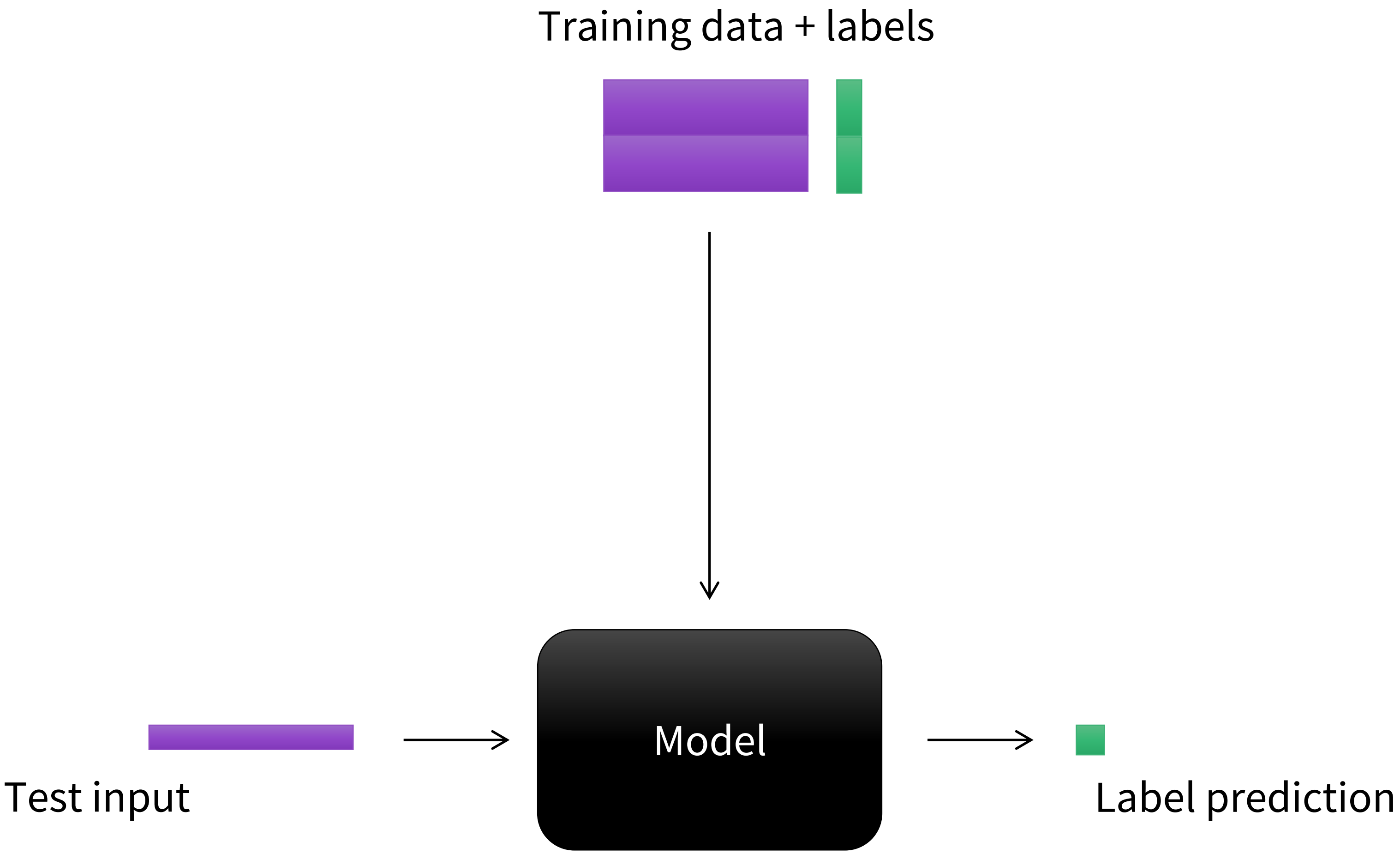
PRACTICAL ML- PART 2

COMMON MISTAKES
TO AVOID

Data Leakage



Training data + labels

Test input

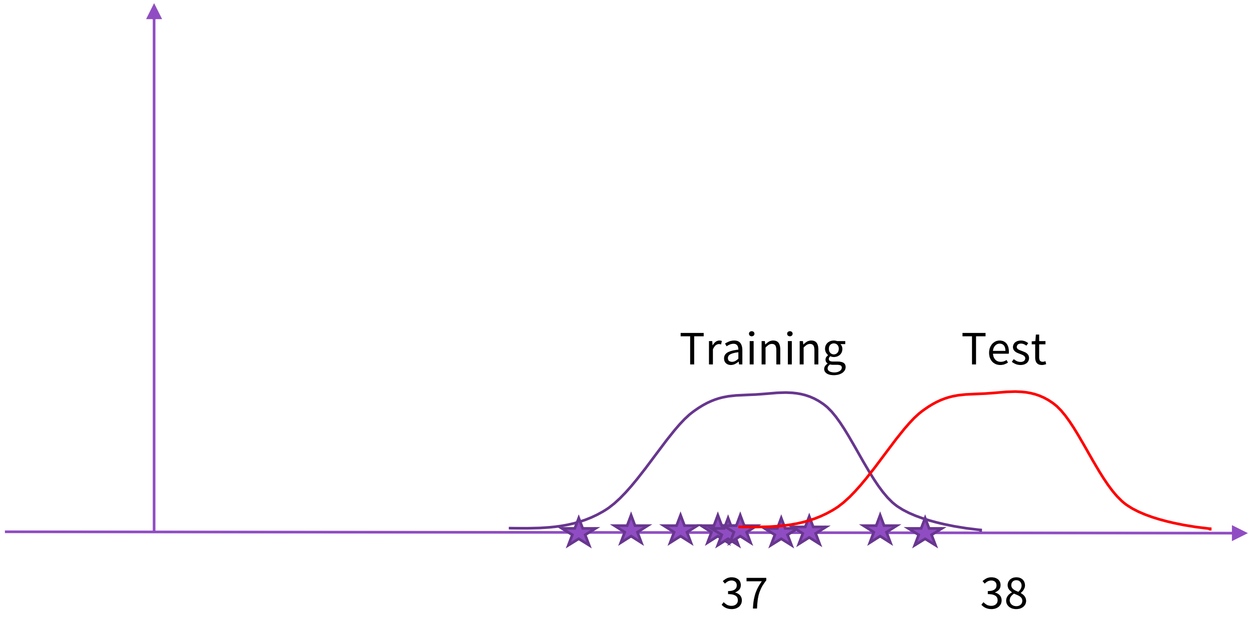
Model

Label prediction

○ Data Leakage

Data leakage occurs when we use accidentally use information from the test set when we are training a machine learning model. It leads to artificially better