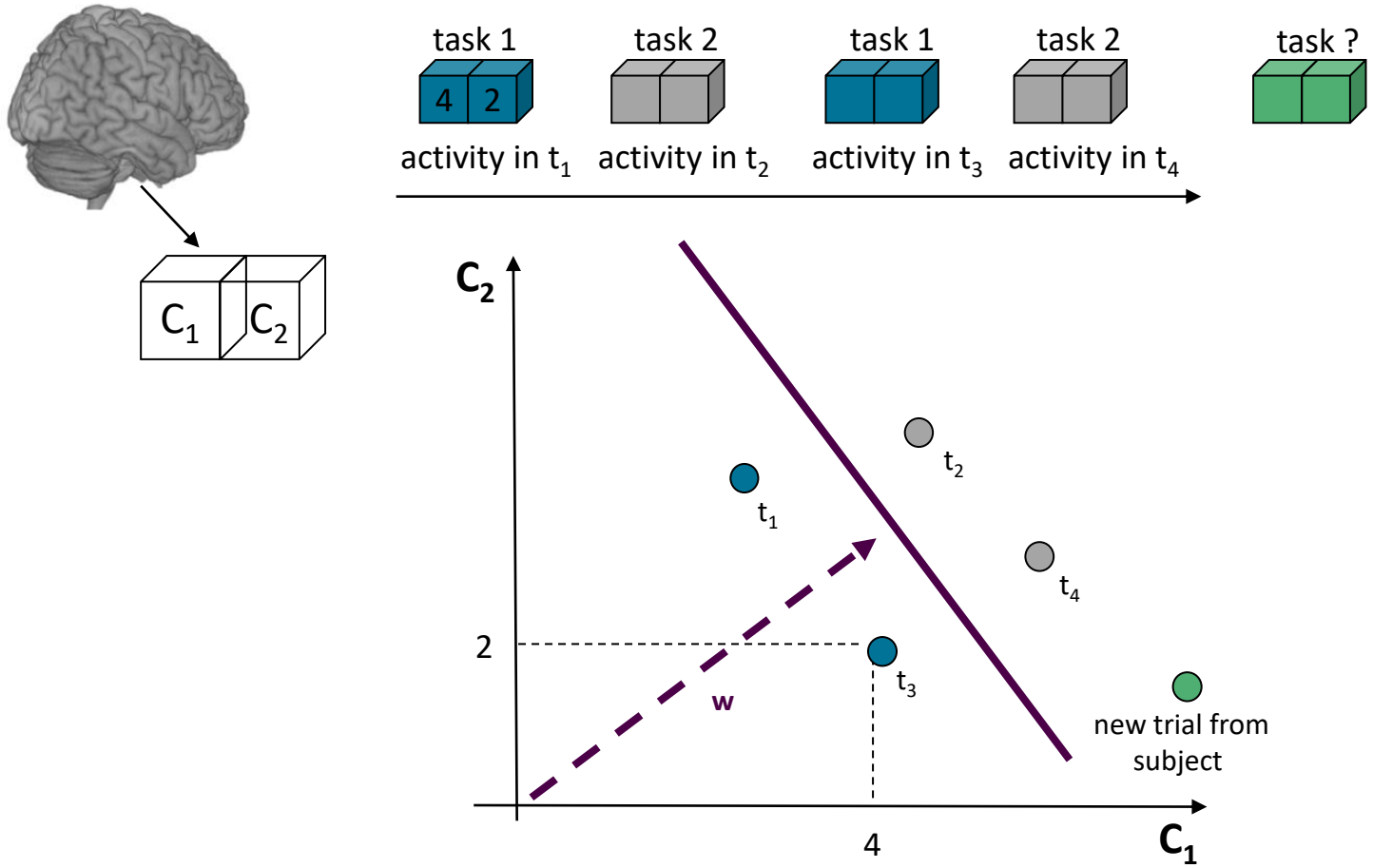
September 15 - 17

Outline

Is my model good?

- Measures of performance for classification
- Measures of performance for regression
- Validation set and cross-validation
- Nested cross-validation
- Assessing significance

Classification: reminder



Classification: confusion matrix

		Predicted	
		P	N
True	P	TP	FN
	N	FP	TN

Accuracy statistics can be shown in a confusion matrix:

Class 1 (P) accuracy, sensitivity = TP/(TP+FN)
Class 2 (N) accuracy, specificity = TN/(FP+TN)
Total Accuracy = (TP+TN)/(TP+FP+FN+TN)
Balanced Accuracy (BA) = mean of classes accuracy
Class 1 predictive value: TP/(TP+FP)
Class 2 predictive value: TN/(FN+TN)

Perfect: FN = FP = 0. Be suspicious if this happens!
Random: TP = TN = FP = FN. Same as flipping a coin.

Classification: accuracy

Total accuracy vs. balanced accuracy

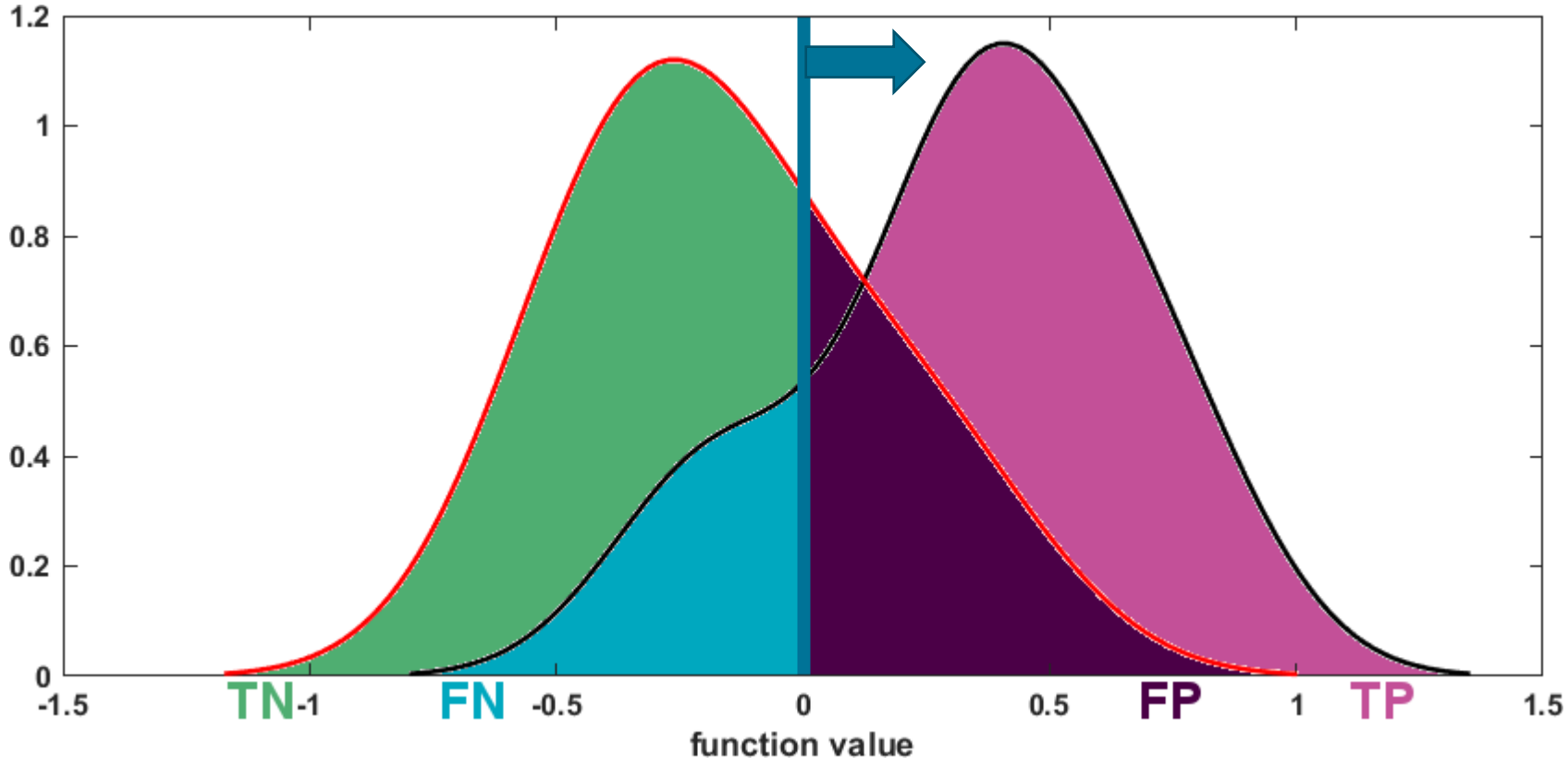
- If classes don't have the same number of examples
- Total accuracy may seem to be above chance whereas the minority classes are sacrificed and below chance
- A common strategy is to subsample the majority class, but data is lost
- Subsample many times (computationally intensive)
- Reporting class accuracies (p_0,..., p_C) is good practice
- Balanced accuracy is the average of class accuracies

