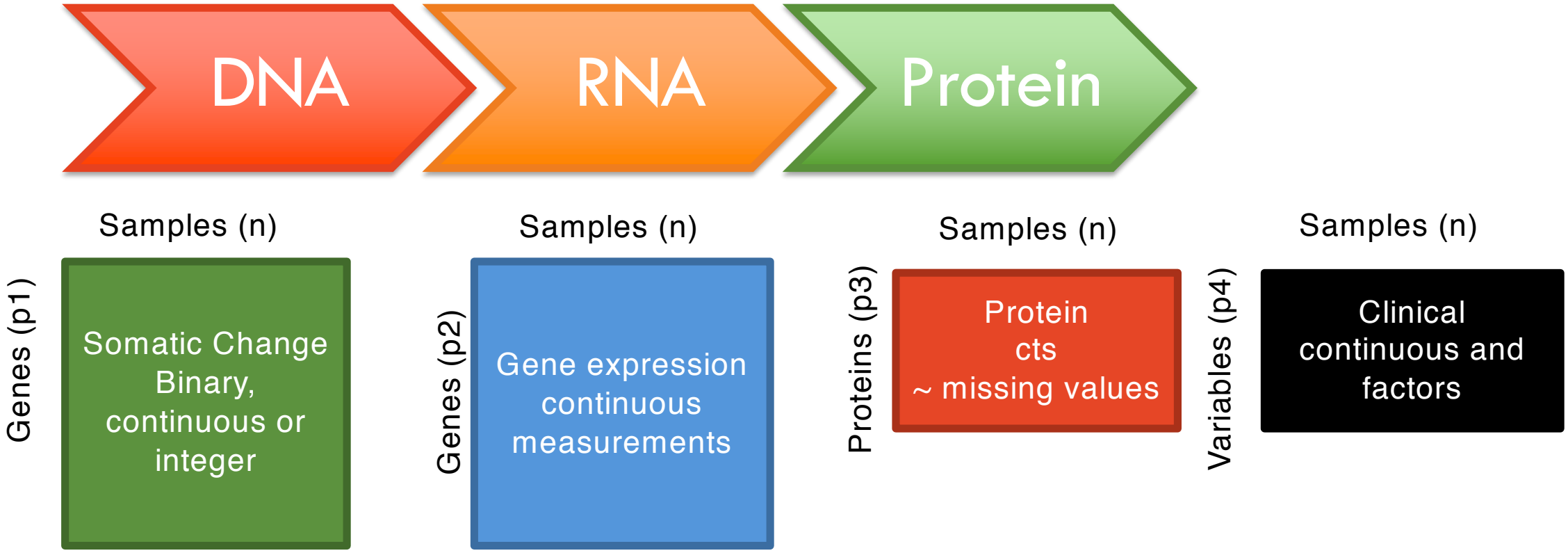
What does biomedical data look like?



Typical questions

How can we find **meaningful biological relationships** between these multiple datasets?

Possible input ?

Samples (n)

Genes (p1)

Somatic Change
Binary, continuous
or integer

Samples (n)

Genes (p2)

Gene expression
continuous
measurements

Samples (n)

Proteins (p3)

Protein
cts
~ missing values

Samples (n)

Variables (p4)

Clinical
continuous and
factors

Possible input ?

Samples (n)

Genes (p1)

Somatic Change
Binary, continuous
or integer

Samples (n)

Genes (p2)

Gene expression
continuous
measurements

Samples (n)

Proteins (p3)

Protein
cts
~ missing values

Samples (n)

Variables (p4)

