# Classification / Prediction

# Classification

"Supervised Learning"

Use a "training set" of examples to create a model that is able to predict, given an unknown sample, which of two or more classes that sample belongs to.

?

TUMOUR CLASSIFIC

Recognizing differences

# What we'll cover

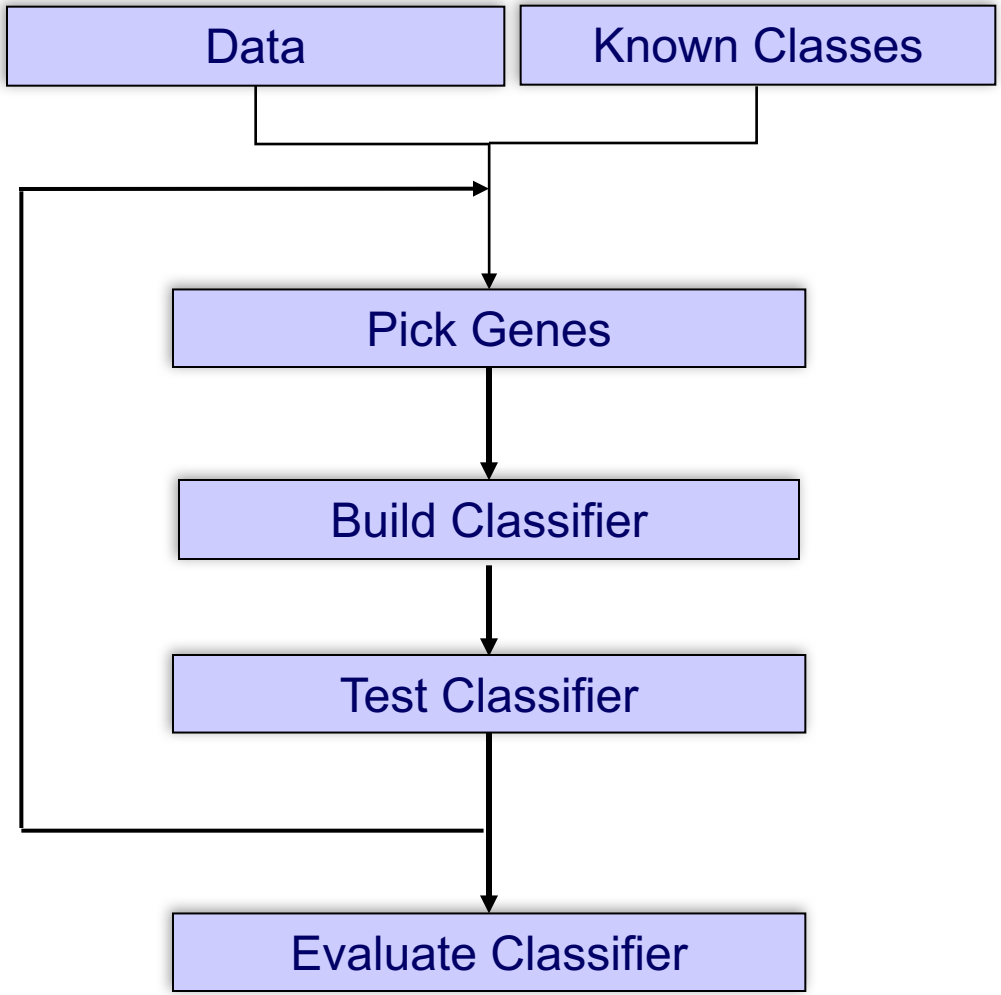
- How to build a classifier.
- How to evaluate a classifier.
- Using GenePattern to classify expression data.

# What Is a Classifier

- A **predictive** rule that uses a set of **inputs** (genes) to predict the values of the **output** (phenotype).
- Known examples (train data) are used to build the predictive rule.
- Goal:
  - Achieve high predictive power.
  - Avoid over-fitting: i.e. classifier memorizes the training data and is not generalizable to other test data

