Analysis of publicly available microarray data

20th-21st, February 2017

University of Cambridge, Cambridge, UK

# Classification and Survival analysis

***butions by Matt Ritchie, Christina Curtis, Jean Yang and Stephen Eglen***)

*Oscar M. Rueda*

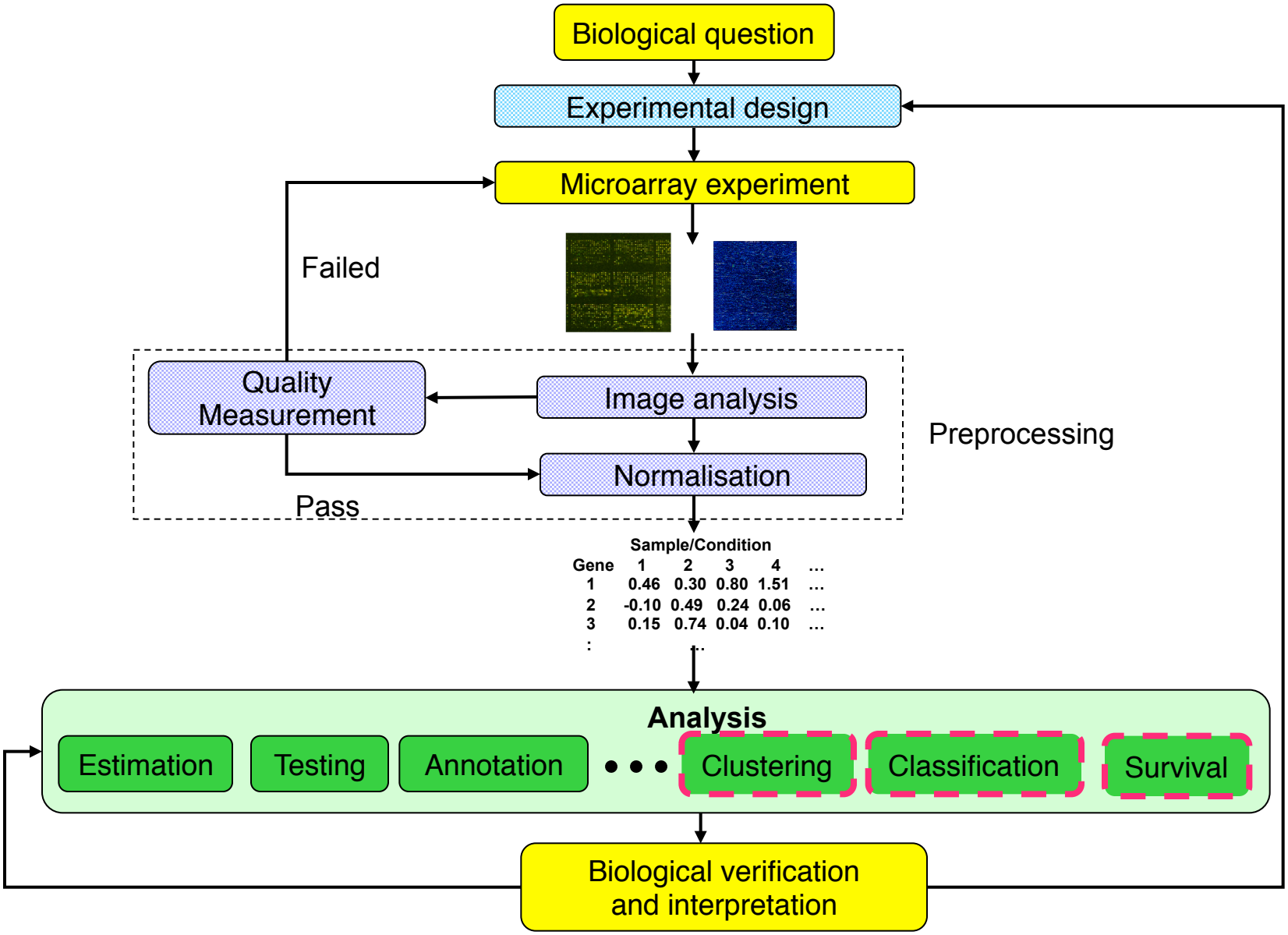
Breast Cancer Functional Genomics Group.