Clinical
continuous and
factors

Gene expression
continuous
measurements

Protein
cts
~ missing values

Gene expression
continuous
measurements

Clinical
continuous and
factors

Protein
cts
~ missing values

Somatic Change
Binary, continuous
or integer

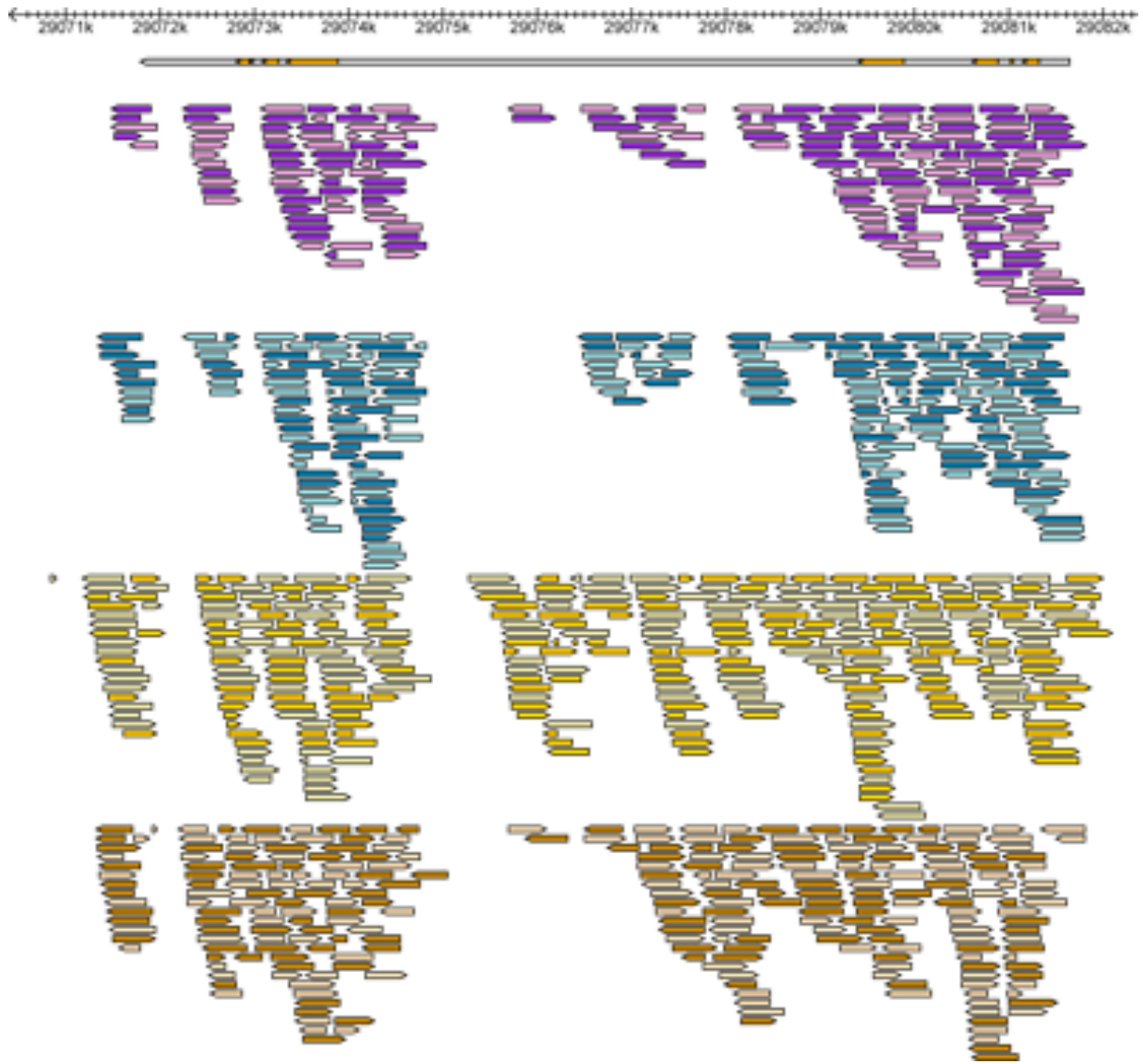
Gene expression
continuous
measurements

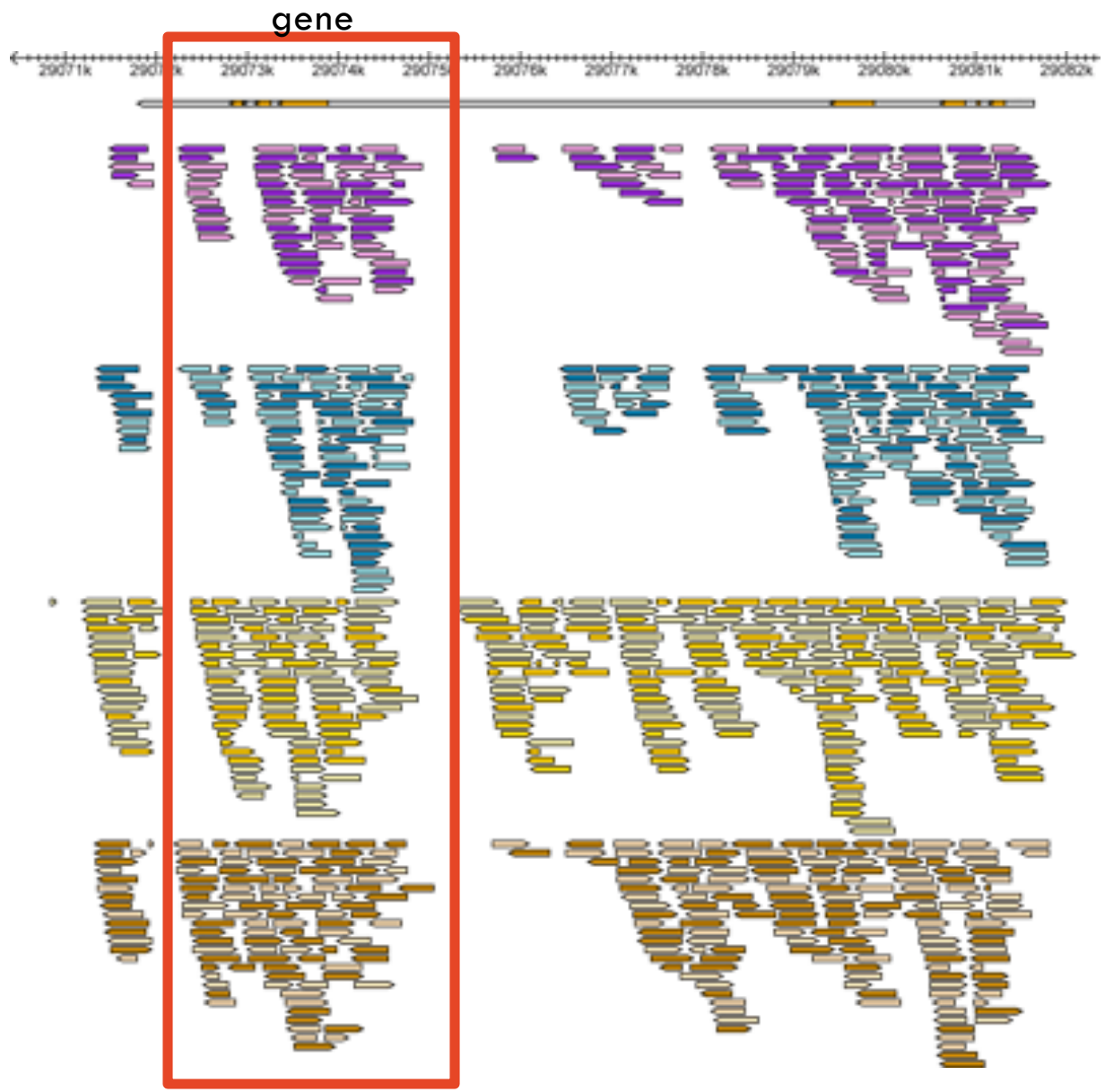
Clinical
continuous and
factors

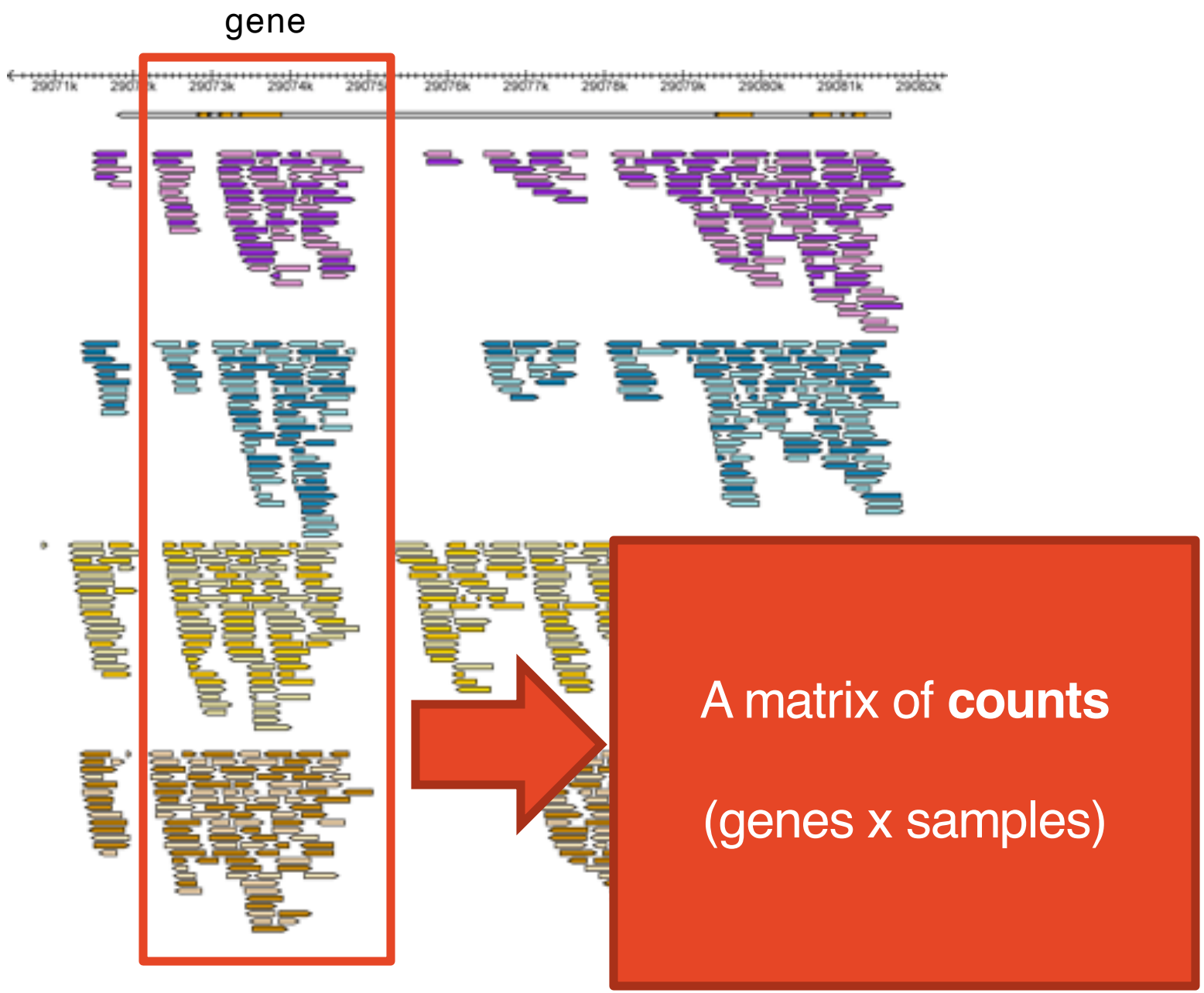
Steps in preparing an RNA-seq library

1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Fragment RNA (200 bp)
4. Convert to cDNA by random priming
5. Apply adaptors and sequence
6. Analyze millions of 25 bp reads

geospiza
FROM SAMPLE TO RESULTS™


29071k
29072k
29073k
29074k
29075k
29076k
29077k
29078k
29079k
29080k
29081k
29082k


gene
29071k
29073k
29074k
29076k
29077k
29078k
29079k
29080k
29081k
29082k


gene
29071k
29073k
29074k
29076k
29077k
29078k
29079k
29080k
29081k
29082k
A matrix of **counts**
(genes x samples)

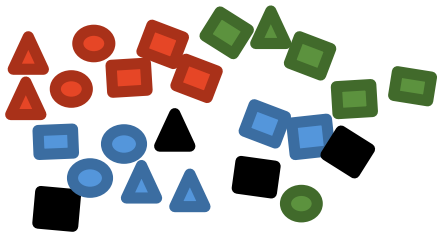
Basic principles of discrimination

Each object associated with

- a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and
- a feature vector of P measurements: $\mathbf{X} = (X_1, \ldots, X_P)$

Aim: predict Y from **X**.

Predefined class {1,2,…K}



Objects

Y = Class Label = 2

X = Feature vector
{colour, shape}

Classification rule ?

X = {red, square}

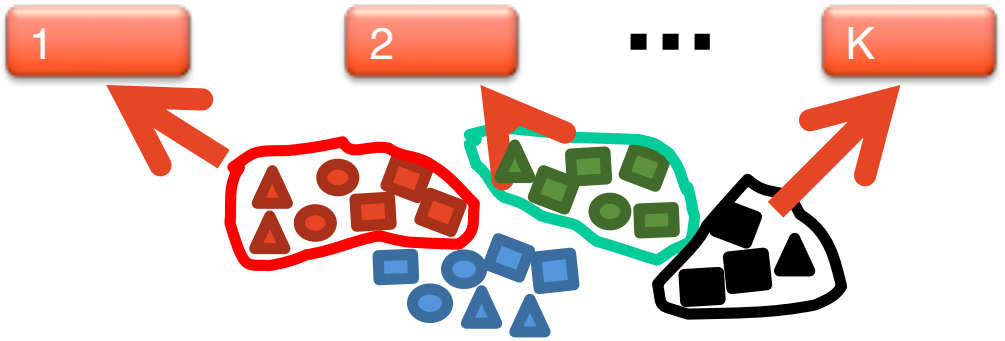
Y = ?

Basic principles of discrimination

Each object associated with

- a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and
- a feature vector of P measurements: $\mathbf{X} = (X_1, \ldots, X_P)$

Aim: predict Y from **X**.



Predefined class {1,2,…K}

Objects

Y = Class Label = 2

X = Feature vector
{colour, shape}

Classification rule ?

X = {red, square}

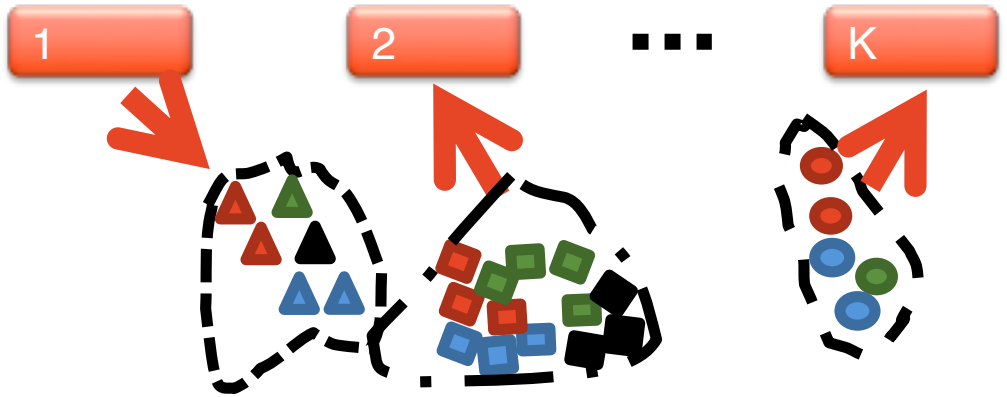
Y = ?

Basic principles of discrimination

Each object associated with

- a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and
- a feature vector of P measurements: $\mathbf{X} = (X_1, \ldots, X_P)$

Aim: predict Y from **X**.



Predefined class {1,2,…K}

Objects

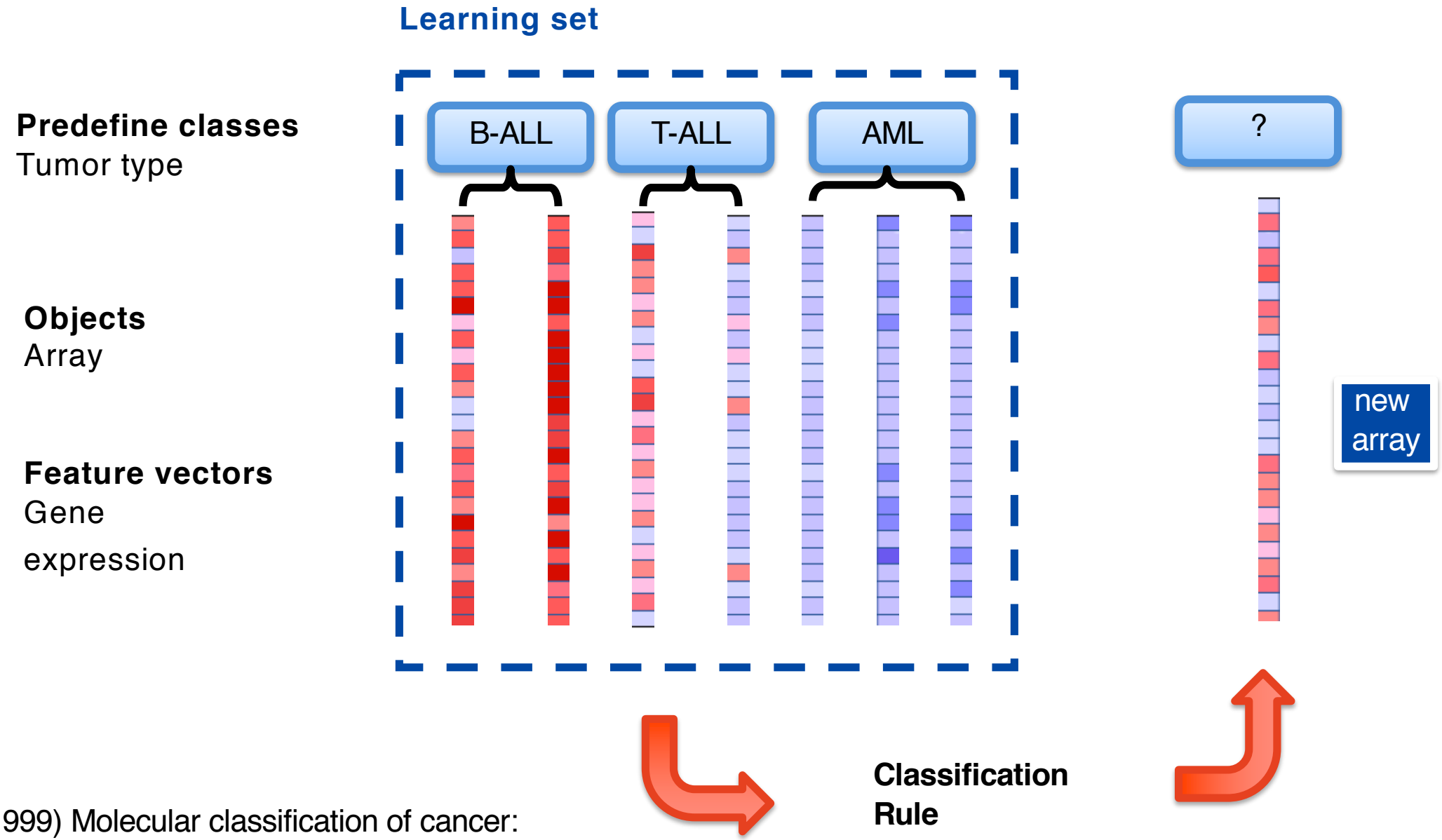
Y = Class Label = 2

X = Feature vector
{colour, shape}

Classification rule ?