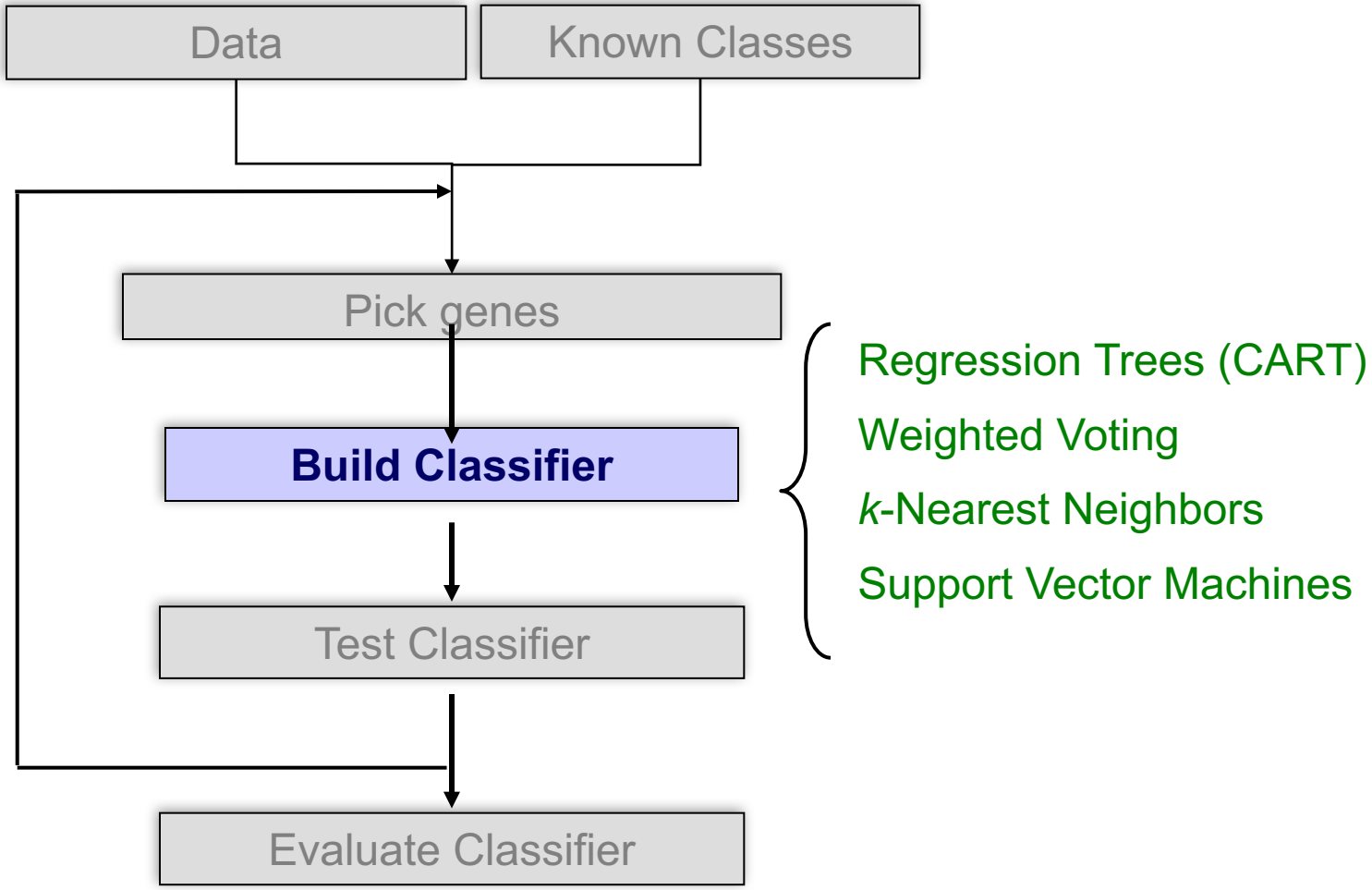
# Classification

## Computational methodology



# Classification

## Computational methodology



# Classifiers

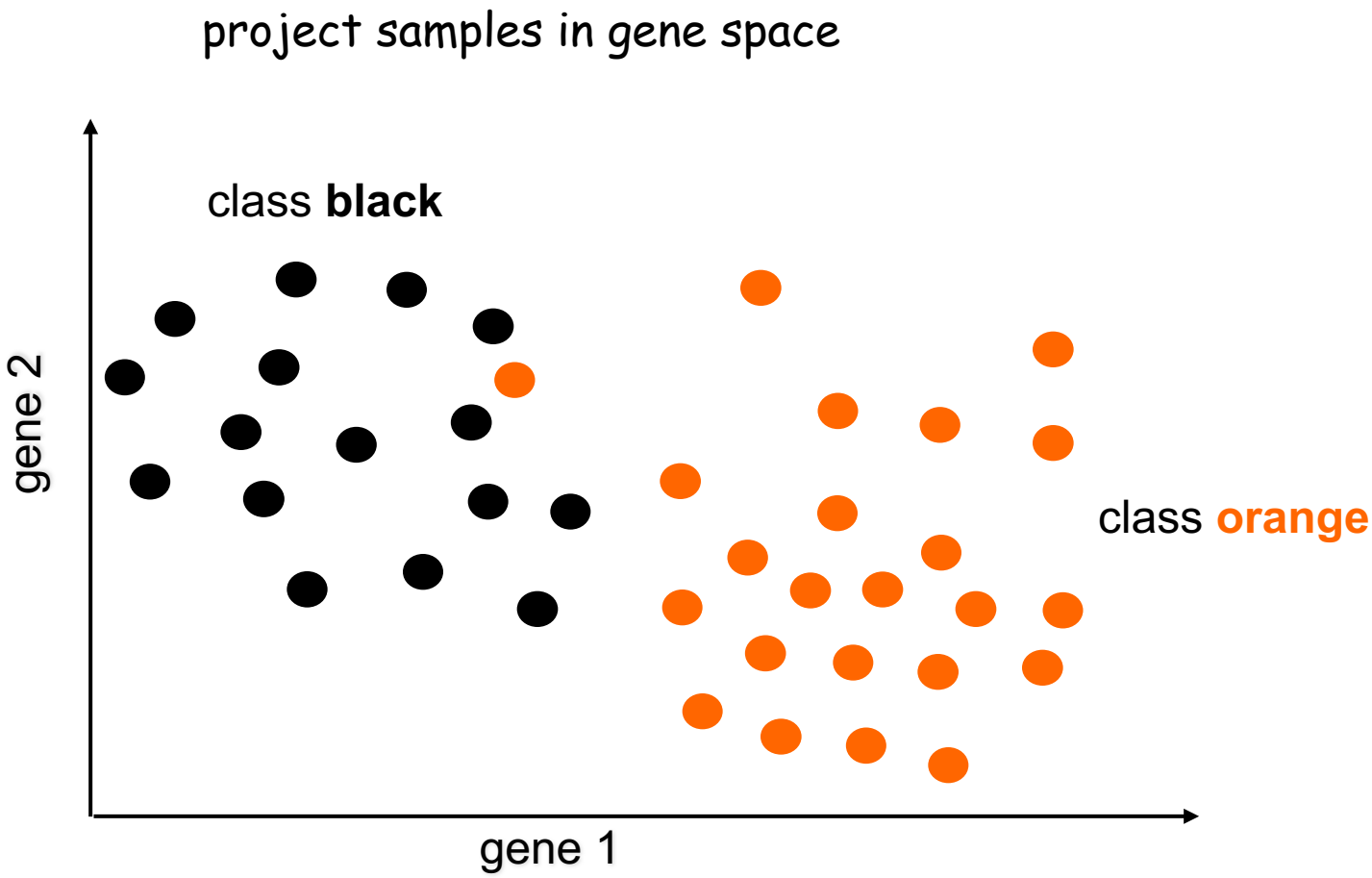
Important issues:

- Few cases, many variables (genes)
- redundancy: many highly correlated genes.
- noise: measurements are very imprecise.
- feature selection: reducing the # of genes is a necessity.