performance in the test data, because information has ‘leaked’ from the test set to the training process.

Data Leakage example (normalisation)

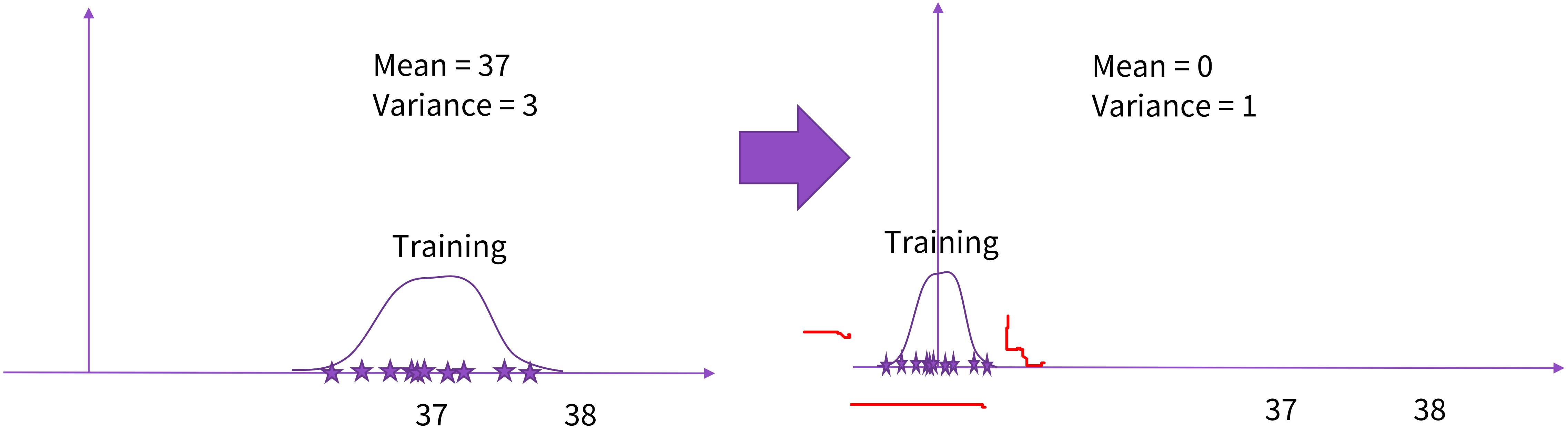


Data from the test set comes from a poorly calibrated sensor

Any model should perform poorly on the test data

A review of novelty detection: http://www.robots.ox.ac.uk/~davidc/pubs/NDreview2014.pdf