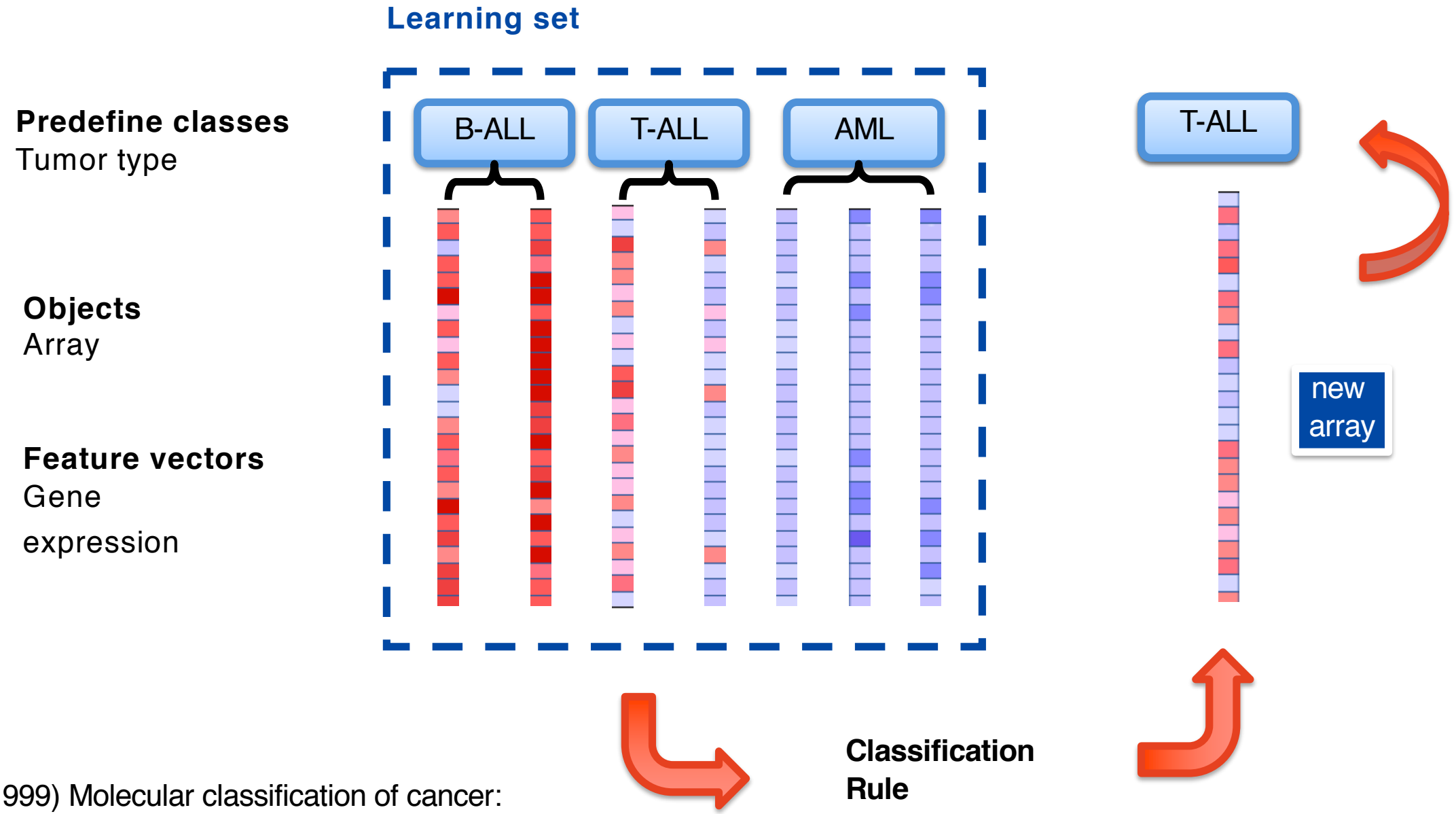
X = {red, square}

Y = ?



Reference

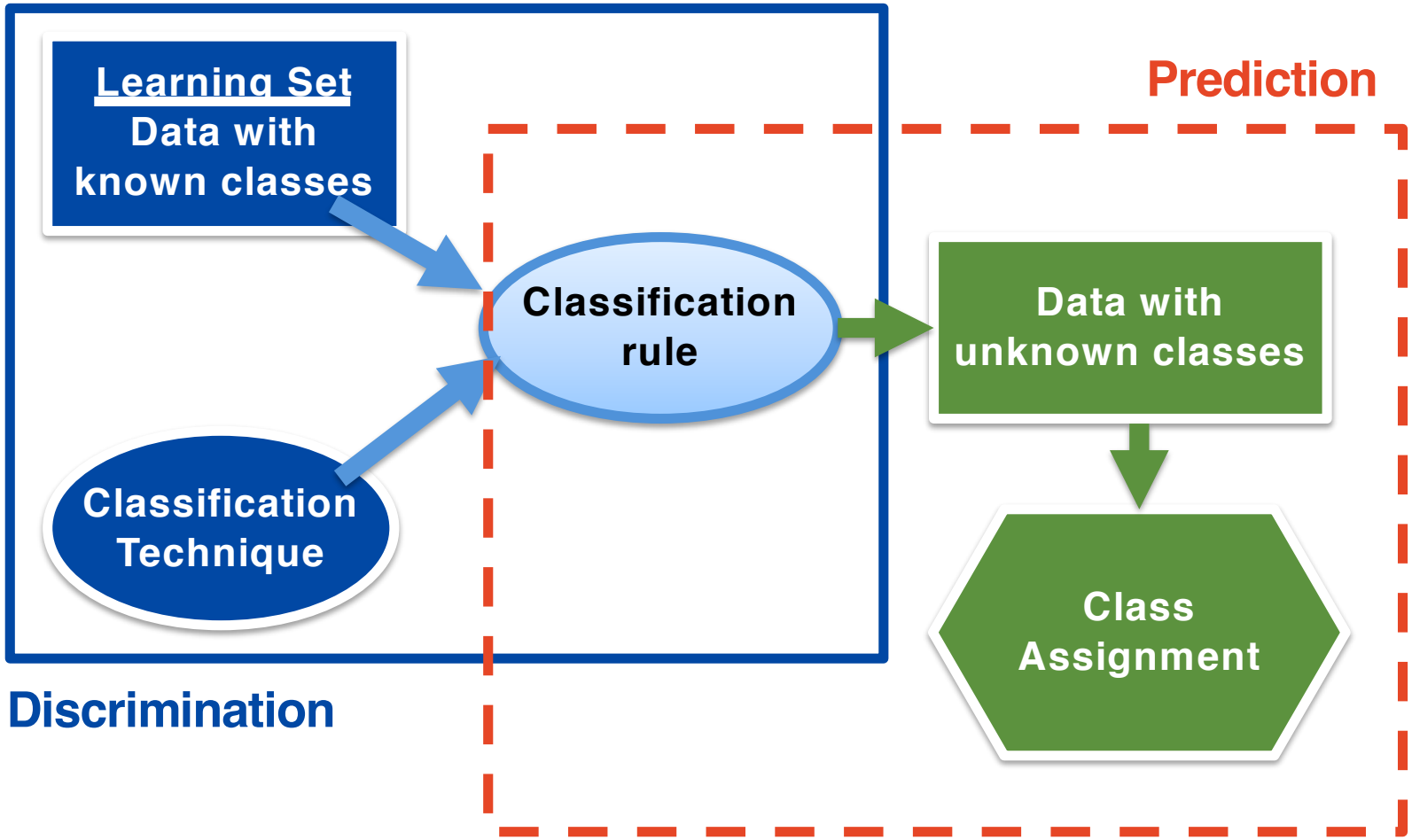
Golub et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537.



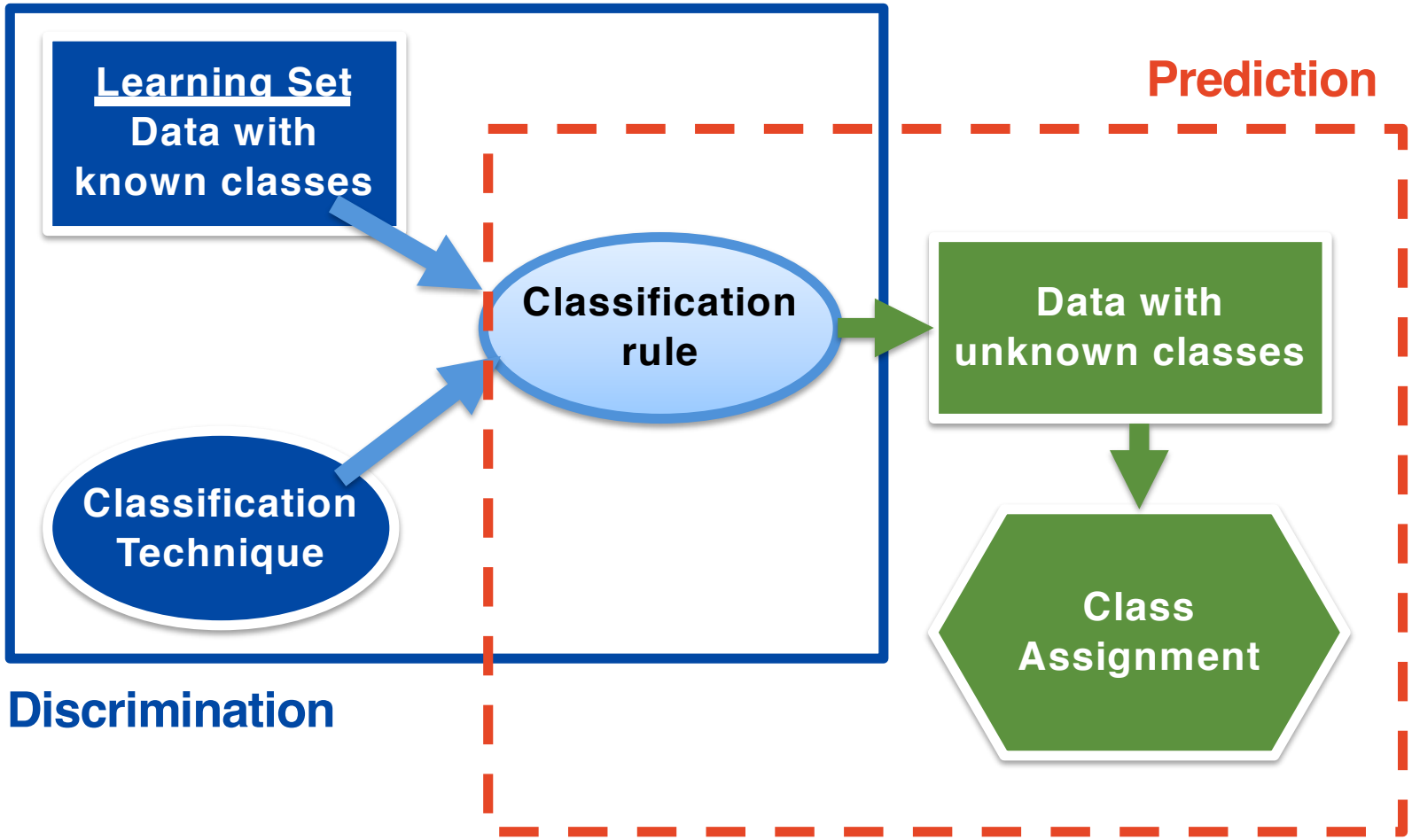
Reference

Golub et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537.