Avoid over-fitting

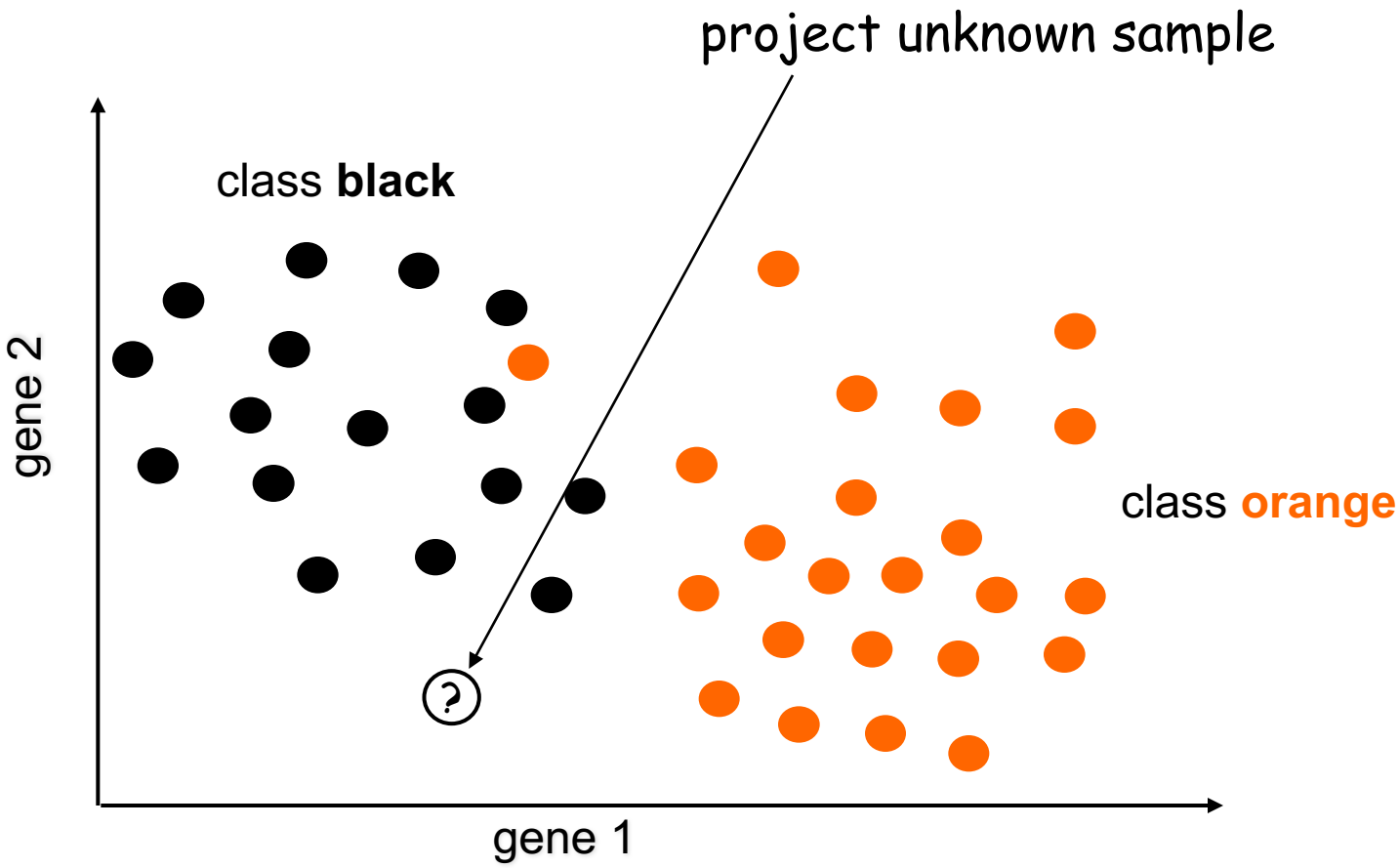
# k Nearest Neighbors (kNN) Classifier

Example: k=5, 2 genes, 2 classes



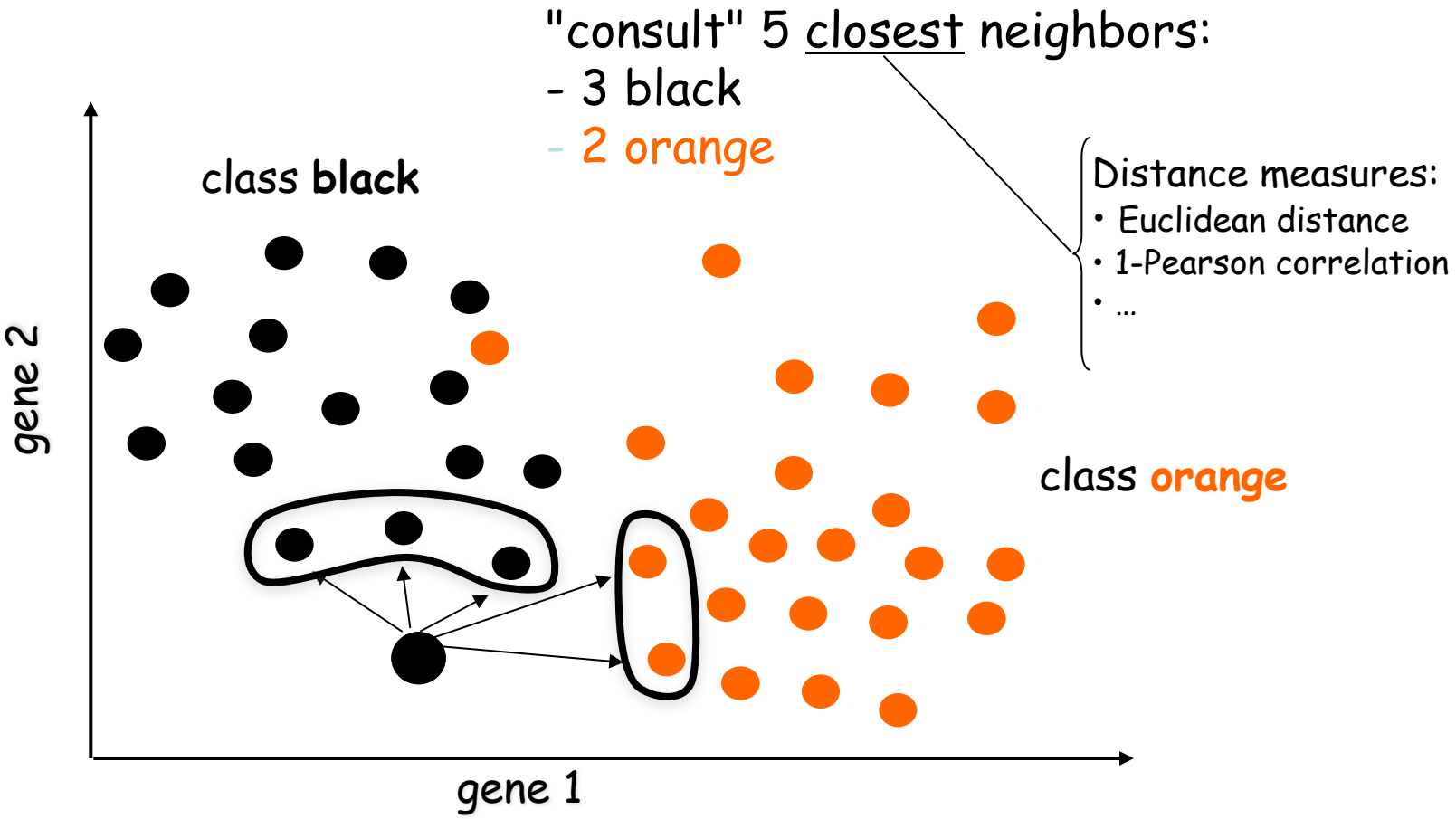
# kNN Classifier

Example: k=5, 2 genes, 2 classes



# kNN Classifier

Example: K=5, 2 genes, 2 classes



# Testing the Classifier

- Evaluation on independent test set