CRUK Cambridge Research Institute (a.k.a. Li Ka Shing Centre)

Oscar.Rueda@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK


Biological question
Experimental design
Microarray experiment
Failed
Quality Measurement
Image analysis
Normalisation
Pass
Preprocessing
Sample/Condition
Gene 1 2 3 4 …
1 0.46 0.30 0.80 1.51 …
2 -0.10 0.49 0.24 0.06 …
3 0.15 0.74 0.04 0.10 …
Analysis
Estimation
Testing
Annotation
Clustering
Classification
Survival
Biological verification and interpretation

# Gene expression as a data matrix

Gene expression data on *p* genes (rows) for *n* samples (columns)

mRNA samples

| Genes | sample1 | sample2 | sample3 | sample4 | sample5 | |
|---|---|---|---|---|---|---|
| 1 | 0.46 | 0.30 | 0.80 | 1.51 | 0.90 | ... |
| 2 | -0.10 | 0.49 | 0.24 | 0.06 | 0.46 | ... |
| 3 | 0.15 | 0.74 | 0.04 | 0.10 | 0.20 | ... |
| 4 | -0.45 | -1.03 | -0.79 | -0.56 | -0.32 | ... |
| 5 | -0.06 | 1.06 | 1.35 | 1.09 | -1.09 | ... |

Gene expression level of gene *i* in mRNA sample *j*

= $\log_2$(Red intensity / Green intensity)

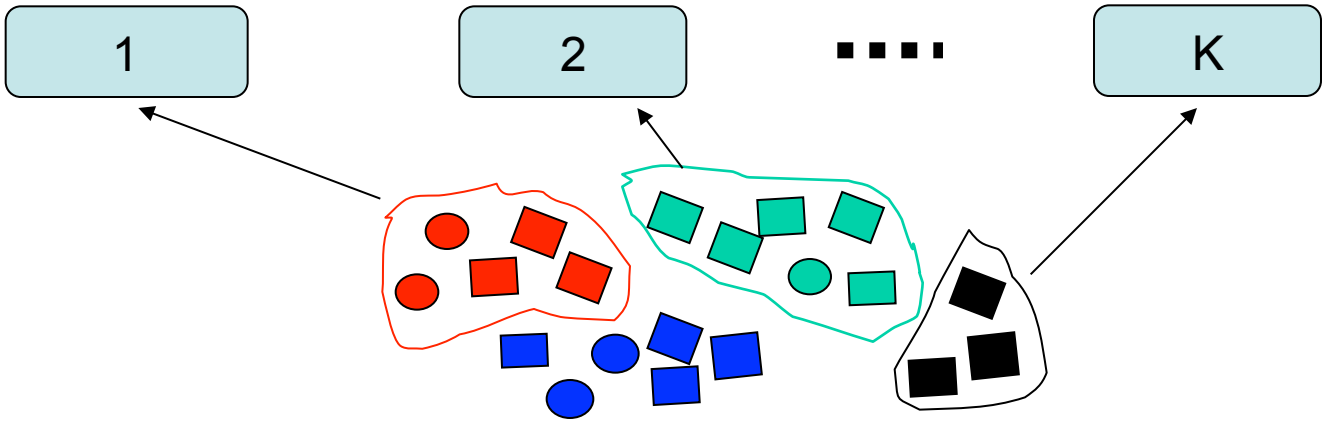
# Classification

# Discrimination - basic principles

• Each object associated with a class label (or **response**) **Y** ∈ {1, 2, …, K} and a feature vector (vector of predictor variables) of G measurements:

$\mathbf{X} = (X_1, \ldots, X_G)$

Aim: predict Y from **X**.