Classification: ROC

For a fixed classifier, increasing sensitivity can only come at the cost of decreasing specificity, and vice-versa.

sensitivity = TP/(TP+FN)

specificity = TN/(FP+TN)

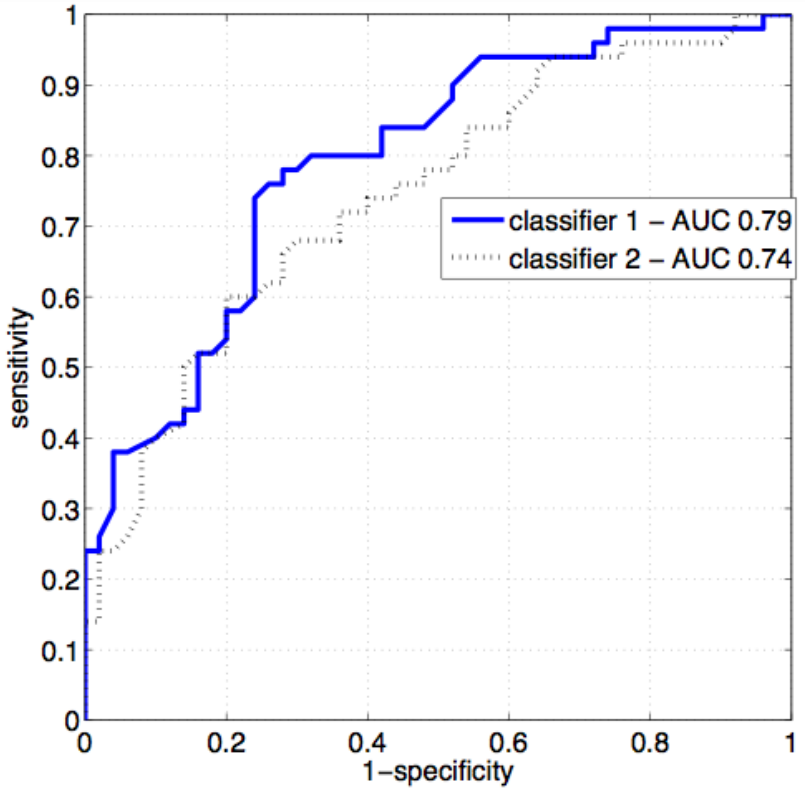


Classification: ROC

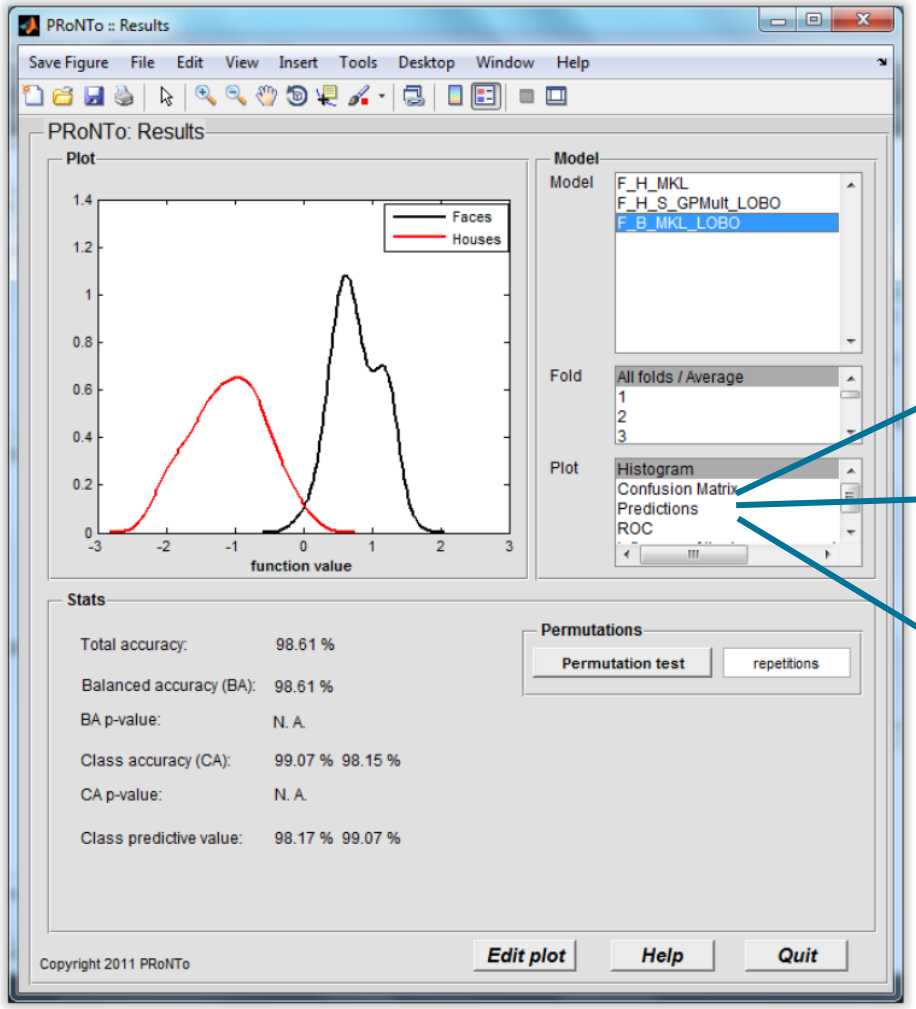
The **Receiver Operating Characteristic (ROC) curve** is a good way of seeing the sensitivity/specificity tradeoff over the operating range of a classifier.

Classifier comparison via **Area Under Curve (AUC)**

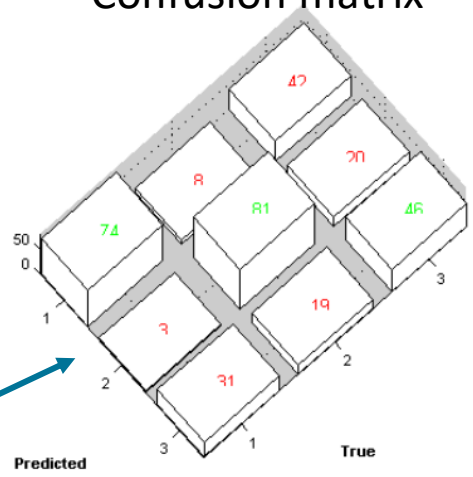
AUC = 1.0: perfect
AUC = 0.5: chance



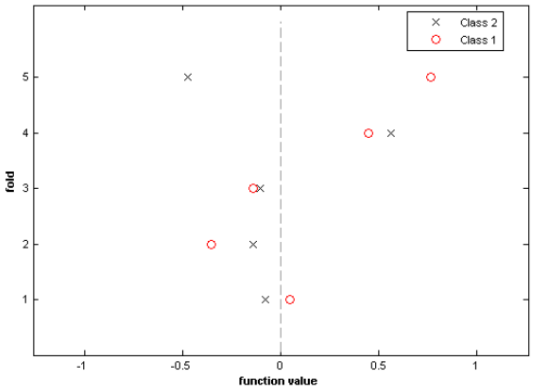
Classification: PRoNTo



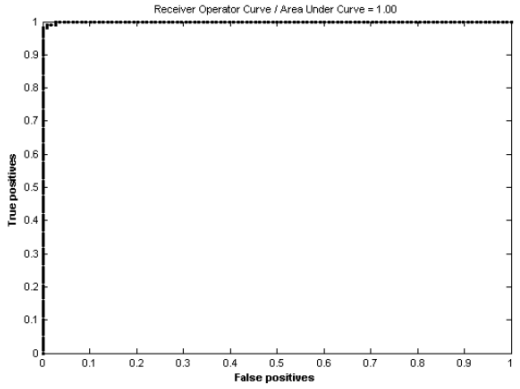
Confusion matrix



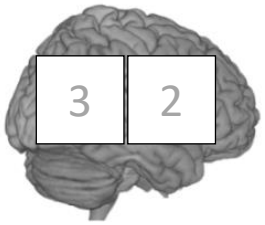
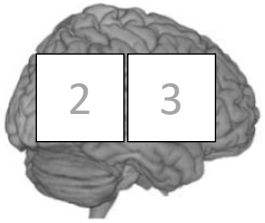
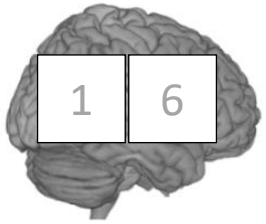
Predictions