  - Build the classifier on the train set.
  - Assess prediction performance on test set.

- Performance measure

$$\text{error rate} = \frac{\text{\# of cases correctly classified}}{\text{total \# of cases}}$$

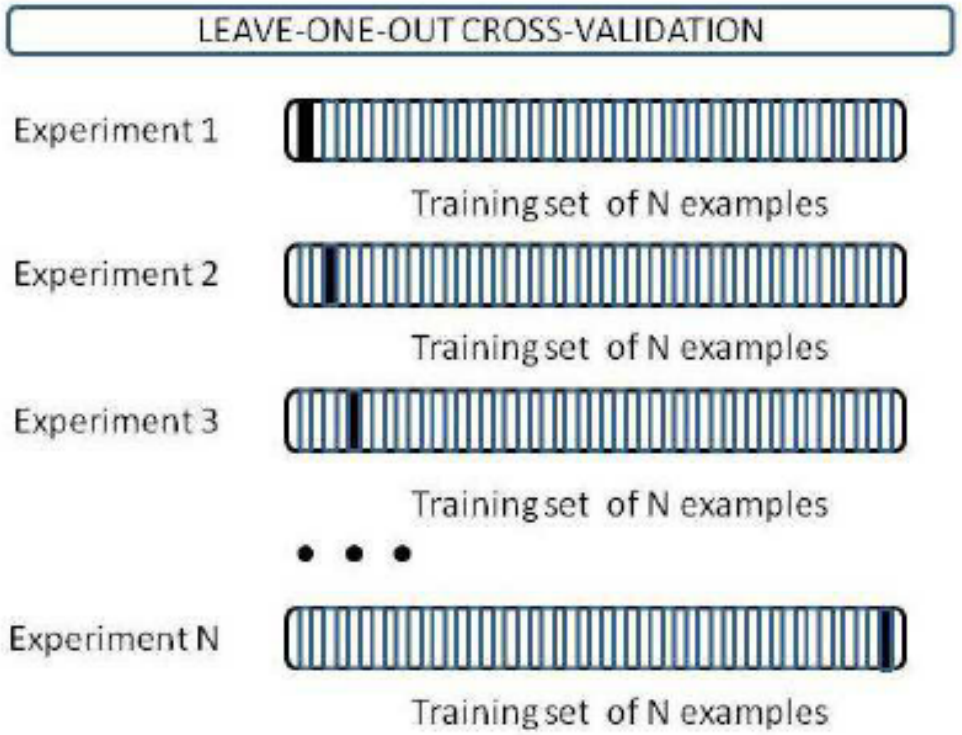
# Testing the Classifier

- Evaluation on independent test set
  - What if we don't have an independent test set?