Discrimination and prediction



Discrimination and prediction

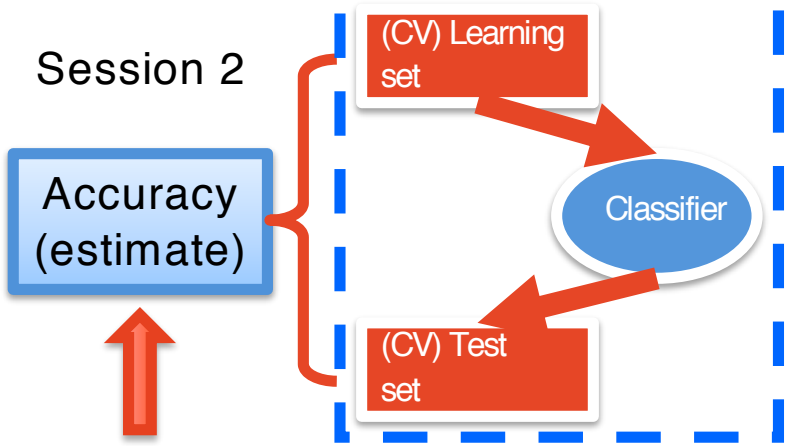


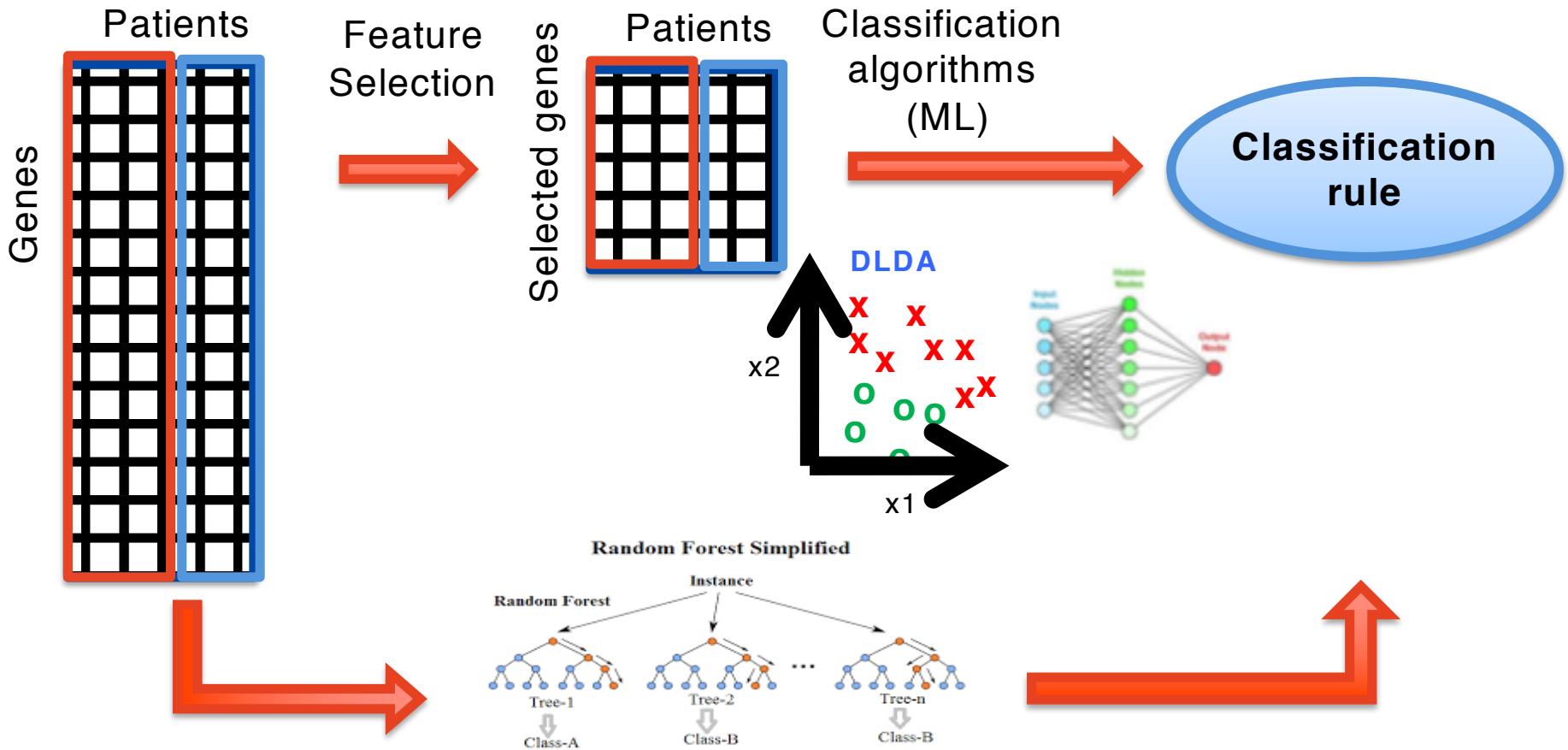
One can think of the classification rule as a black box, some methods provides more insight into the box.

Classification Rule

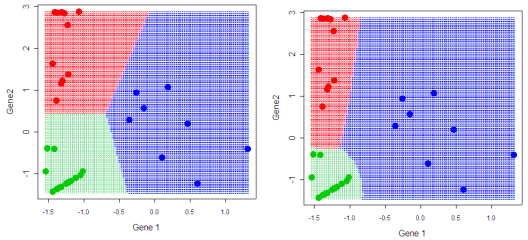
- Classification procedure,
- Feature selection,
- Parameters
- Distance measure,
- Aggregation methods.
- pthers …

Classification





Many algorithms out there …



y_i

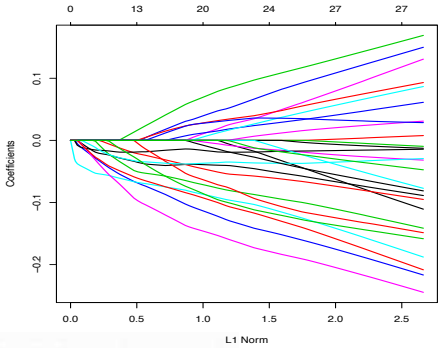
KNN

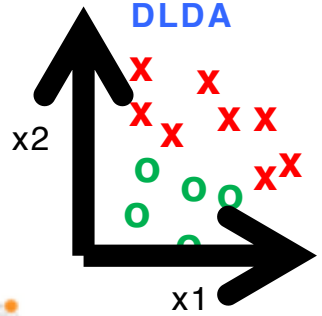
$$\text{where } i = \underset{i}{\text{argmin}}\ K(\hat{\mathbf{x}}_i, \mathbf{x})$$

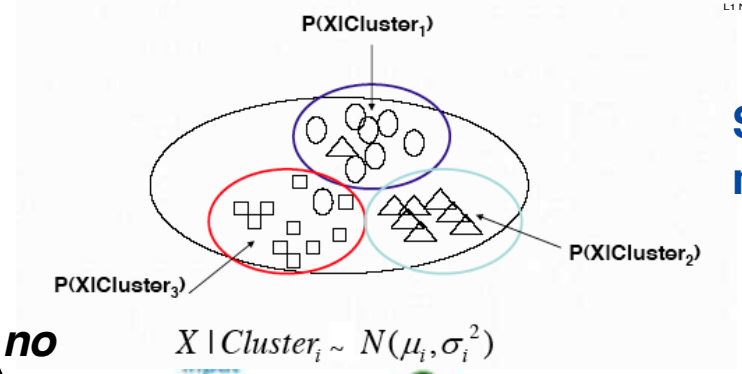
logistic / lasso

$$\text{maximize} \sum_i \log(P(y_i \mid \mathbf{x}, \boldsymbol{\theta})) + \lambda \|\boldsymbol{\theta}\|$$

$$\text{where } P(y_i \mid \mathbf{x}, \boldsymbol{\theta}) = 1/(1+\exp(-y_i \boldsymbol{\theta}^T \mathbf{x}))$$





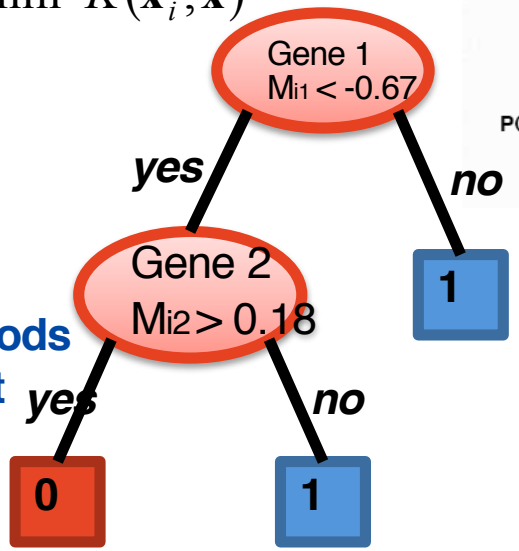


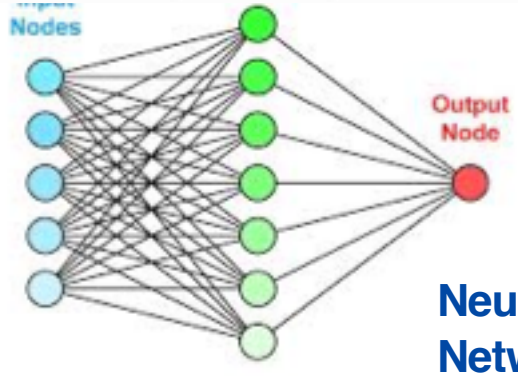
$$X \mid Cluster_i \sim N(\mu_i, \sigma_i^2)$$

Statistical models

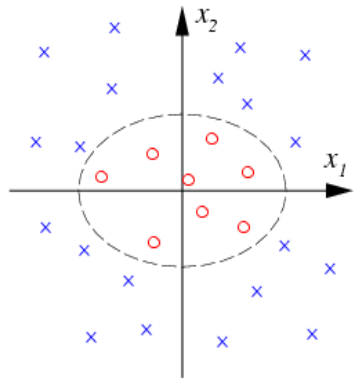
Tree based methods

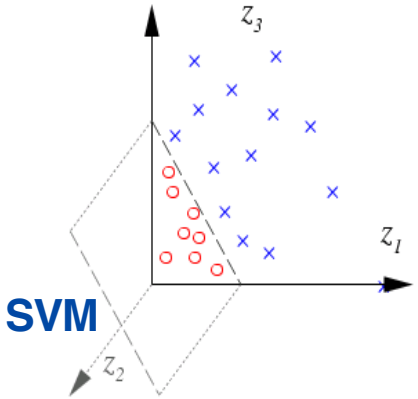
- random forest
- adaboost
- bagging





Neural Network





SVM

Some code limma stuff DE for feature selection

- Do show MA-plot + vocano plot and describe what it means.
- The difference between DE and biomarker.
- Don’t forget logistic regression

ML discriminant rules - special cases

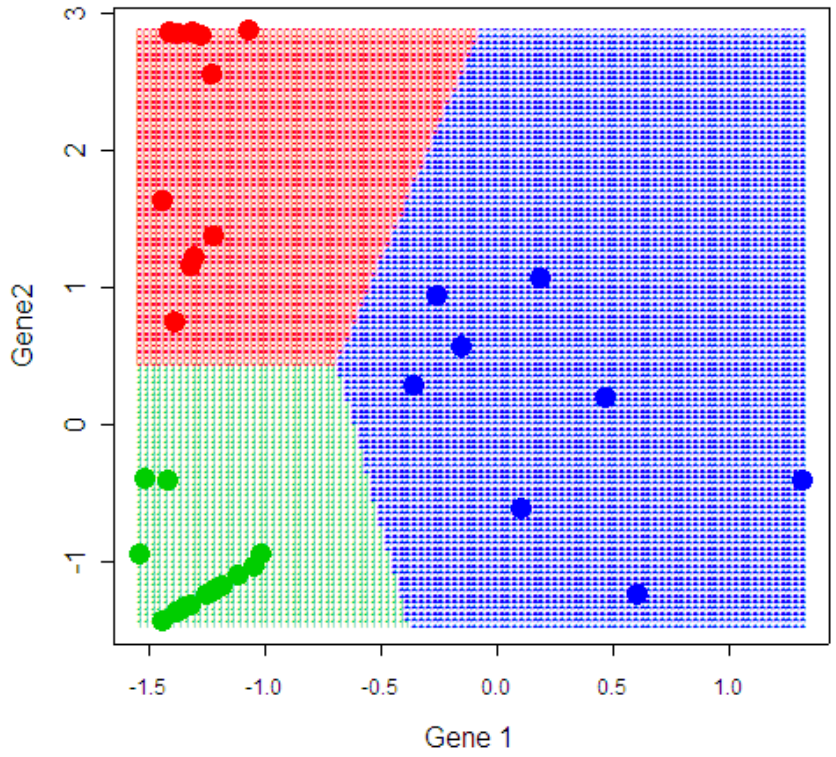
Gaussian ML discriminant rules

- For multivariate Gaussian (normal) class densities $\mathbf{X}|Y=k \sim N(\mu_k, \Sigma_k)$, the ML classifier is

$$C(\mathbf{X}) = \text{argmin}_k \{(\mathbf{X} - \mu_k)\ \Sigma_k^{-1}\ (\mathbf{X} - \mu_k)' + \log|\ \Sigma_k\ |\}$$