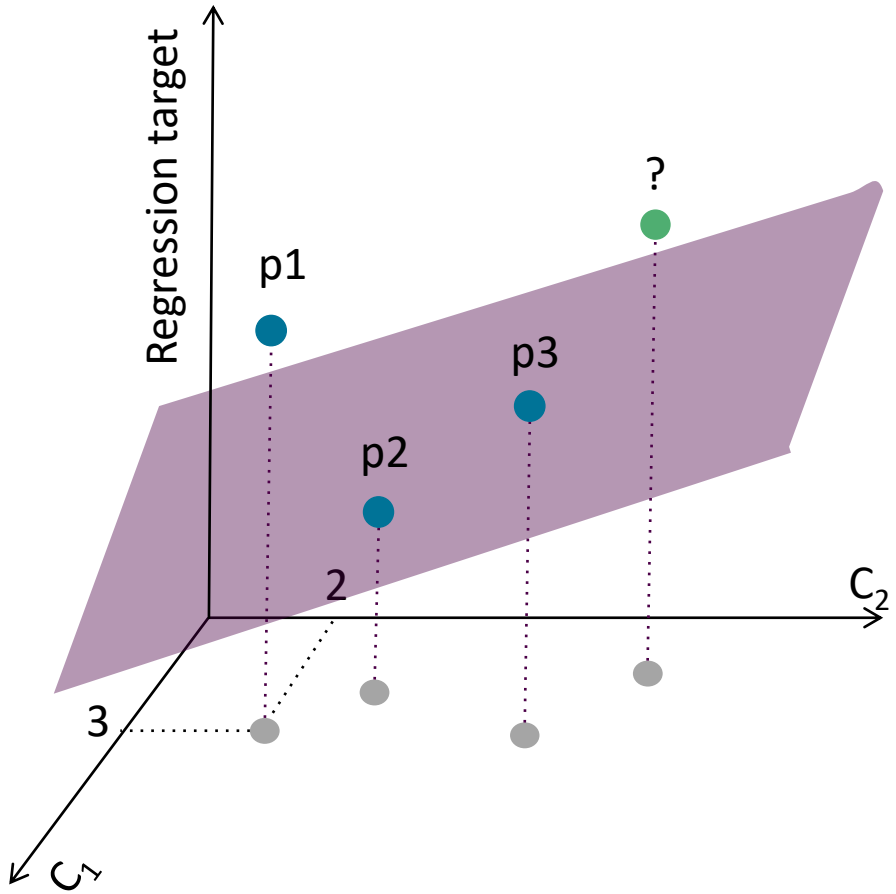


ROC curve



Regression: reminder

Pattern of brain activation	Target
	p1
	p2
	?



Regression: performance

- Correlation:

$$\mathrm{corr}\big(y, f(x)\big) = \frac{\sum_n (y_n - \mu_y)(f(x_n) - \mu_f)}{\sqrt{\sum_n (y_n - \mu_y)^2 \sum_n (f(x_n) - \mu_f)^2}}$$

- Coefficient of determination:

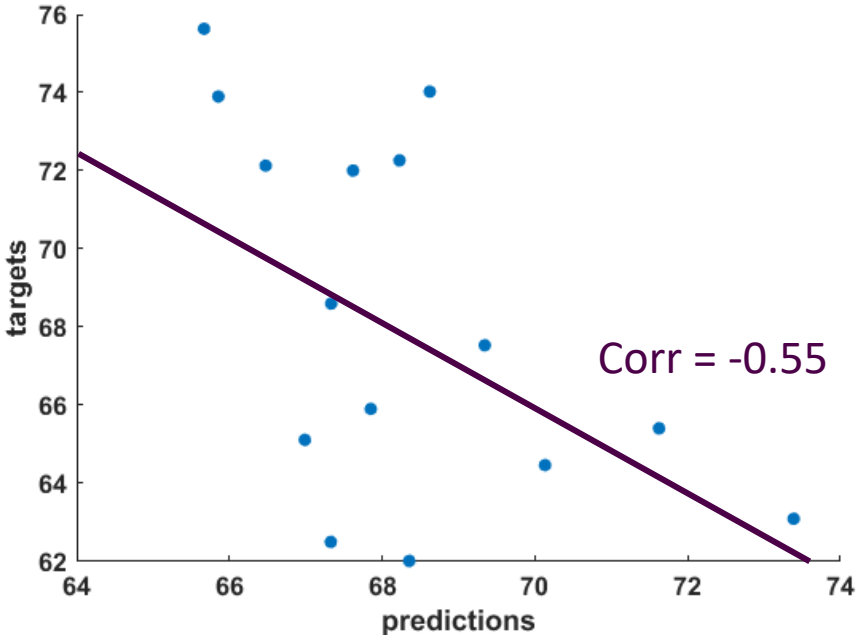
$$R^2 = \mathrm{corr}(y, f(x))^2$$

- Mean Squared Error:

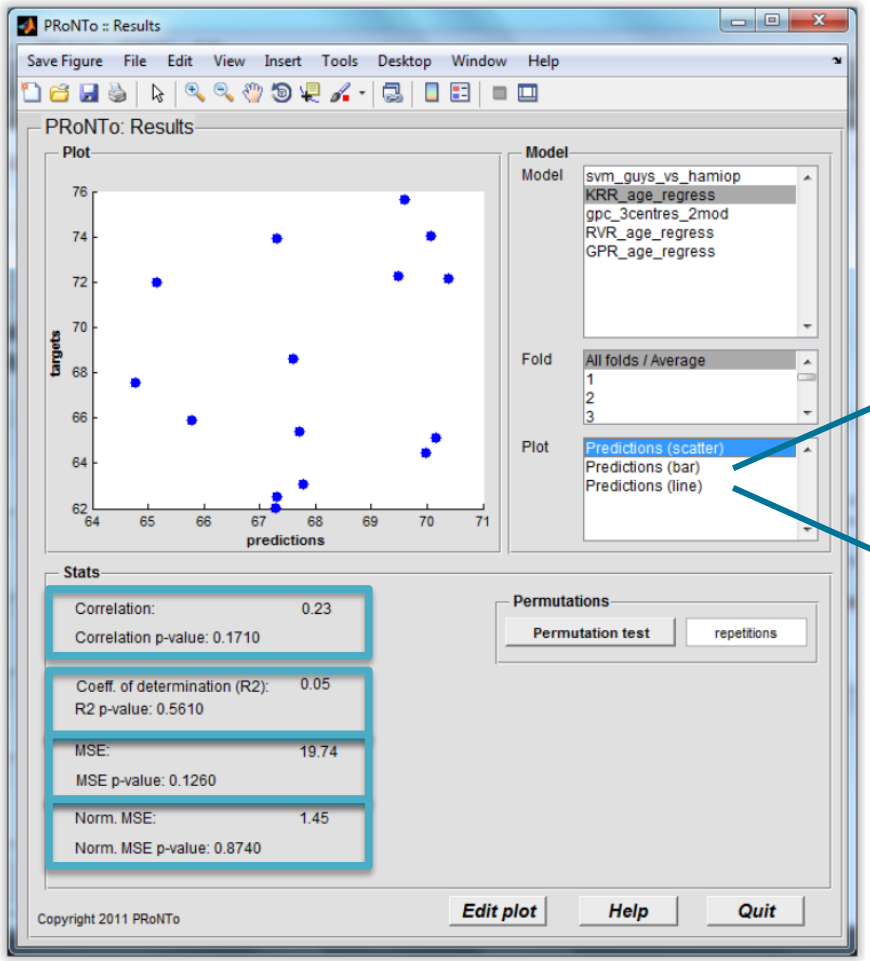
$$\mathrm{MSE} = \frac{1}{N}\sum_n (y_n - f(x_n))^2$$

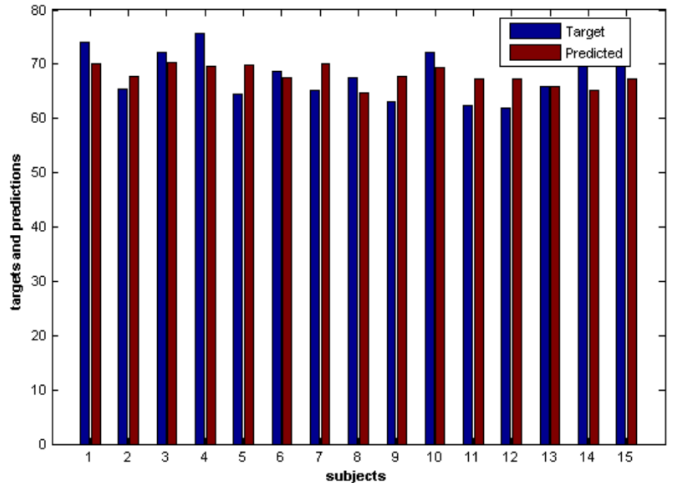
- Normalized MSE:

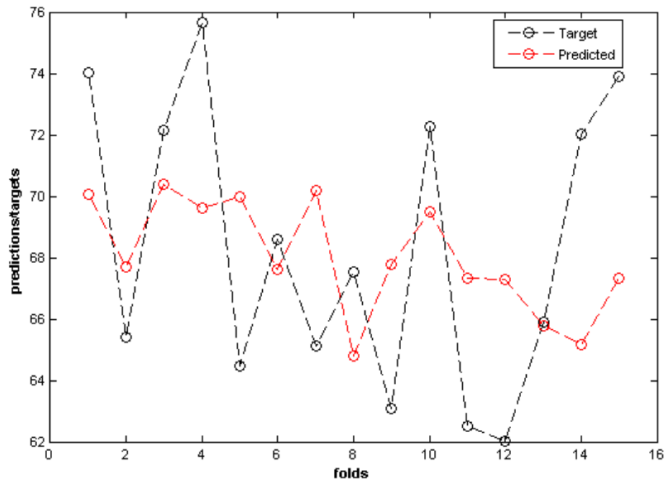
$$\mathrm{NMSE} = \mathrm{MSE}/(y_{max} - y_{min})$$



