



**Cross Validation: Right or Wrong?**

• We want to apply a classifier to a dataset with two classes

Proposed strategy:

Start with the 5000 available predictions and 50 samples

Select 100 predictors based on highest correlation with the class labels

Apply a classifier, e.g. logistic regression, using only the 100 predictors

*How do we estimate the test performance?*



- We want to apply a classifier to a data set with two classes
- Proposed strategy:
  - Start with the 5000 available predictors and 50 samples
  - Select 100 predictors based on highest correlations with the class labels
  - Apply a classifier, e.g. logistic regression, using only these 100 predictors

- The wrong way: apply cross validation for the classifier only

- The right way: apply cross validation in both steps

# Cross Validation: Right or Wrong?

- We want to apply a classifier to a data set with two classes
- Proposed strategy:
  - Start with the 5000 available predictors and 50 samples
  - Select 100 predictors based on highest correlations with the class labels
  - Apply a classifier, e.g. logistic regression, using only these 100 predictors

*How do we estimate the test performance?*

- The wrong way: apply cross validation for the classifier only
- The right way: apply cross validation in both steps

# Cross Validation: Right or Wrong?