Data Leakage example (normalisation)



Mean = 37
Variance = 3

Training

37

38

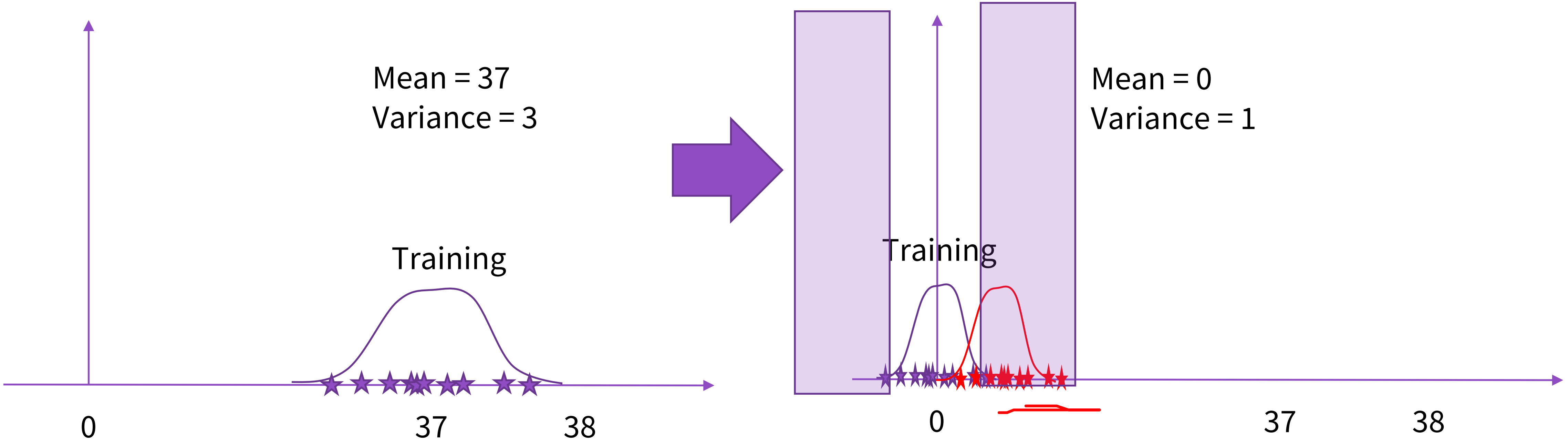
Mean = 0
Variance = 1

Training

37

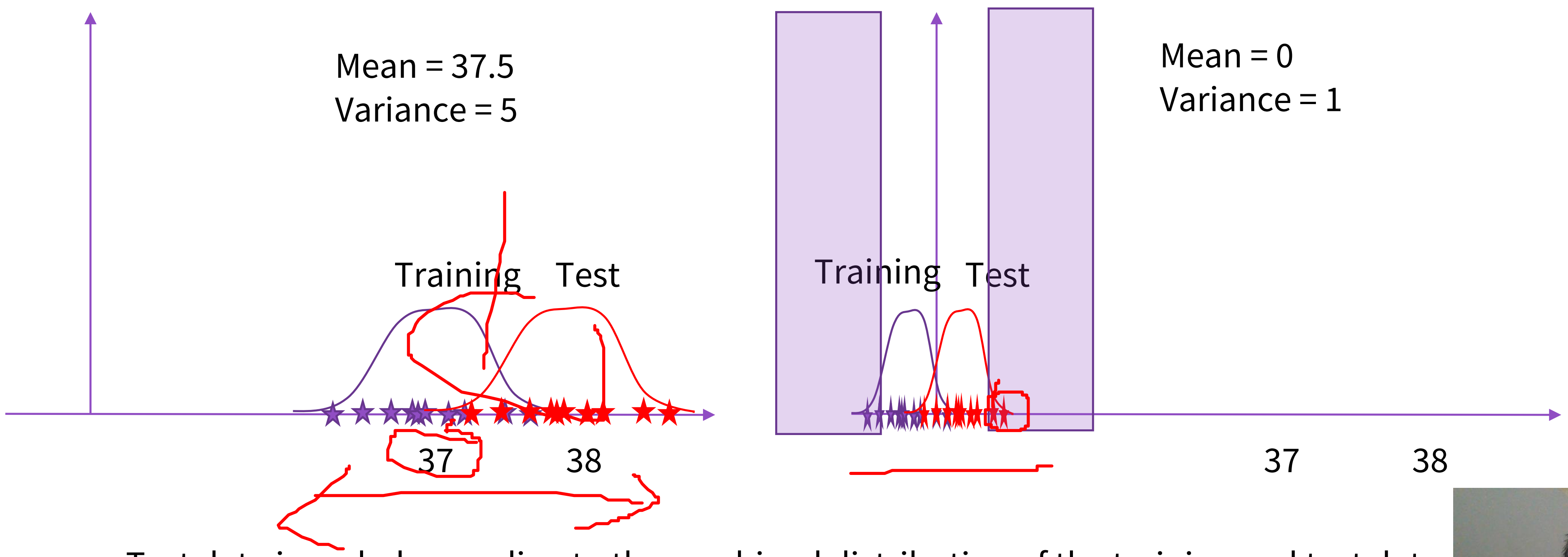
38