Regression performance in PRoNTo

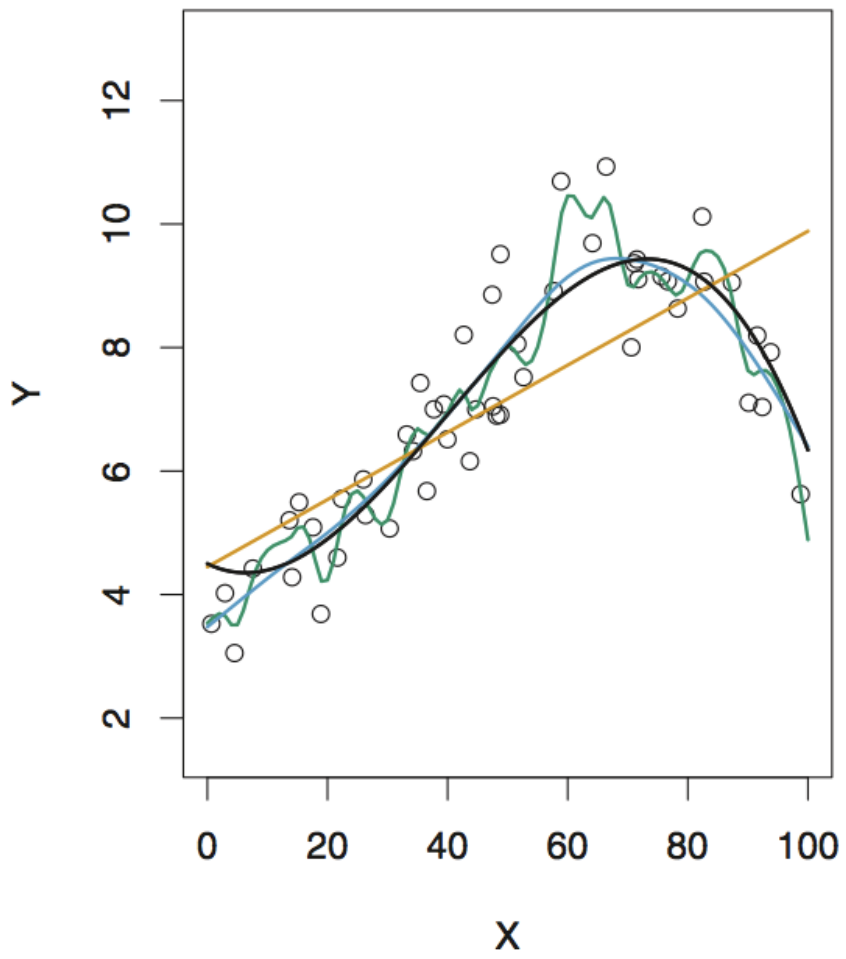




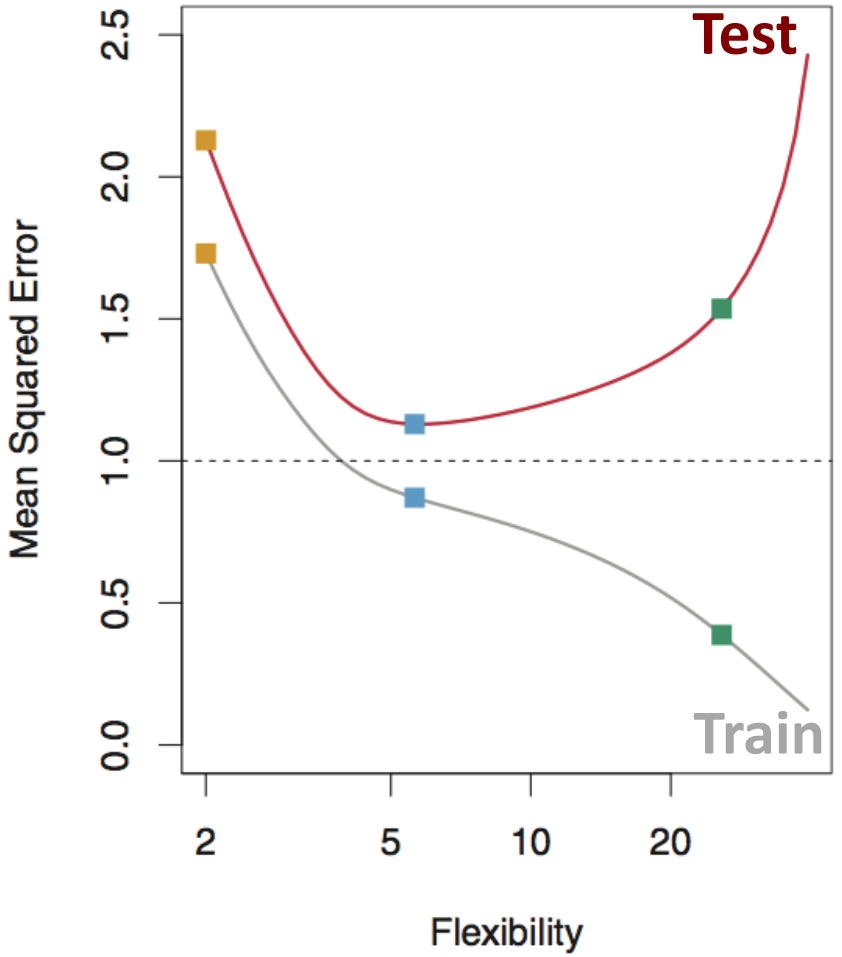


Train and test error

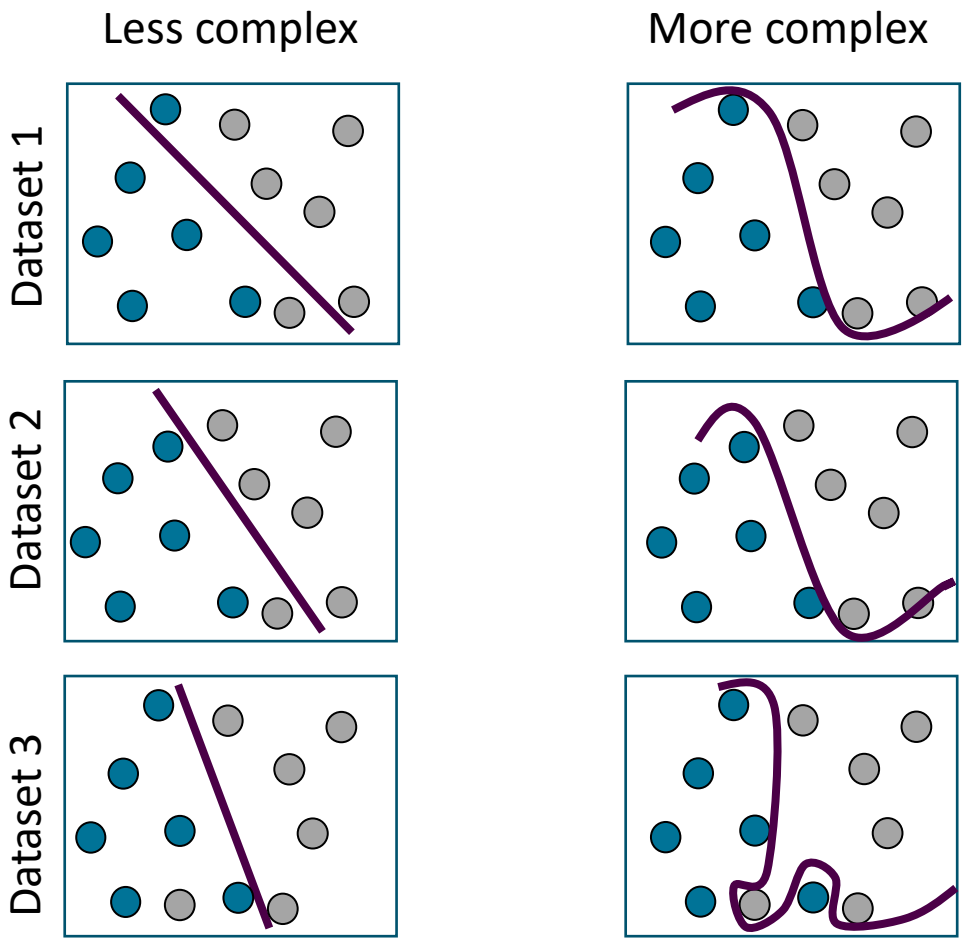
Different models

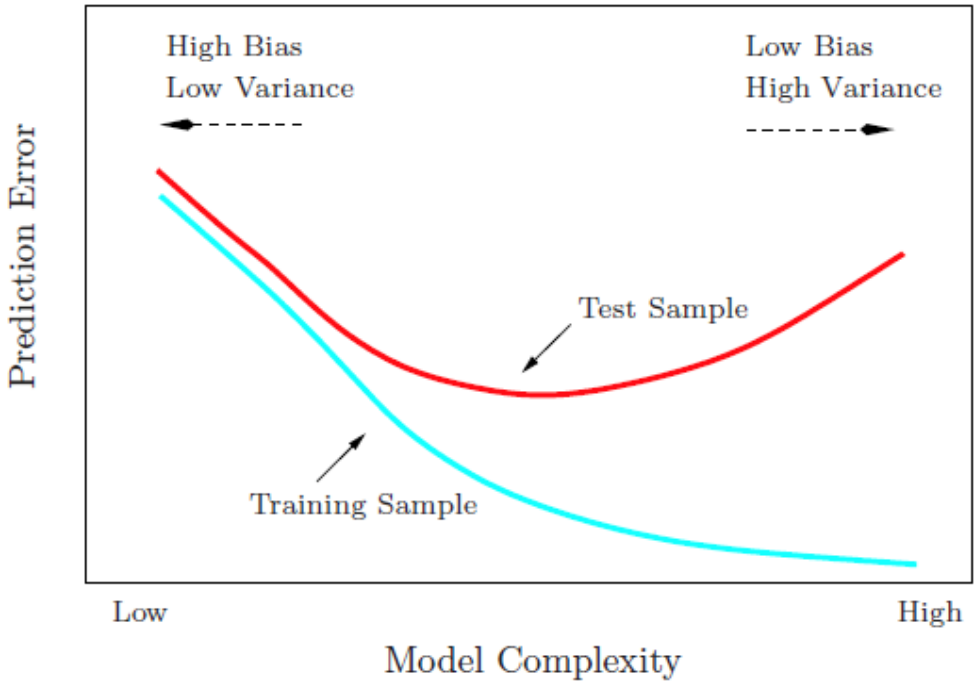


Prediction Error