- In general, this is a quadratic rule (Quadratic discriminant analysis, or QDA)

- In practice, population mean vectors μ_k and covariance matrices Σ_k are estimated by corresponding sample quantities



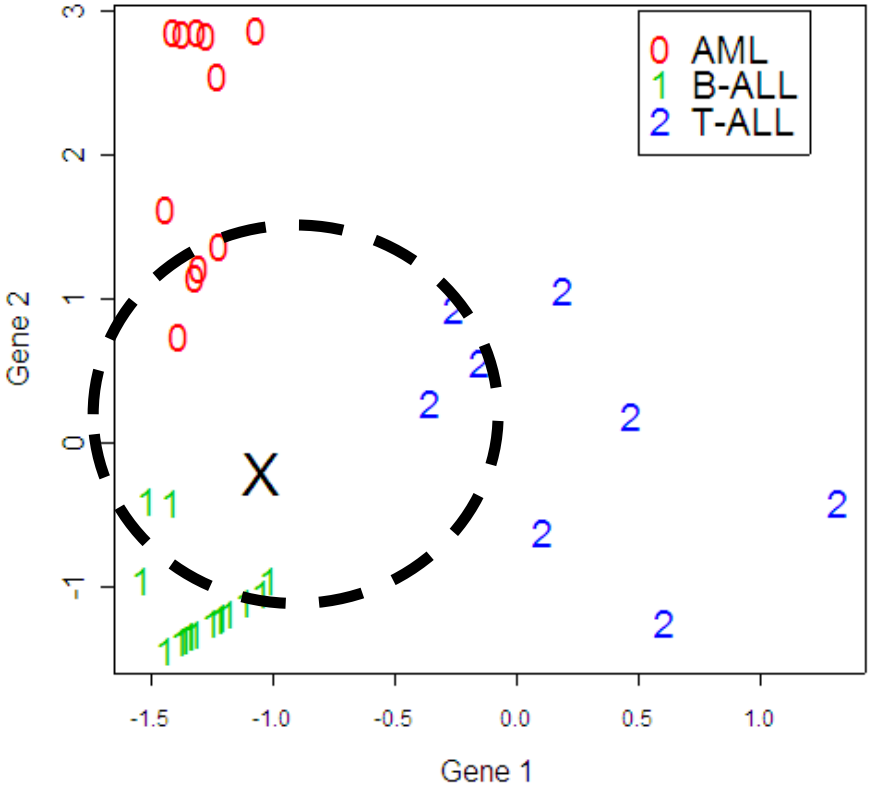
[DLDA]
Diagonal linear discriminant analysis
class densities have the same diagonal covariance matrix $\nabla = \text{diag}(s_1^2, \ldots, s_p^2)$

Nearest neighbor classification

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation).

- k-nearest neighbor rule (Fix and Hodges (1951)) classifies an observation **X** as follows:
  - find the k observations in the learning set closest to **X**
  - predict the class of **X** by majority vote, i.e., choose the class that is most common among those k observations.

- The number of neighbors k can be chosen by cross-validation (more on this later).

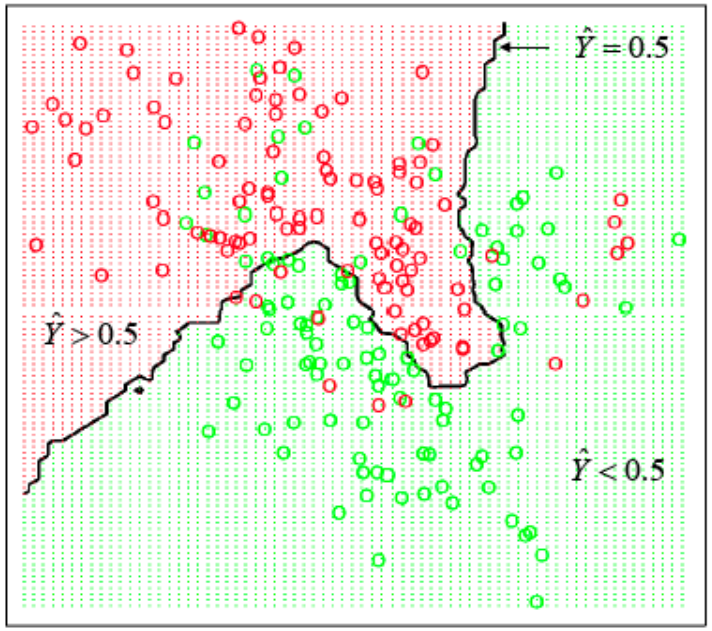


Classification - k Nearest Neighbor

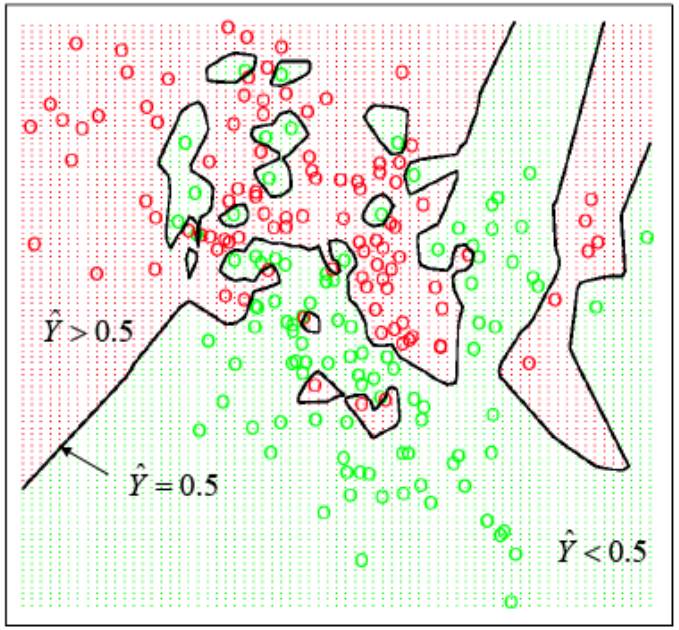
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

$N_k(x)$: the k closest points to x

15-nearest neighbor averaging

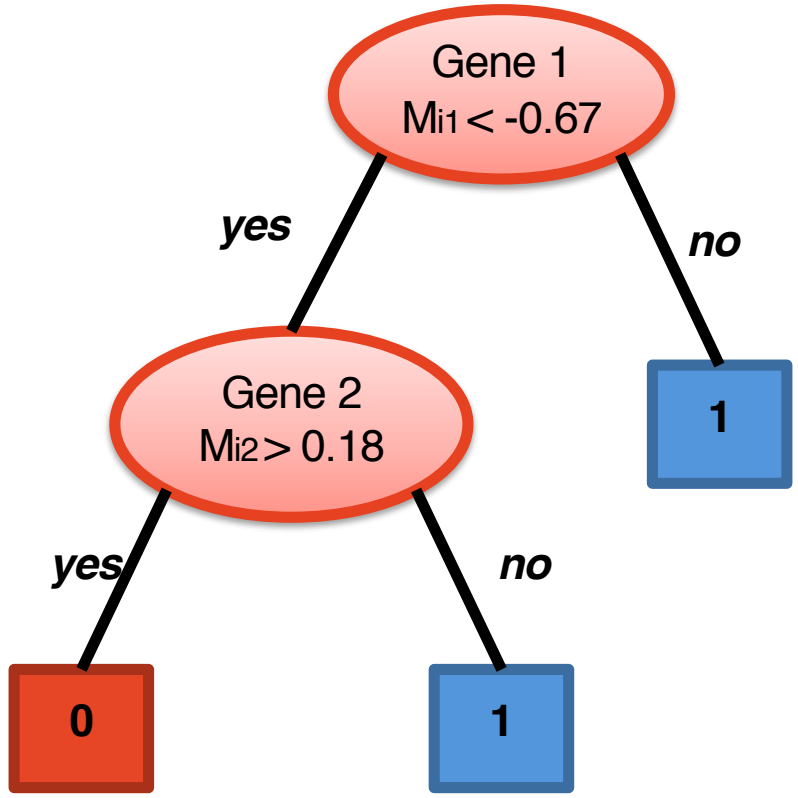


1-nearest neighbor averaging

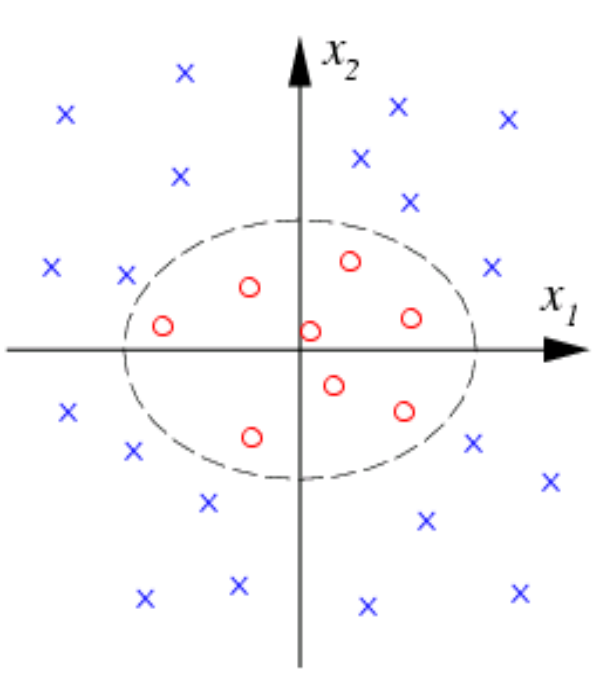


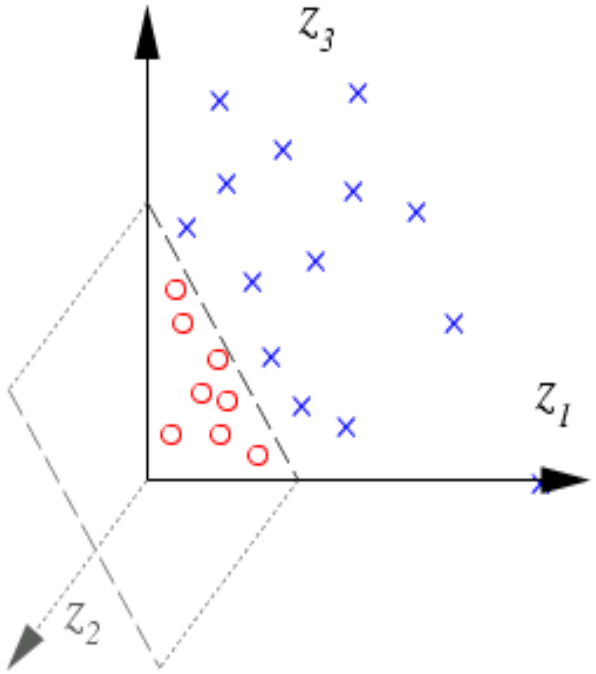
Classification tree

- Partition the feature space into a set of rectangles, then fit a simple model in each one
- Binary tree structured classifiers are constructed by repeated splits of subsets (nodes) of the measurement space $\underline{X}$ into two descendant subsets (starting with $\underline{X}$ itself)
- Each terminal subset is assigned a class label; the resulting partition of $\underline{X}$ corresponds to the classifier