Data Leakage example (normalisation)



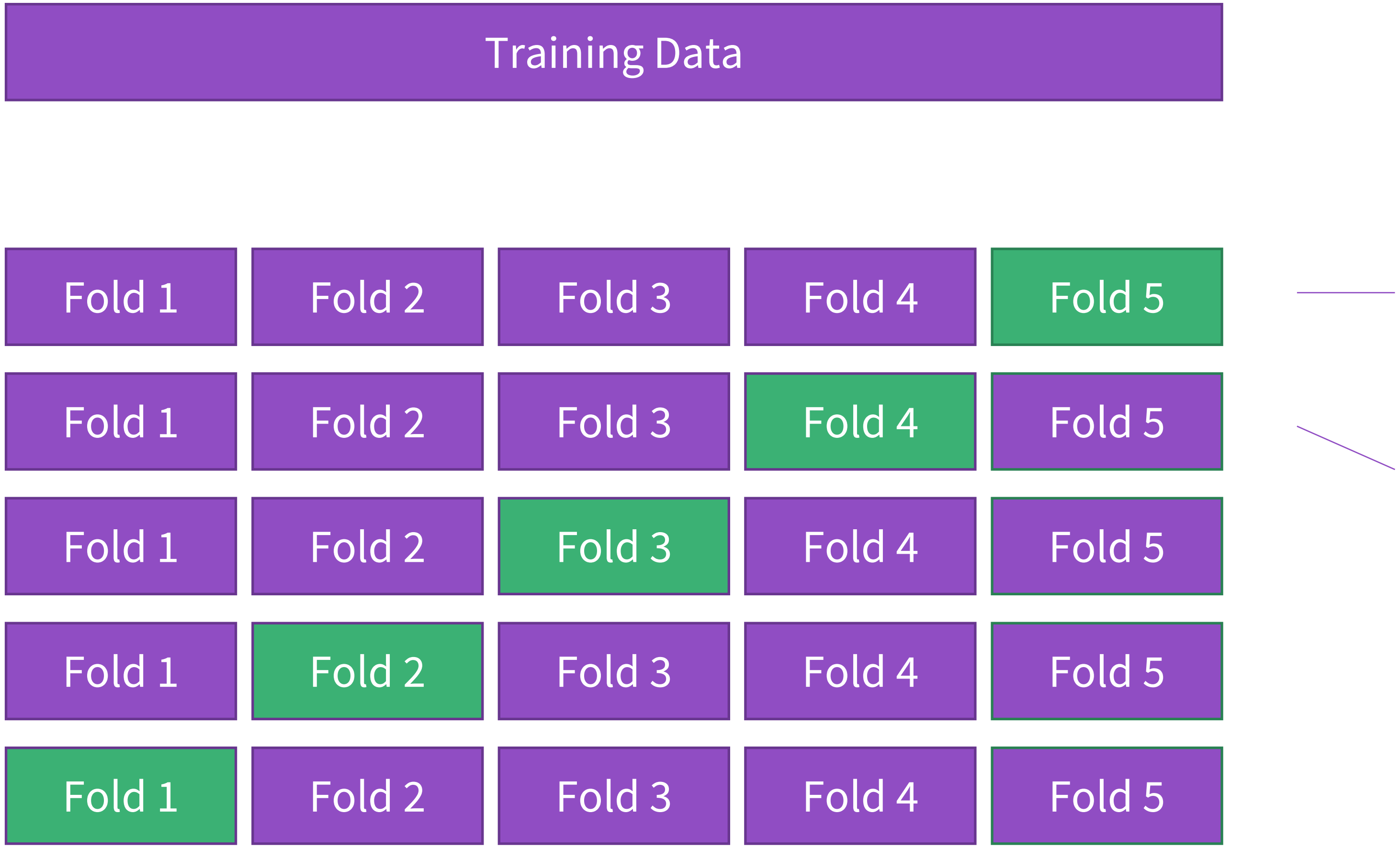
Test data is scaled according to the distribution of the training data, only. Test data is far from zero, so it will be classified as abnormal