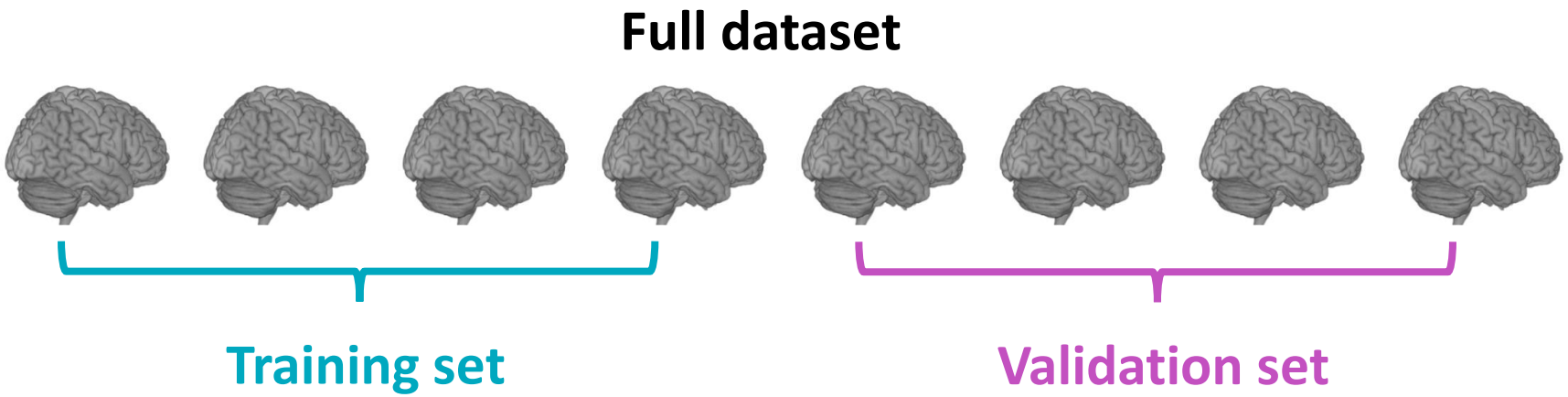
Bias-variance trade-off





Variance: variations in decision functions when the data set is modified (over-fitting)
Bias: error caused by model assumption (under-fitting)

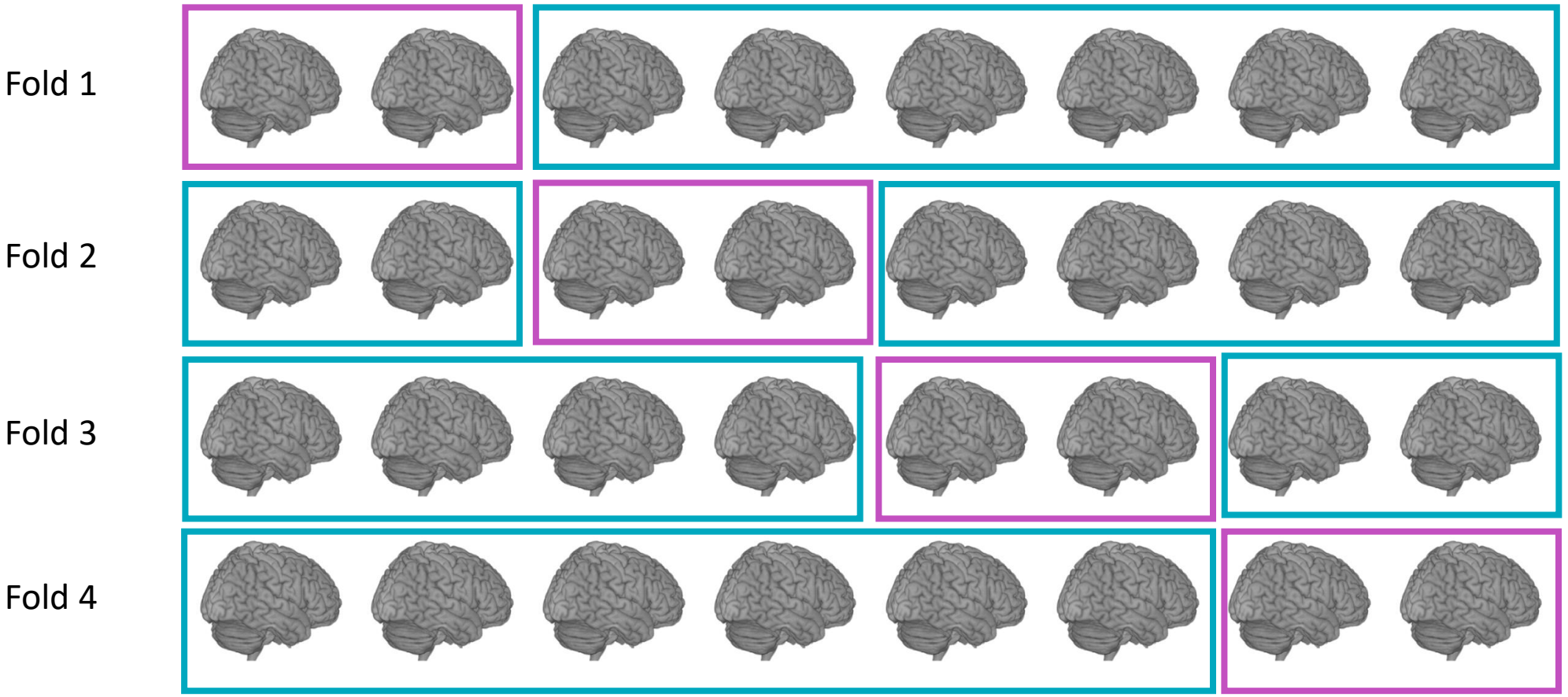
Validation: validation set



Drawbacks:

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

Validation: cross-validation



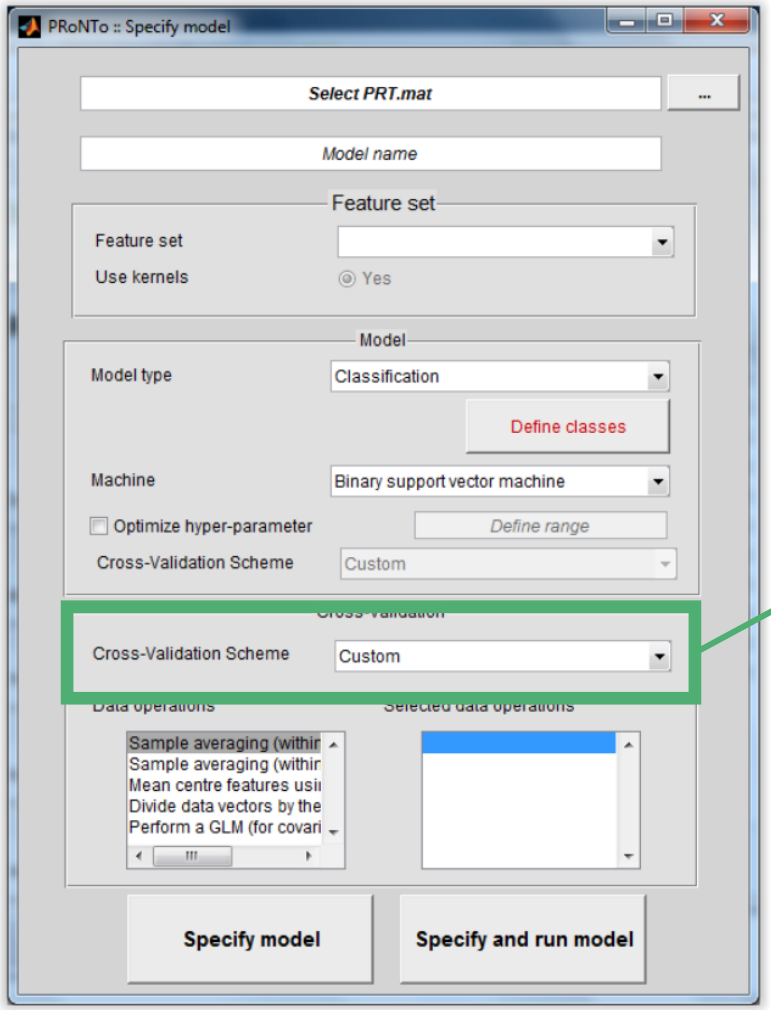
Validation: cross-validation

- Number of folds:
  - = number of samples: Leave-One-Out (but see (Varoquaux, 2017))
  - = user based: typically, leave 10 to 20% of data out

- Data in each fold:
  - Regression: are samples sorted?
  - Classification: Leave-per-Class-Out, keeping frequency distributions in each fold
  - Structured data: correlated blocks in test set

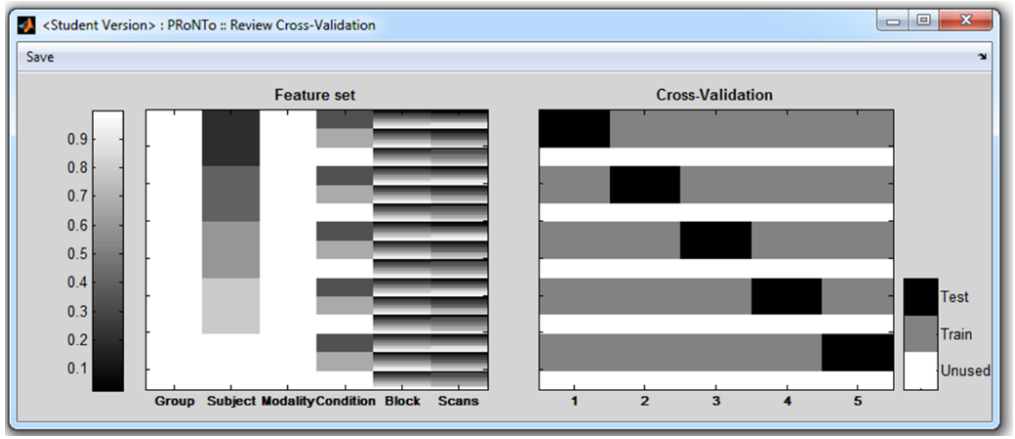
- Results will depend on chosen cross-validation, no cherry picking!
- Good practice to report model performance in average and std

Validation: PRoNTo

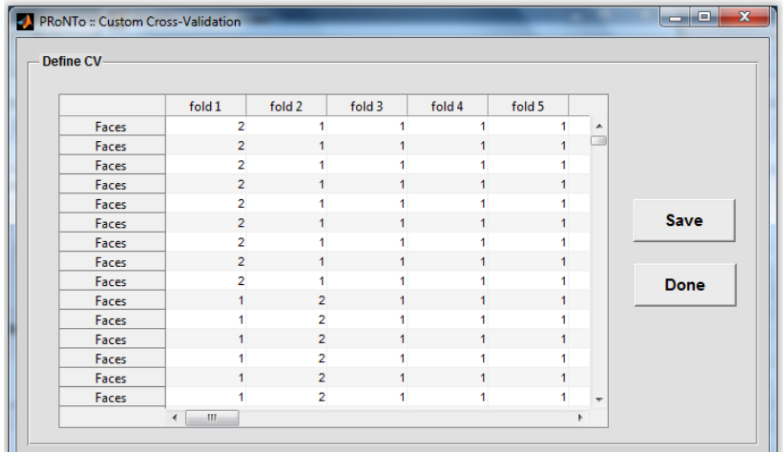


Standard approaches:

- LOSO
- LOBO
- LORO
- LOSGO
- k-fold CV

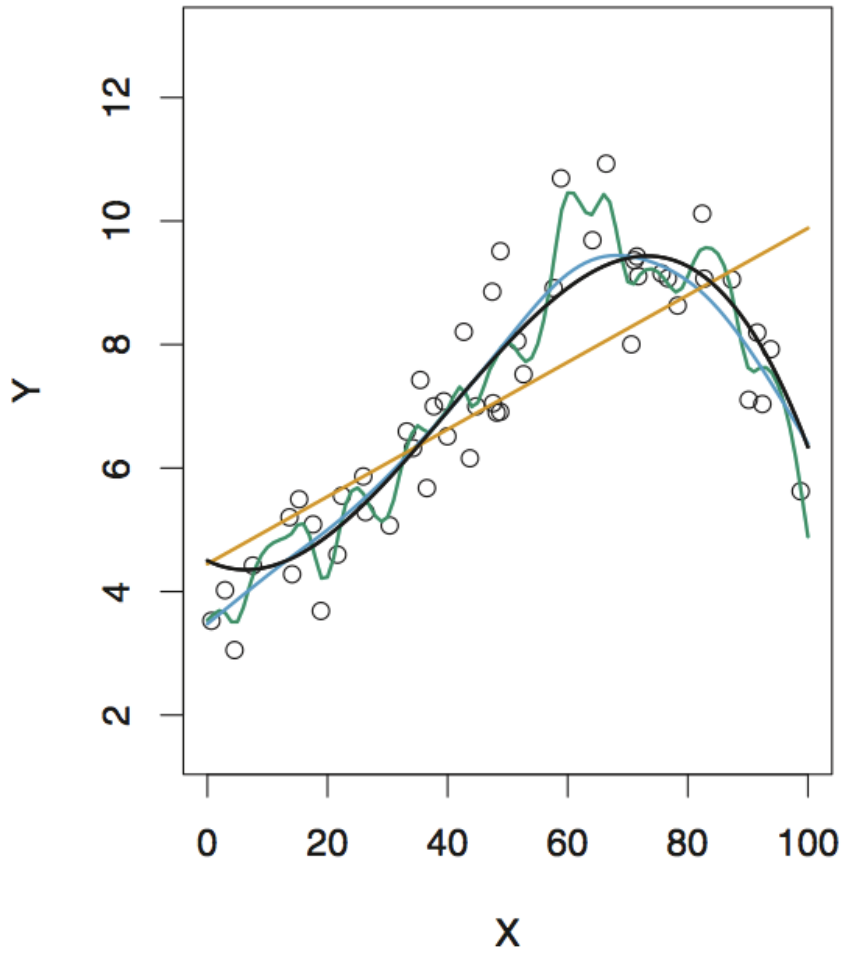


Flexible CV schemes allowed

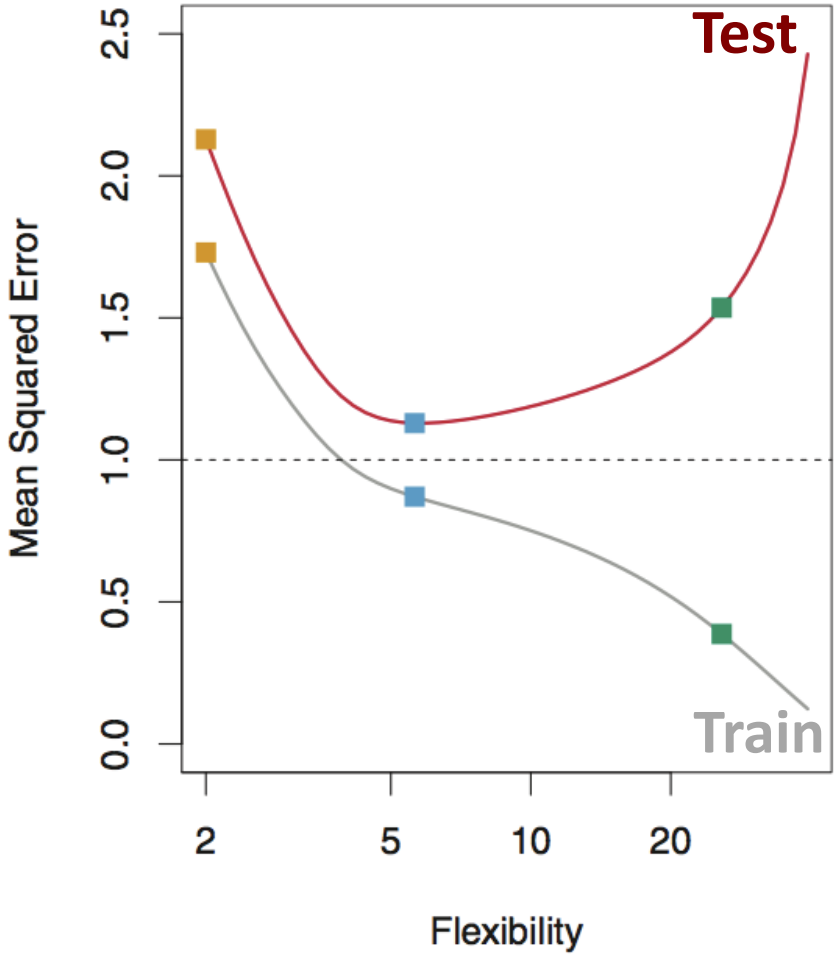


Hyper-parameters

Different models

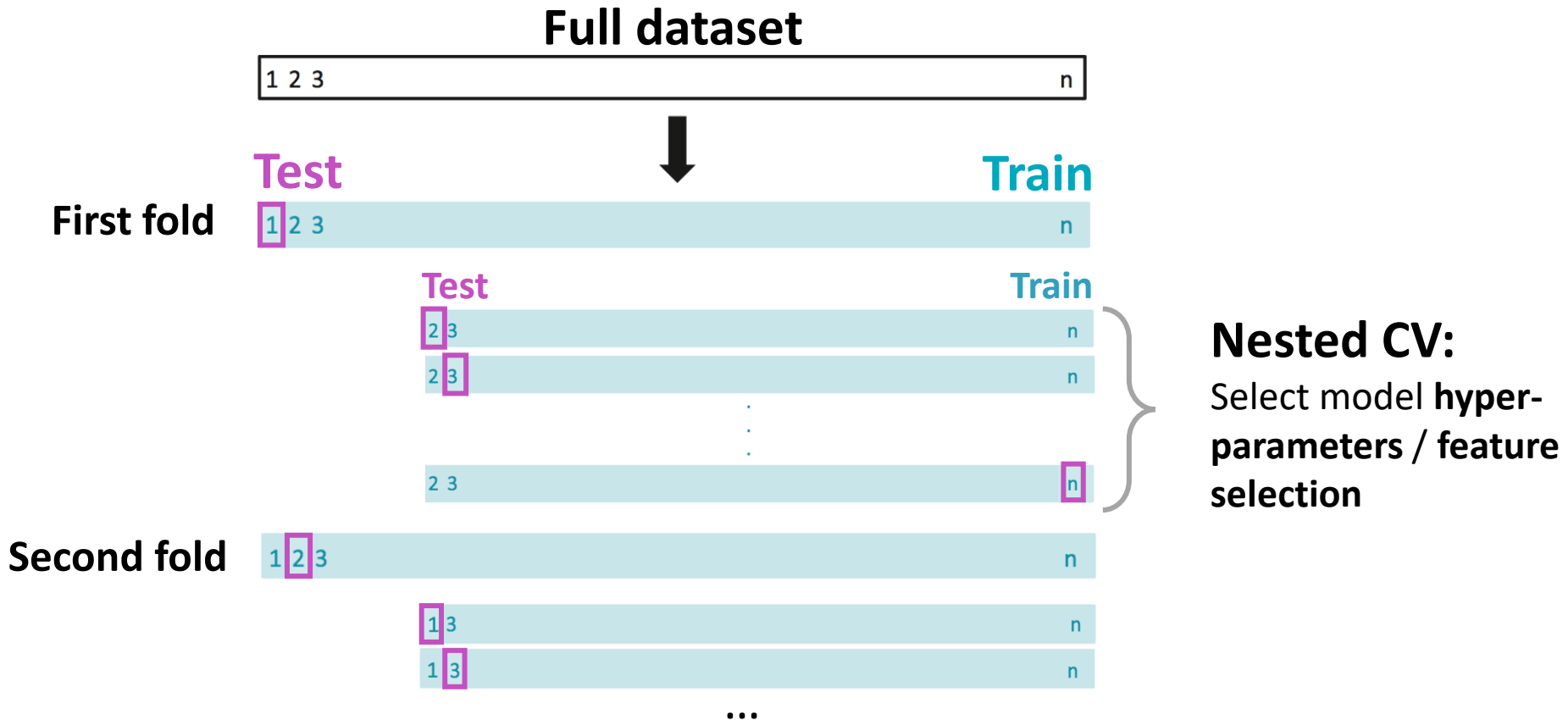


Prediction Error



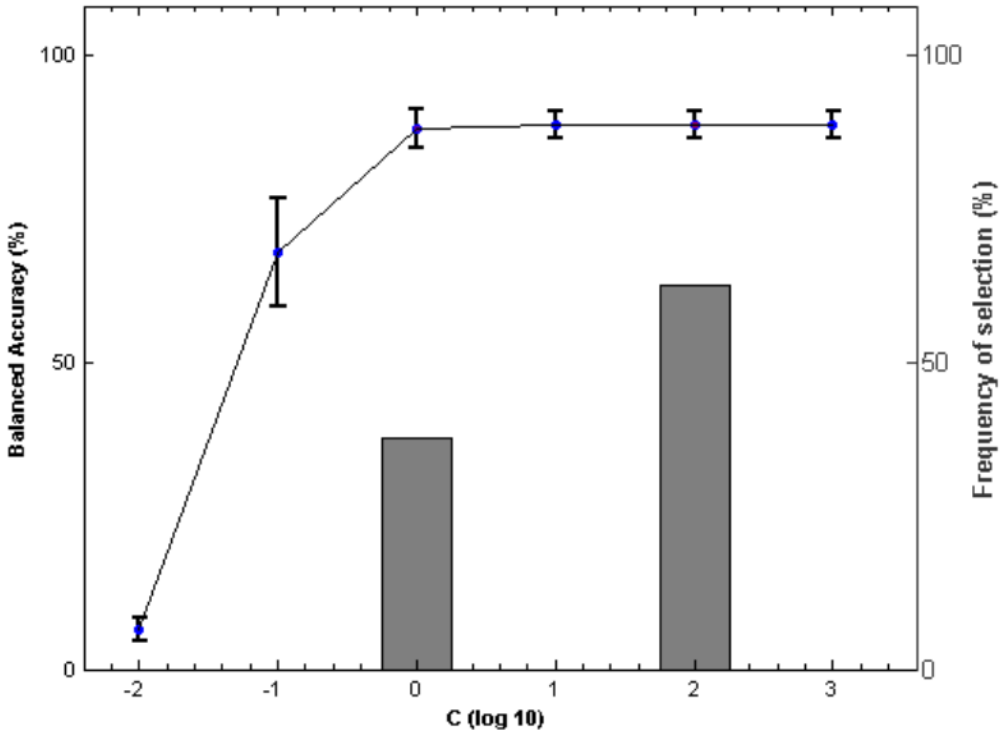
Nested cross-validation

- Problem: use CV to select best model and assess model performance (test error)
- Solution: **Run CV inside CV for model or feature selection**



Model selection in PRoNTo

If hyper-parameter optimisation was performed using nested CV:



Assessing significance

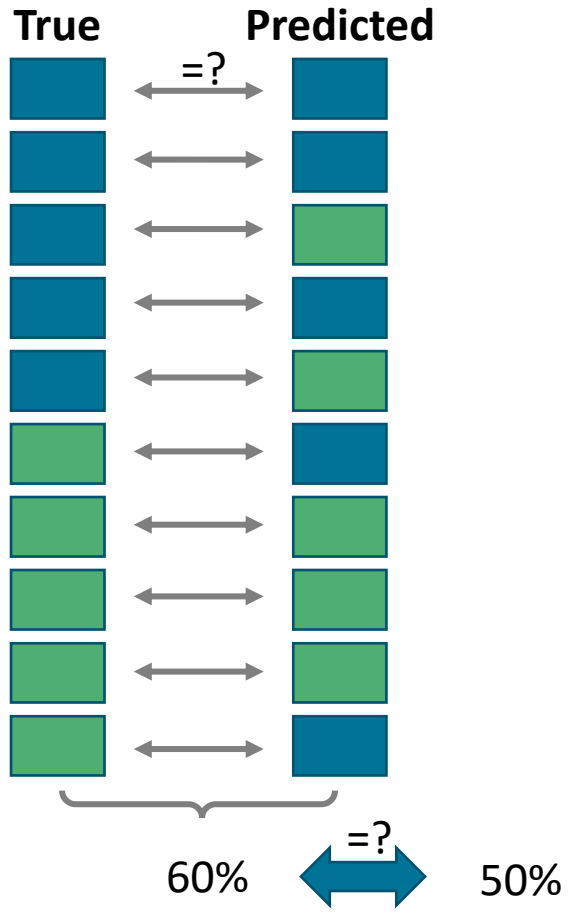
Parametric tests

e.g. Binomial test

- Model decision in two-class problem modeled as Bernoulli trials
- Probability of *k* successes out of *n* trials follows binomial distribution