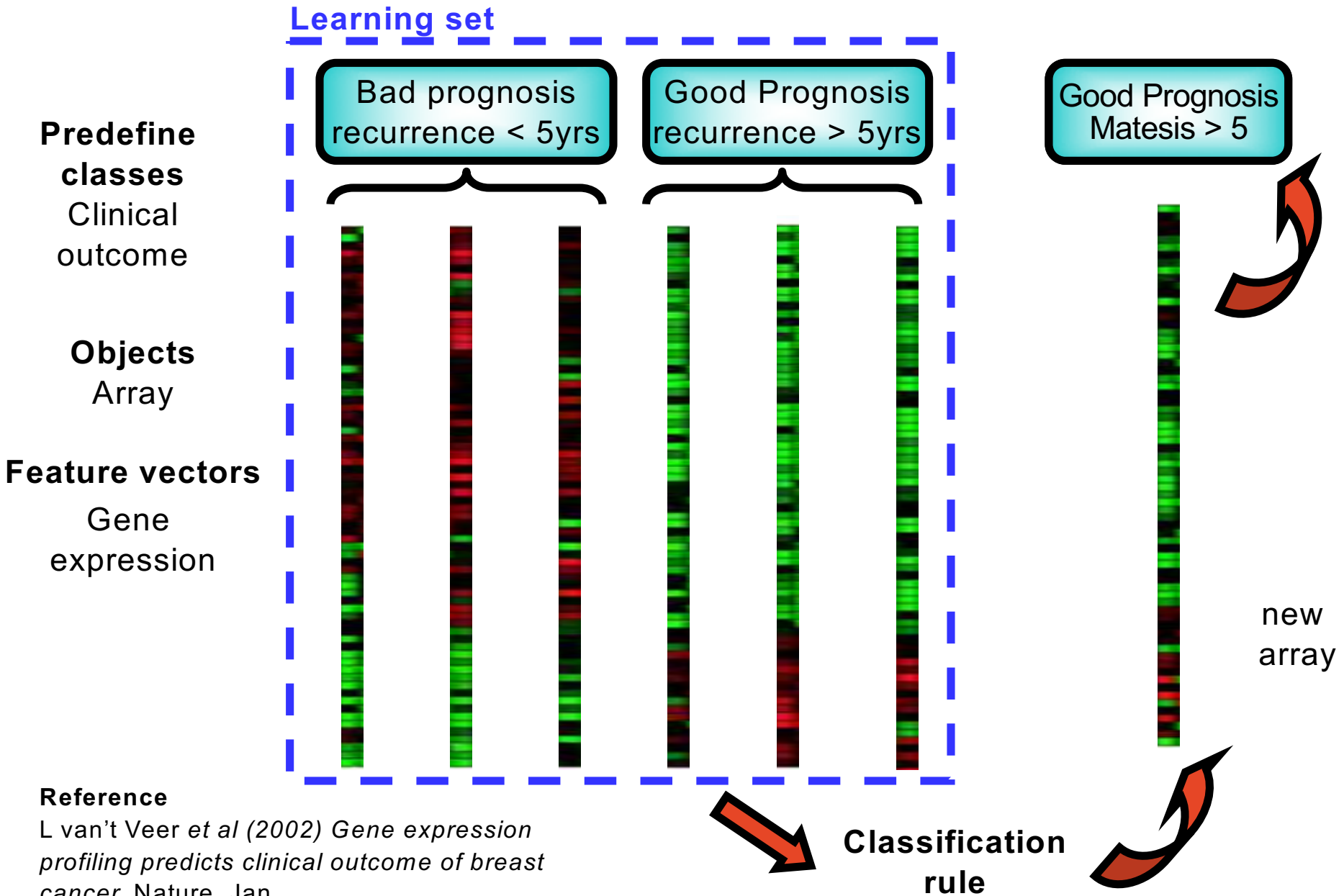
Classification with SVMs





A quick summary (update with package)

- Classical Maximum Likelihood classifiers:
  - Linear Discriminant Analysis (LDA)
  - DLDA
  - DQDA
  - K-Nearest Neighbour Classifiers
- Modern LDA Derivatives:
  - PAMR (http://www-stat.stanford.edu/~tibs/PAM/)
  - SCRDA
- Support Vector Machines (SVM)
- Aggregated Trees (CART)
- Other classifiers:
  - Neural networks (NN)
  - Bayesian belief networks



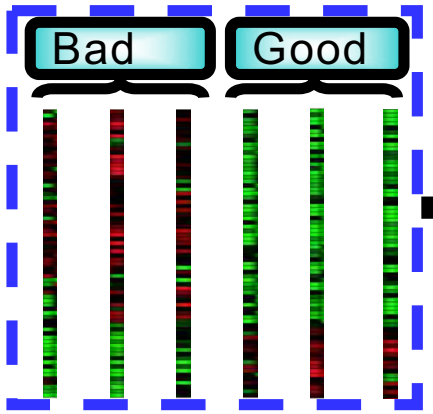
Reference

L van't Veer *et al (2002) Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan.

.

Case studies

Learning set



Classification Rule

Feature selection.
Correlation with class labels, very similar to t-test.

Using cross validation to select 70 genes

295 samples selected
from Netherland Cancer Institute
tissue bank (1984 – 1995).

Results" Gene expression profile is a more powerful predictor then standard systems based on clinical and histologic criteria

Agendia (formed by reseachers from the Netherlands Cancer Institute)
Start in Oct, 2003

1) 3000 subjects [Health Council of the Netherlands]
2) 5000 subjects New York based Avon Foundation.

Custorm arrays are made by Aglient including
70 genes + 1000 controls

Reference 1
Retrospective study
L van't Veer *et al* *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan 2002.
.

Reference 2
Retrospective study
M Van de Vijver *et al*. A gene expression signature as a predictor of survival in breast cancer. The New England Jouranl of Medicine, Dec 2002.

Reference 3
Prospective trials.
Aug 2003
Clinical trials
http://www.agendia.com/

Session 2
Performance assessment

THE UNIVERSITY OF SYDNEY

Performance assessment

- Any classification rule needs to be evaluated for its performance on the future samples. It is almost never the case in microarray studies that a large independent population-based collection of samples is available at the time of initial classifier-building phase.

- One needs to estimate future performance based on what is available: often the same set that is used to build the classifier.

- Assessing performance of the classifier based on
  - Cross-validation.
  - Test set.
  - Independent testing on future dataset.
  - Independent testing on existing dataset (integrative analysis).

Diagram of performance assessment

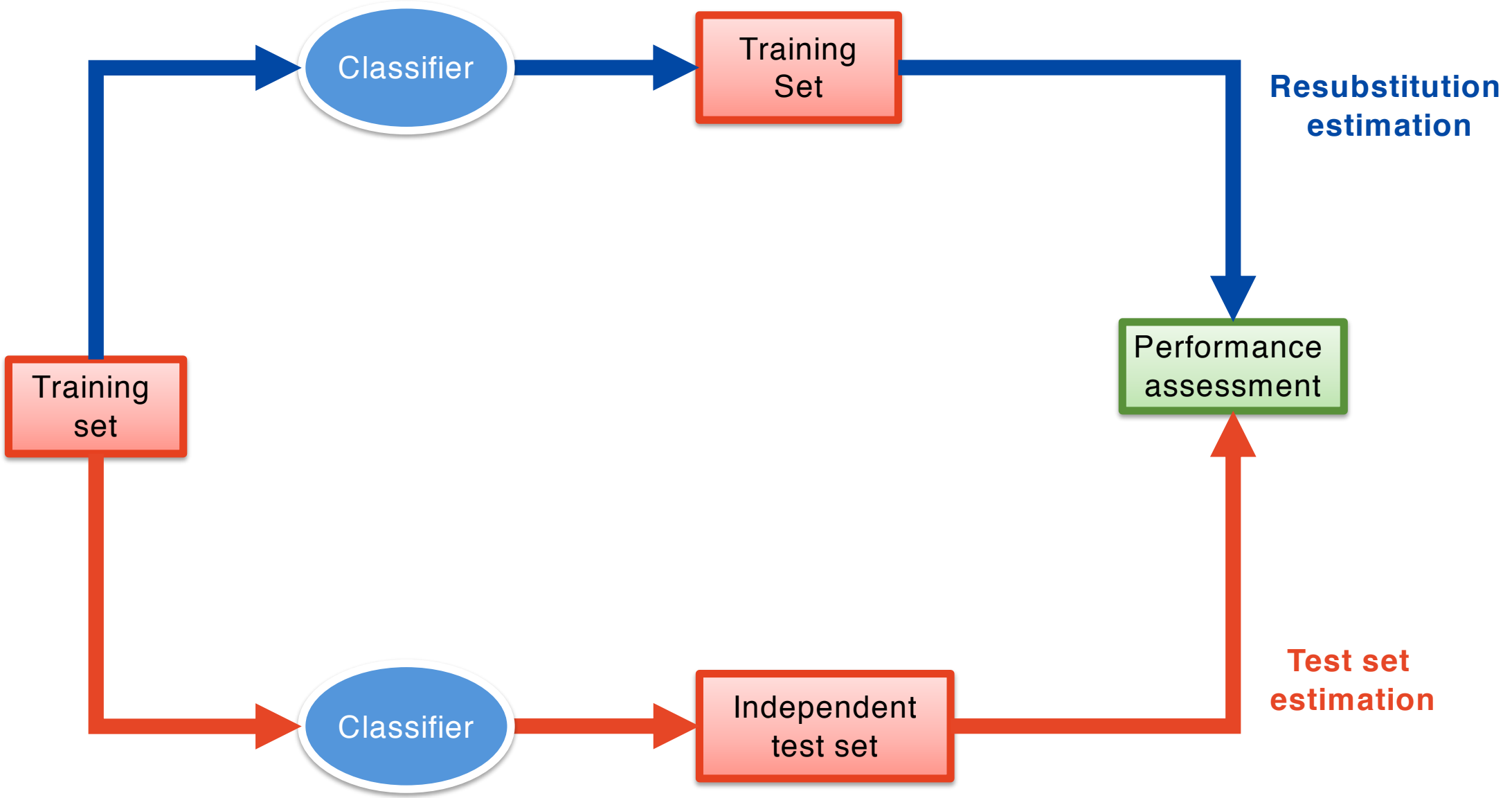
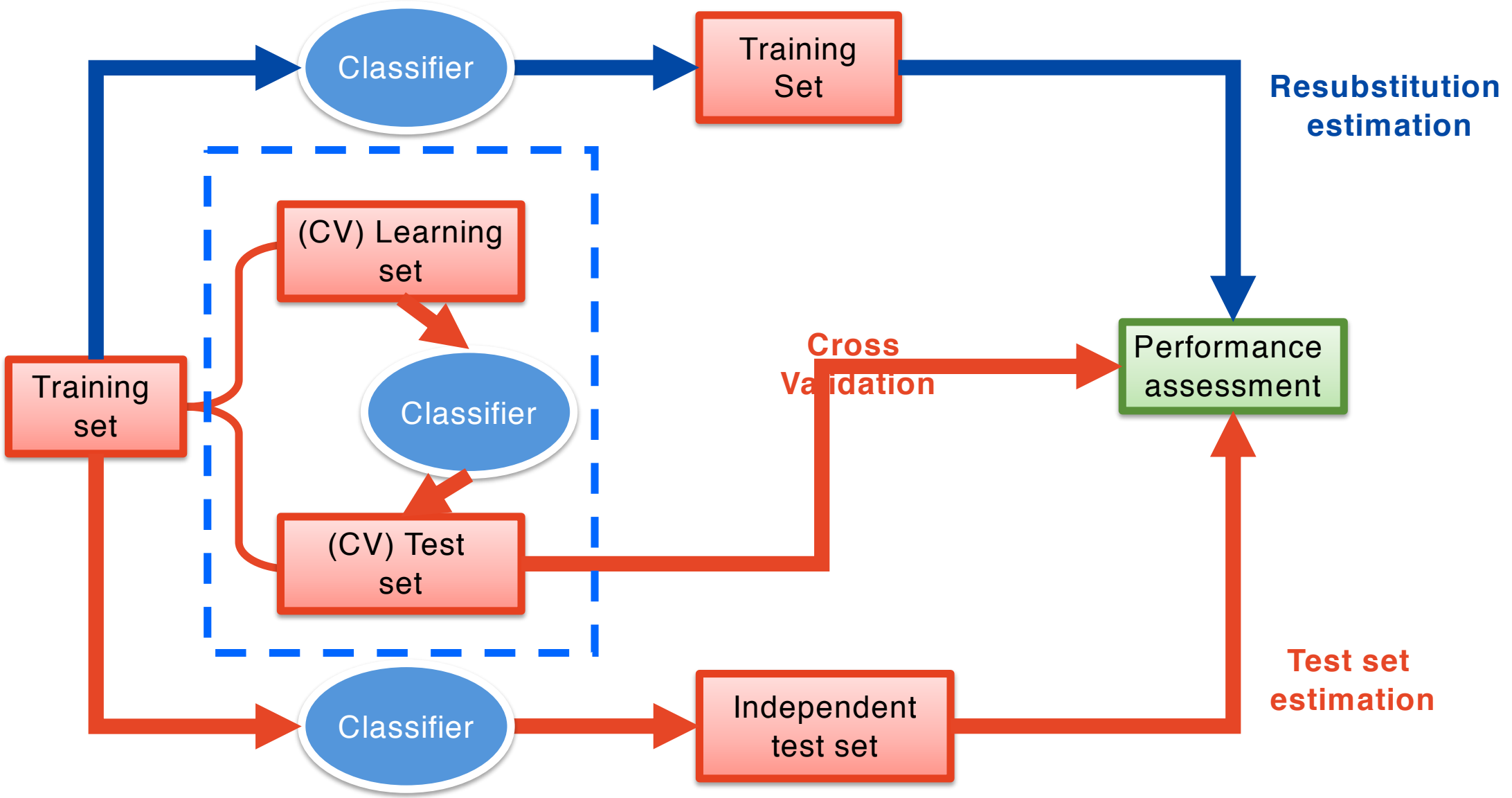
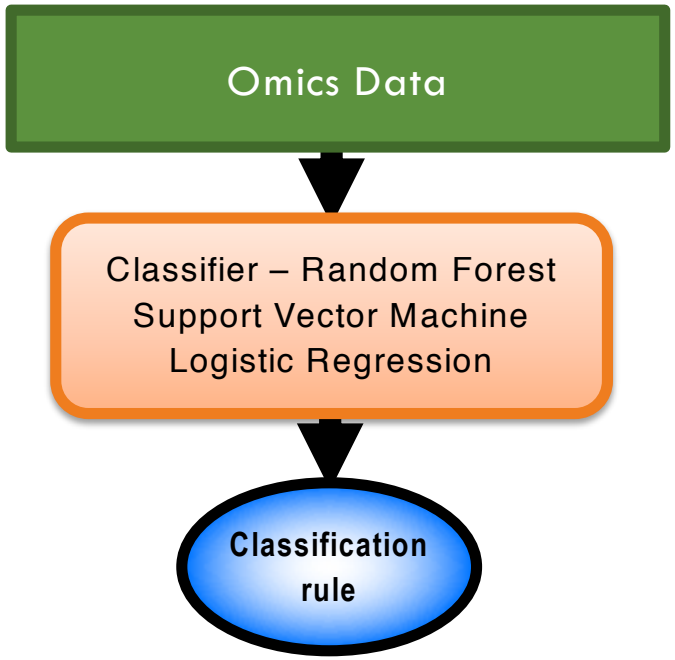
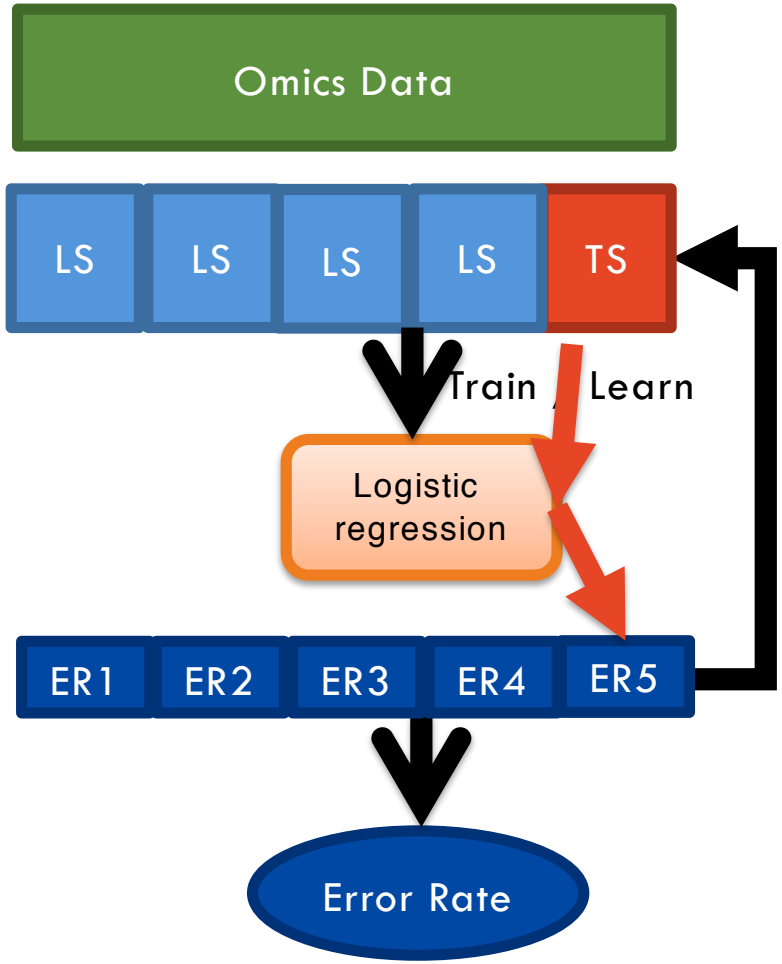