Data Leakage example (normalisation)



Test data is scaled according to the combined distribution of the training and test data. Much less of the test data is far from zero, and so the classifier decides a greater proportion of measurements are normal

Data Leakage in Cross Validation



Training Data

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

- Data normalisation using folds 1-4
- Feature selection using folds 1-4
- Validate on fold 5

- Data normalisation using folds 1,2,3,5
- Feature selection using folds 1,2,3,5
- Validate on fold 4

etc

Report average validation result

Imbalanced Learning

- Imbalance is common

- Extreme imbalance common for many real-life problems:
  - Credit card fraud
  - Health monitoring

- Accuracy paradox
  - Possible to get very high accuracy with a poor classifier
  - E.g. consider data set with:
    - 1,000,000 normal credit card transactions
    - 10 fraud

Resampling Methods

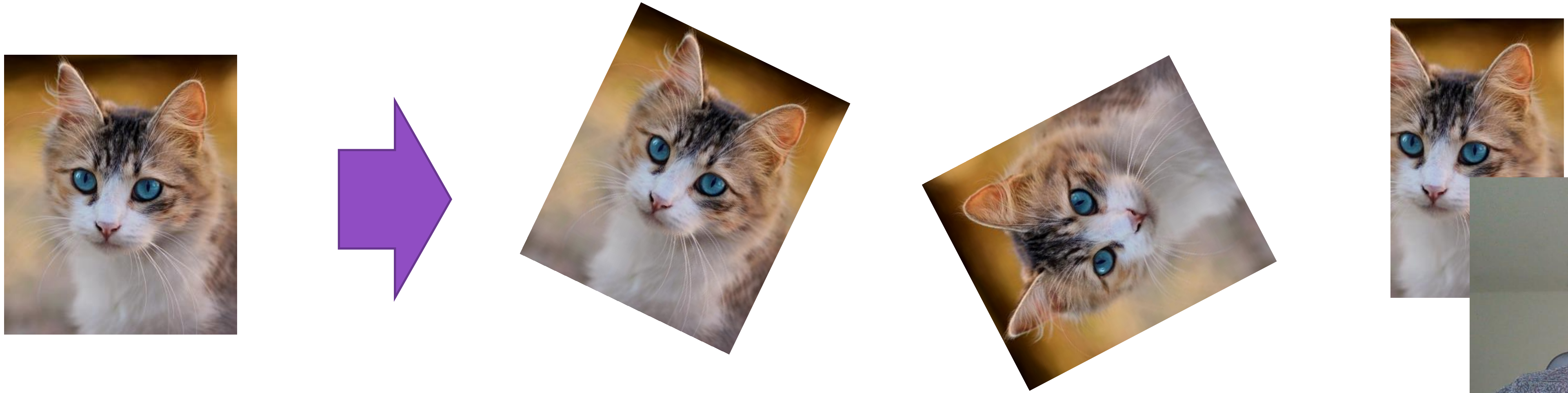
- Undersampling
  - Randomly remove samples of the majority class
  - E.g. if 60,000 in class 1, and 10000 in class 2, then remove 50,000 from the class 1 examples at random

- Oversampling
  - Create extra instances of the minority class
    - E.g. if 60,000 in class 1, and 10000 in class 2, sample randomly 50,000 times from examples in class 2, with replacement, and add to the training set.
    - Data augmentation – e.g. SMOTE

Oversampling via data augmentation

- Synthetic Minority Over-Sampling Technique
  - Creates samples of the minority class that are ‘similar’ to the training data

- For some problems, we can safely create extra instance of the minority class



Summary

- Data Leakage
  - Occurs when information about the test data is inadvertently used during training
  - Avoid by performing all data process on the training set only.

- Class Imbalance
  - Can lead to situations in which accuracy is good, but objective classification is poort
  - Commonly remedied by resampling data