Not a good idea:

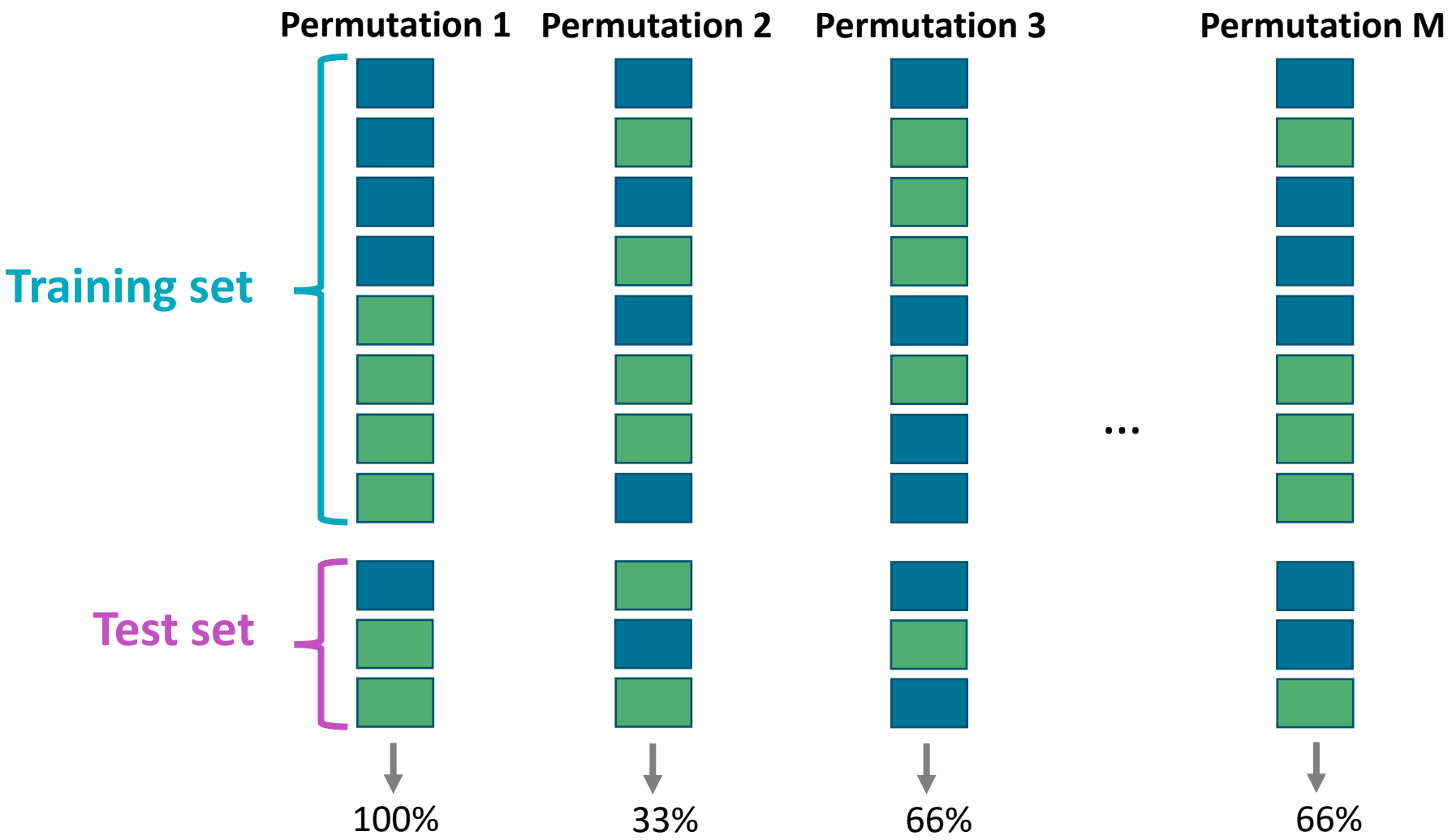
- Assumes IID samples
- Accuracy from cross-validated data does not follow the binomial distribution (Noirhomme et al. 2014)



Assessing significance

- No hypotheses on data distribution
- H_0: “targets are non-informative”
- Test statistic: balanced and class accuracy / MSE / R^2
- Estimate the distribution of the test statistic under H_0 by randomly permuting targets M-1 times, and running the full CV experiment

Assessing significance



p-value:

$$p = \frac{1}{M} \sum_{m=1}^{M} (acc_m^{perm} \geq acc_1)$$

Minimum: $p = \frac{1}{M}$

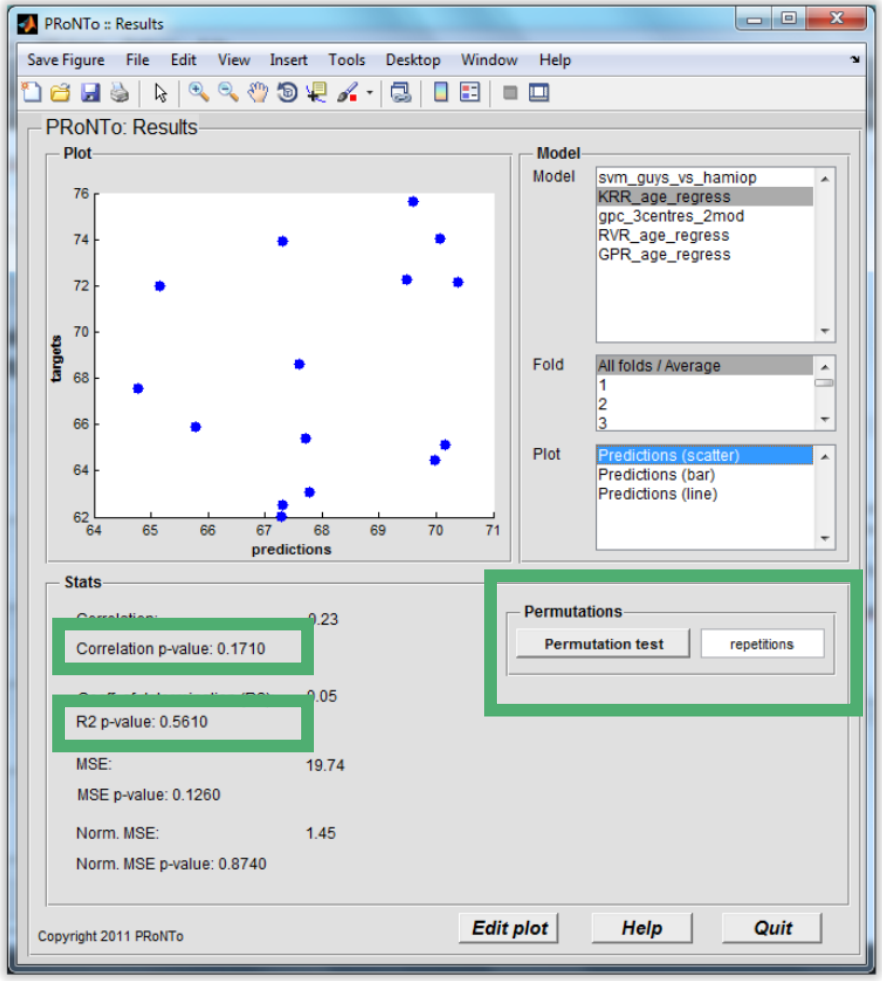
Assessing significance

- No hypotheses on data distribution
- H_0: “targets are non-informative”
- Test statistic: balanced and class accuracy / MSE / R^2
- Estimate the distribution of the test statistic under H_0 by randomly permuting targets M-1 times, and running the full CV experiment

Way to go but…

- Computationally intensive
- Requires “sufficiently large” M
- Returns *estimate* of p-value

Assessing significance



In PRoNTo:

User-input = M-1

Take-home on performance

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

Recommended reading: performance

- Duda et al., *Pattern Recognition*, Wiley, 2001.
- Hastie et al., *The elements of statistical learning*, Springer, 2009.
- James et al., *Introduction to Statistical Learning*, Springer, 2014.
- Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, IJCAI, 1995.
- Kriegeskorte et al., *Circular analysis in systems neuroscience: the dangers of double dipping*, Nature Neuroscience 12, 2009.
- Noirhomme et al., Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions, Neuroimage Clin., 2014
- Pereira et al., *Machine learning classifiers and fMRI: A tutorial overview*, NeuroImage 45, 2009.
- Varoquaux, *Cross-validation failure: Small sample sizes lead to large error bars*, NeuroImage, 2017.

Thank you!

Questions?

LIÈGE université
GIGA
CRC In vivo Imaging

FREEDOM TO RESEARCH