Diagram of performance assessment

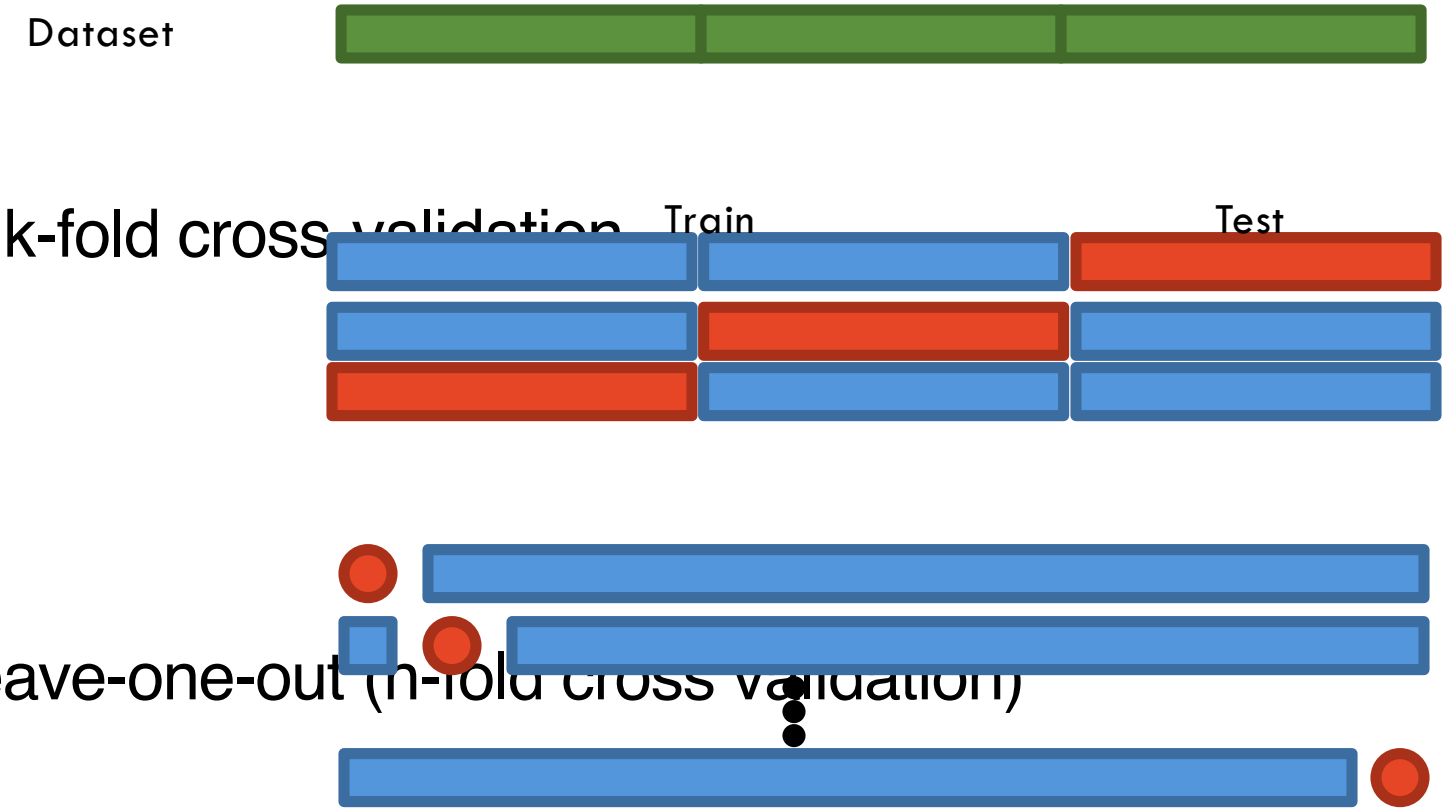


5-fold CV


Omics Data
Classifier – Random Forest
Support Vector Machine
Logistic Regression
Classification rule


Omics Data
LS
LS
LS
LS
TS
Train
Learn
Logistic regression
ER1
ER2
ER3
ER4
ER5
Error Rate