- Cross Validation (XV):
  - Split the dataset into n folds (e.g., 10 folds of 10 samples each).
  - For each fold (e.g., for each group of 10 samples),
    - train (i.e., build model) on n-1 folds (e.g., on 90 samples),
    - test (i.e., predict) on left-out fold (e.g., on remaining 10 samples).
  - Combine test results.

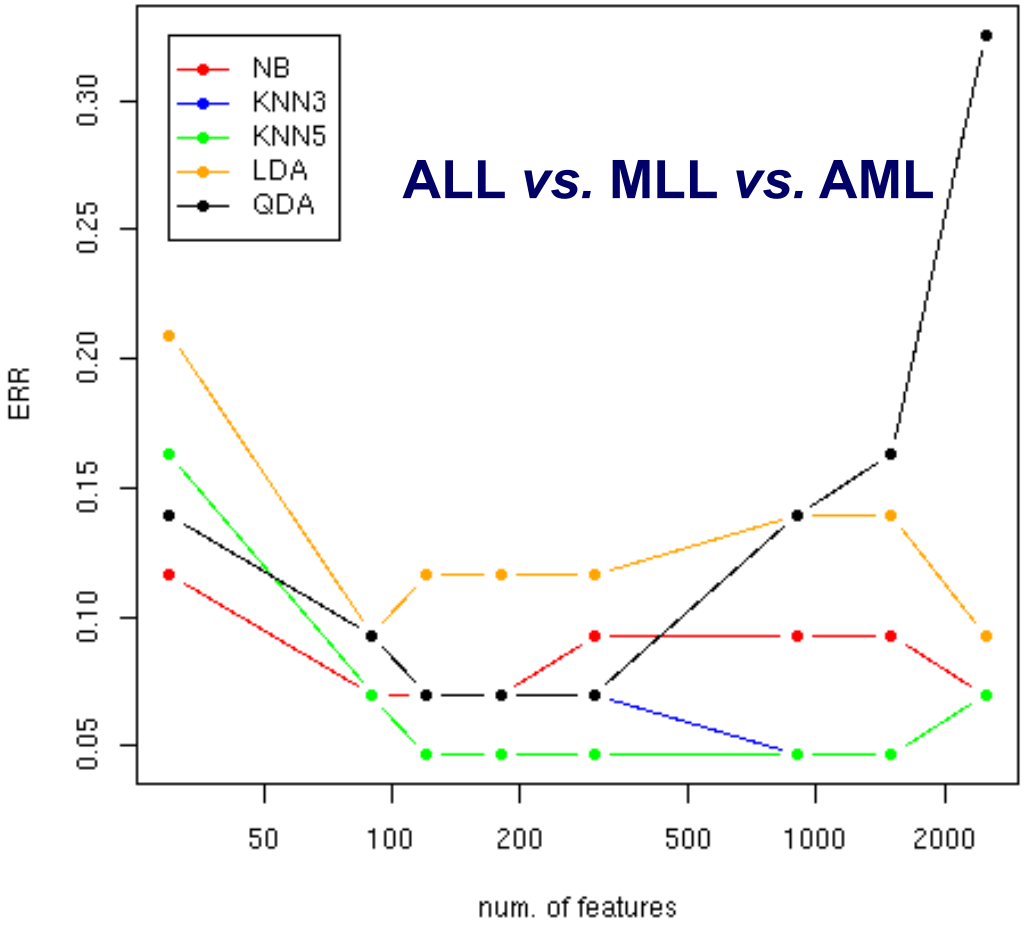
# Testing the Classifier

- Usually we us leave one out cross-validation



# Testing the Classifier

## Learning curves – leave one out cross validation




More features are not always better…

# Classification Notebook

- In the Notebook Repository, locate the notebook titled **2018-01-23_07 UBIC Classification and Prediction – RNAseq**

- Click on this notebook.
- Select “Get a copy”.
- Follow along with me