Common Splitting Strategies

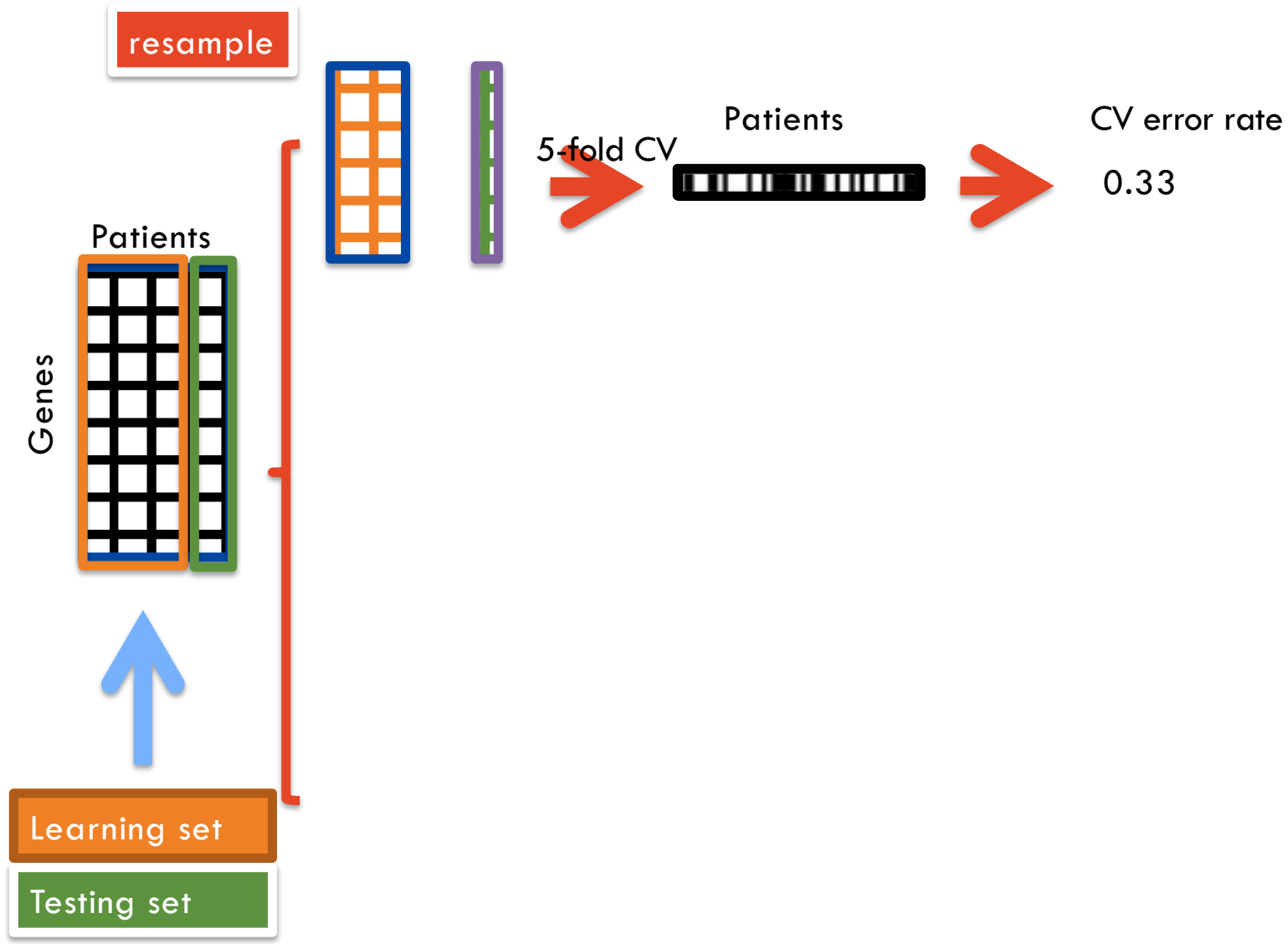


- k-fold cross validation

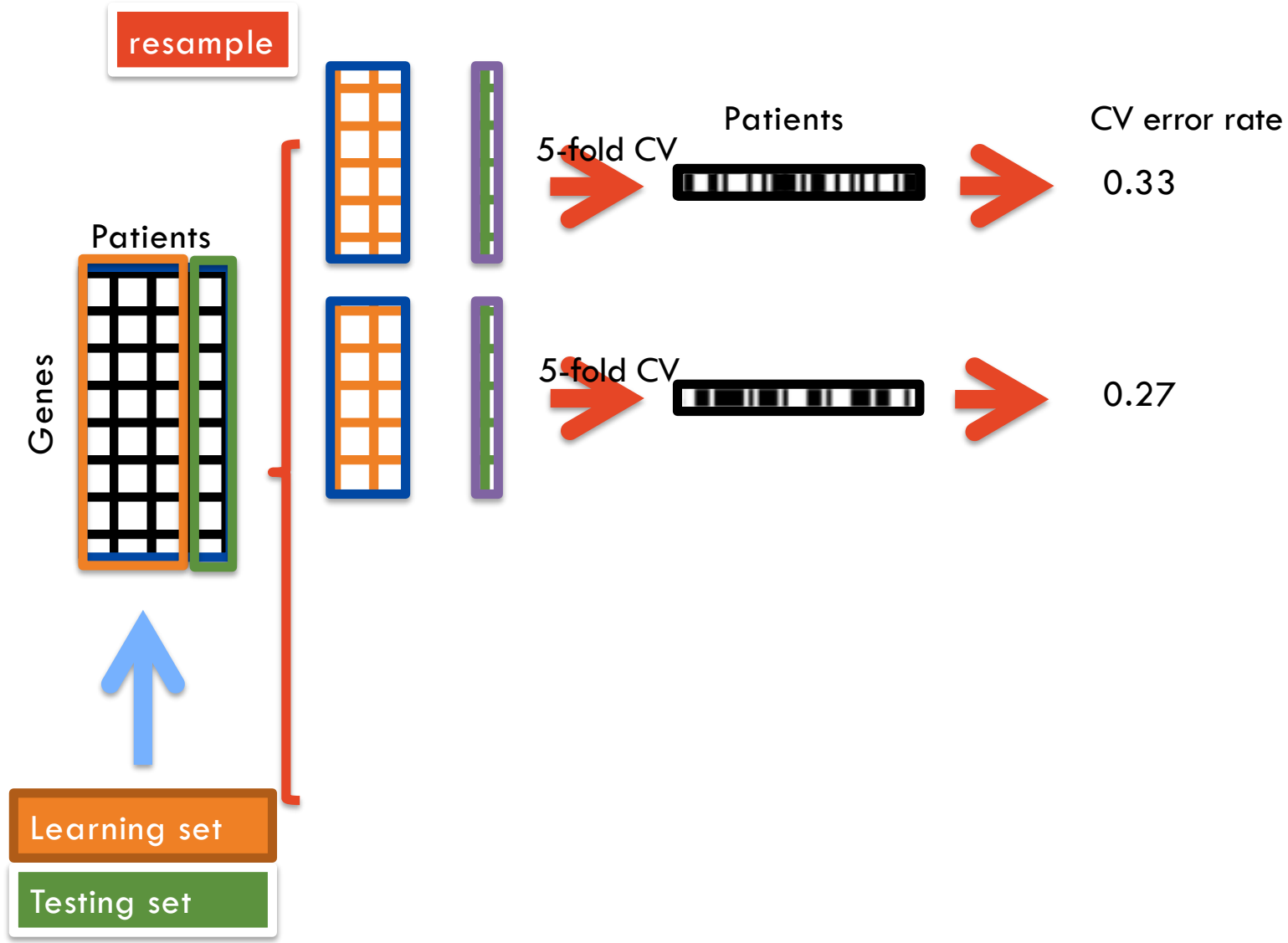
- Leave-one-out (n-fold cross validation)

Common question

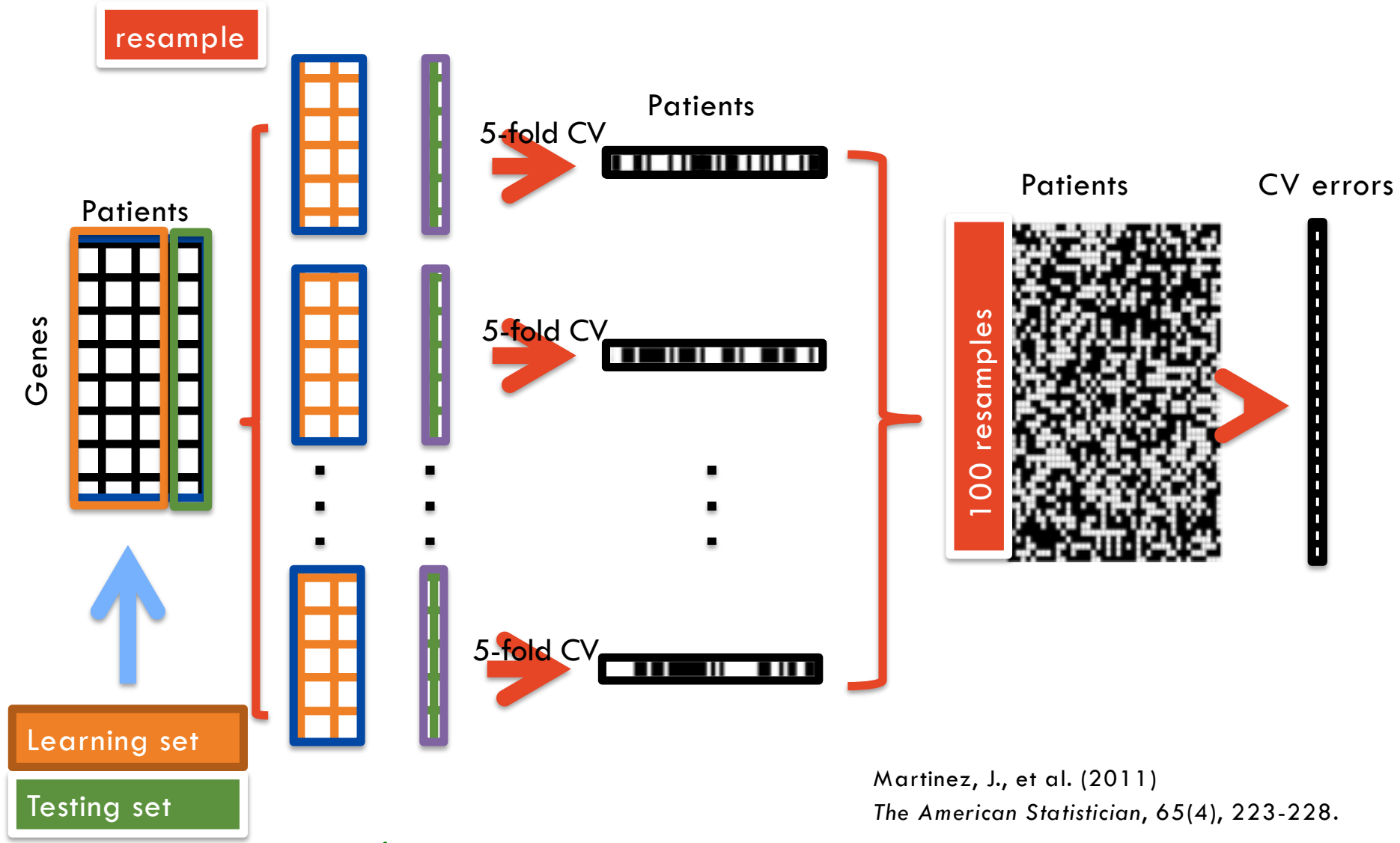
Average 5-fold cross validation



Average 5-fold cross validation



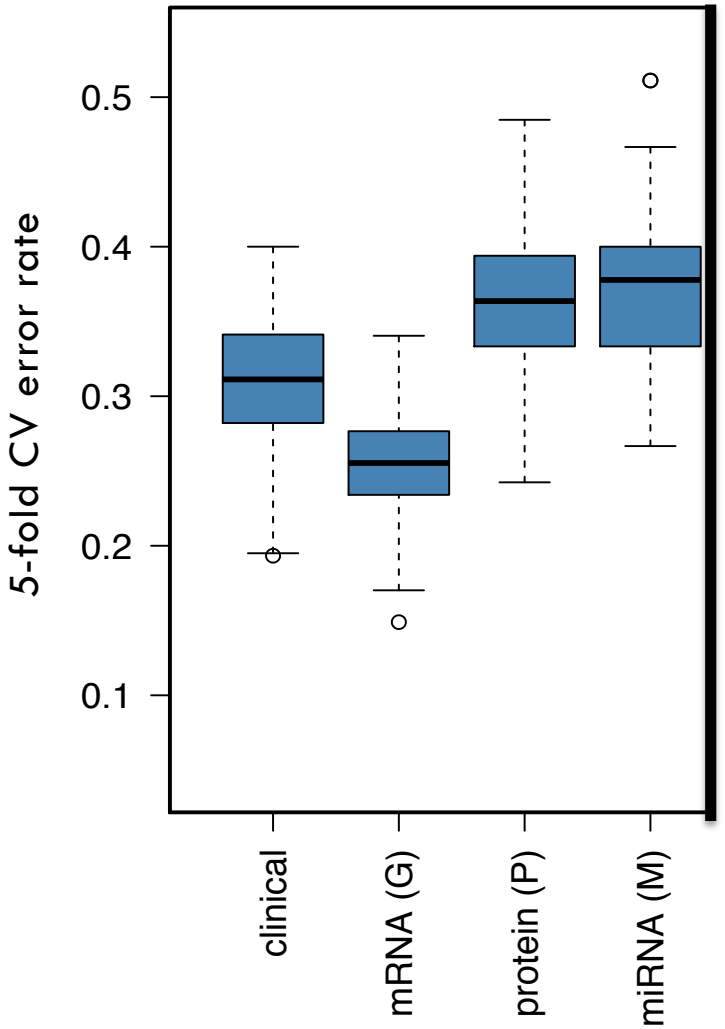
Average 5-fold cross validation



Martinez, J., et al. (2011)
The American Statistician, 65(4), 223-228.

`Rpackage: ClassifyR`

Platforms comparison



Individual platform

- Clinical: CPM
- G: mRNA
- P: protein
- M: microRNA

Jayawardana et al, IJC, 2014

Average 5-fold cross validation