Predefined Class {1,2,…K}

Objects

Y = Class Label

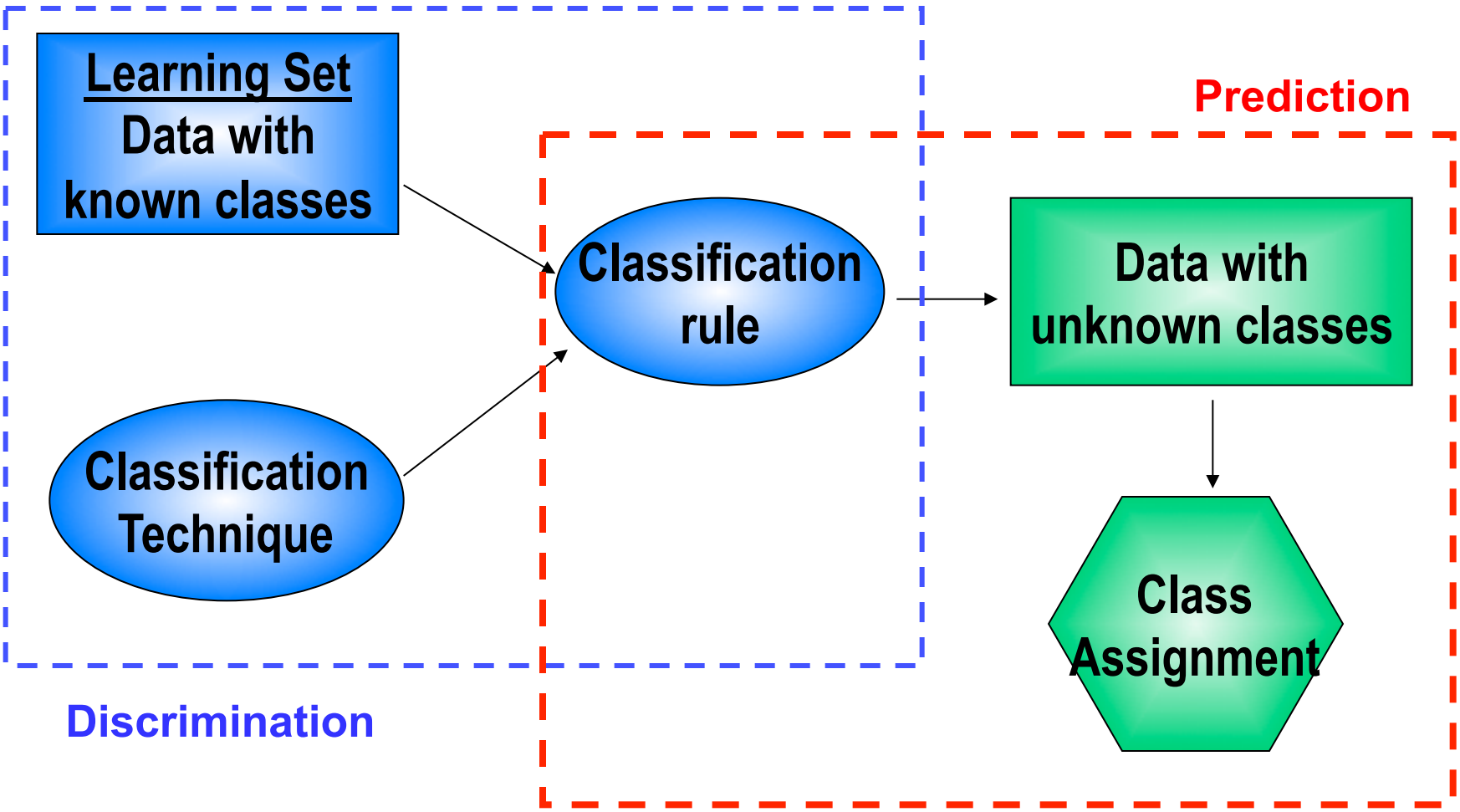
X = Feature vector {colour, shape}

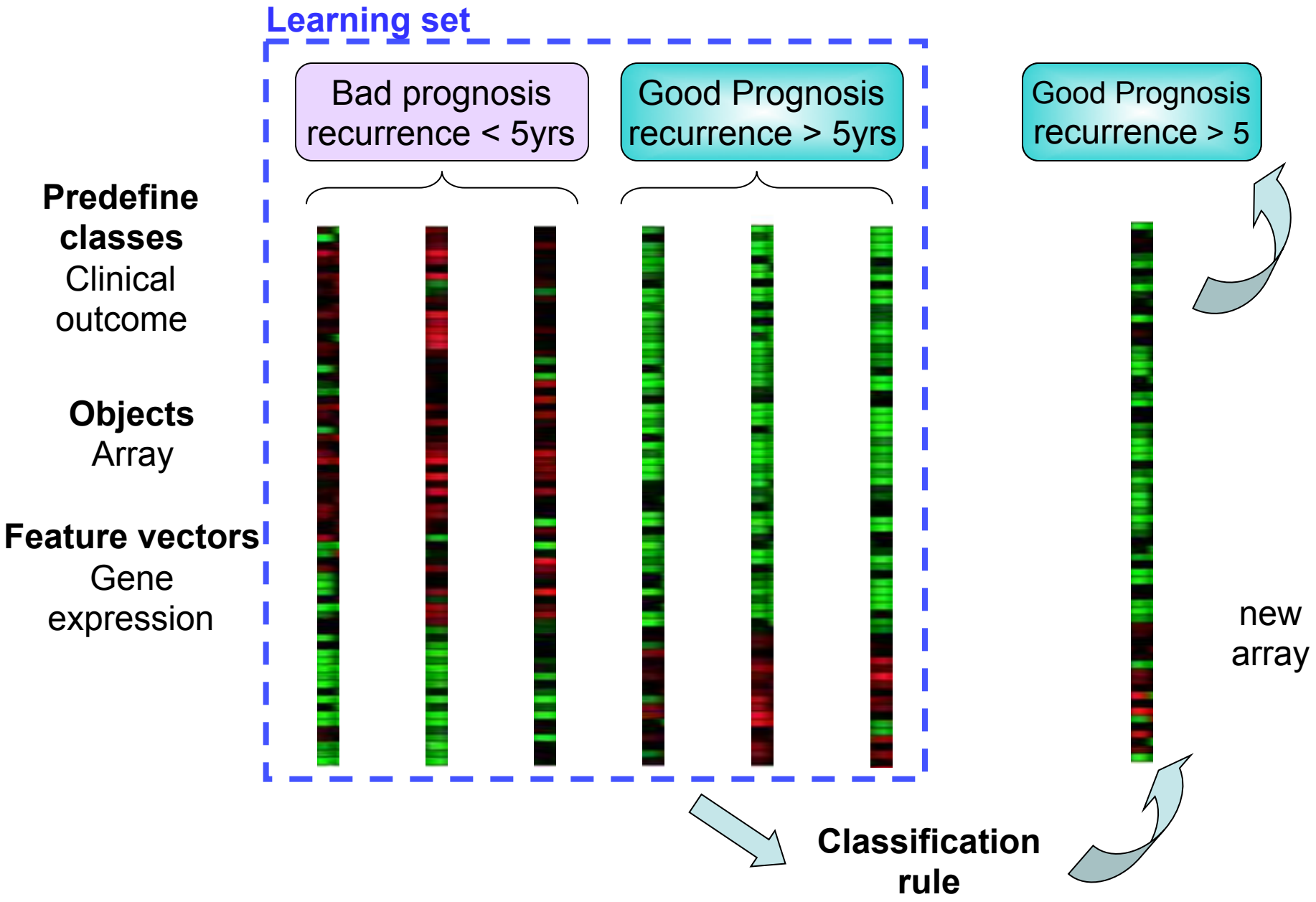
**Classification rule ?**

X = {red, square}

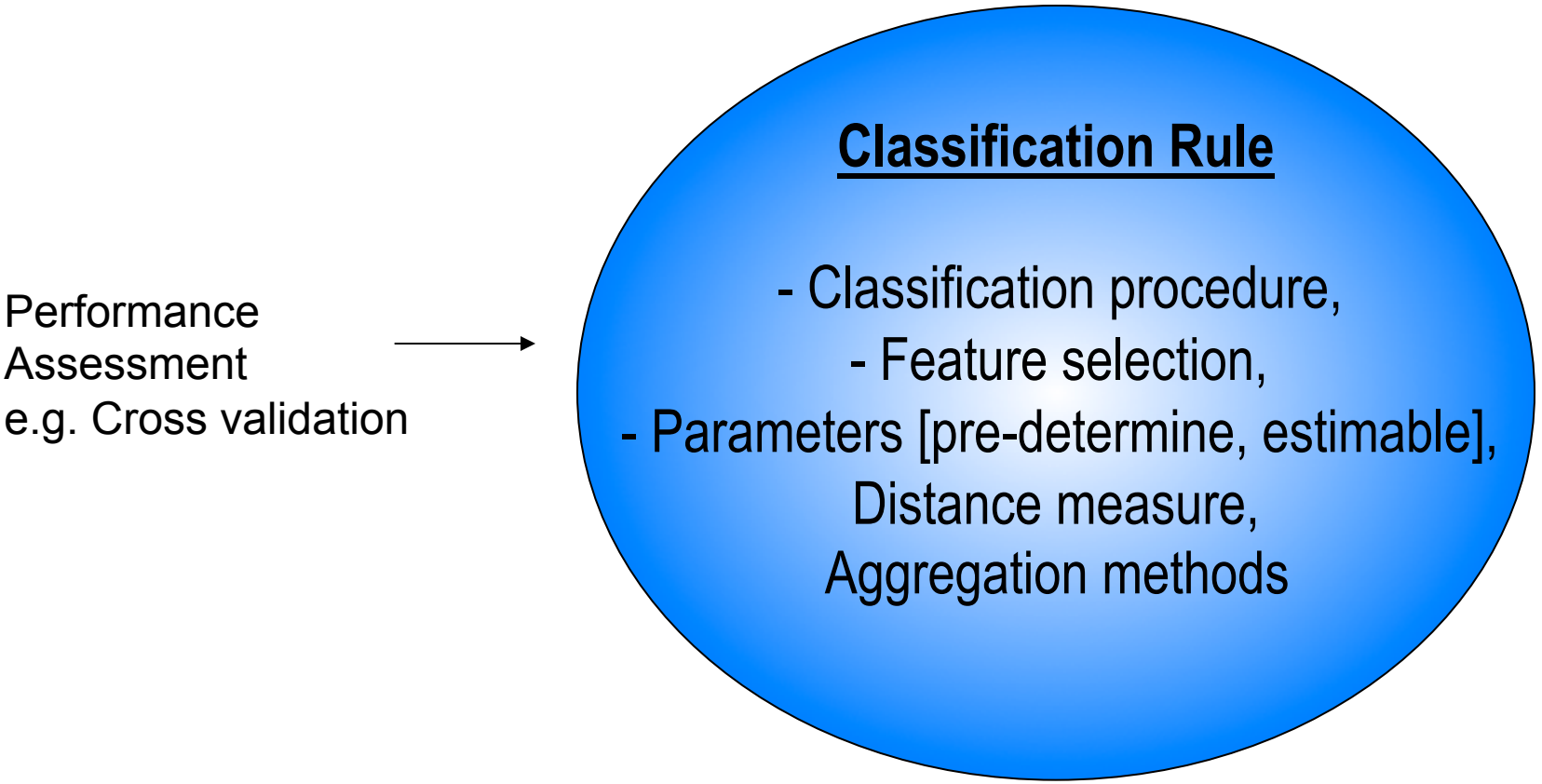
Y = ?

# Classification






van 't Veer LJ, et al. "Gene expression profiling predicts clinical outcome of breast cancer". *Nature*, 2002 Jan 31;415(6871):530-6.




- One can think of the classification rule as a black box, some methods provide more insight into the box.
- Performance assessment needs to be looked at for all classification rules.

# Why feature selection?

- Removing variables that are noise with respect to the outcome leads to better classification performance
- May provide useful insights into etiology of a disease
- Can eventually lead to a diagnostic test (e.g., "breast cancer chip")

# Common methods

Many classifiers available including:

- Linear Discriminant Analysis (LDA).
- Logistic regression.
- Single/multi-layer neural networks.
- Nearest-neighbour methods (k-NN).
- Classification and regression trees (CART).
- Prediction Analysis for microarrays (PAM)
- Many others: Support Vector Machines (SVM), Bayesian networks, logic regression…

# Discriminant Analysis

- Assumption: data follows a **multivariate normal distribution**:

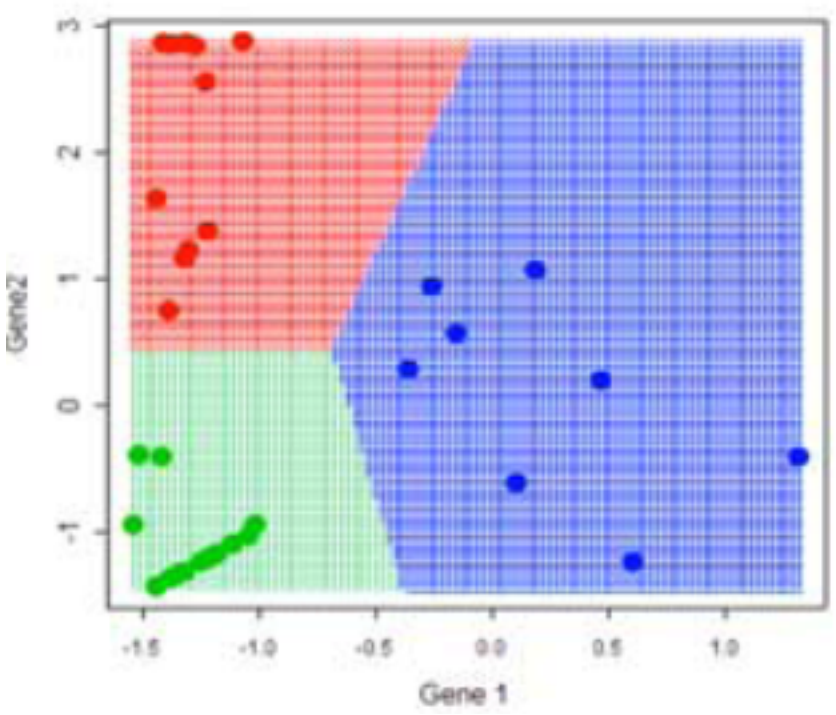
$$(X \mid Y = k) \sim N(\mu_k, \Sigma_k)$$

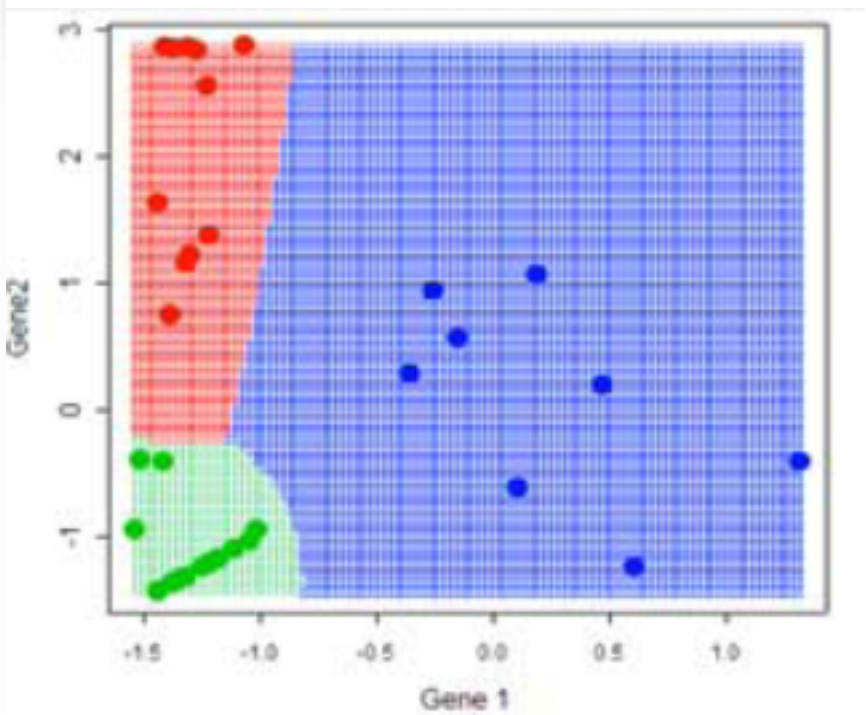
- Classification rule:

$$C(X) = \arg\min_k \left\{ (X - \mu_k)\Sigma_k^{-1}(X - \mu_k)^T + \log|\Sigma_k| \right\}$$

In general, this is a quadratic rule

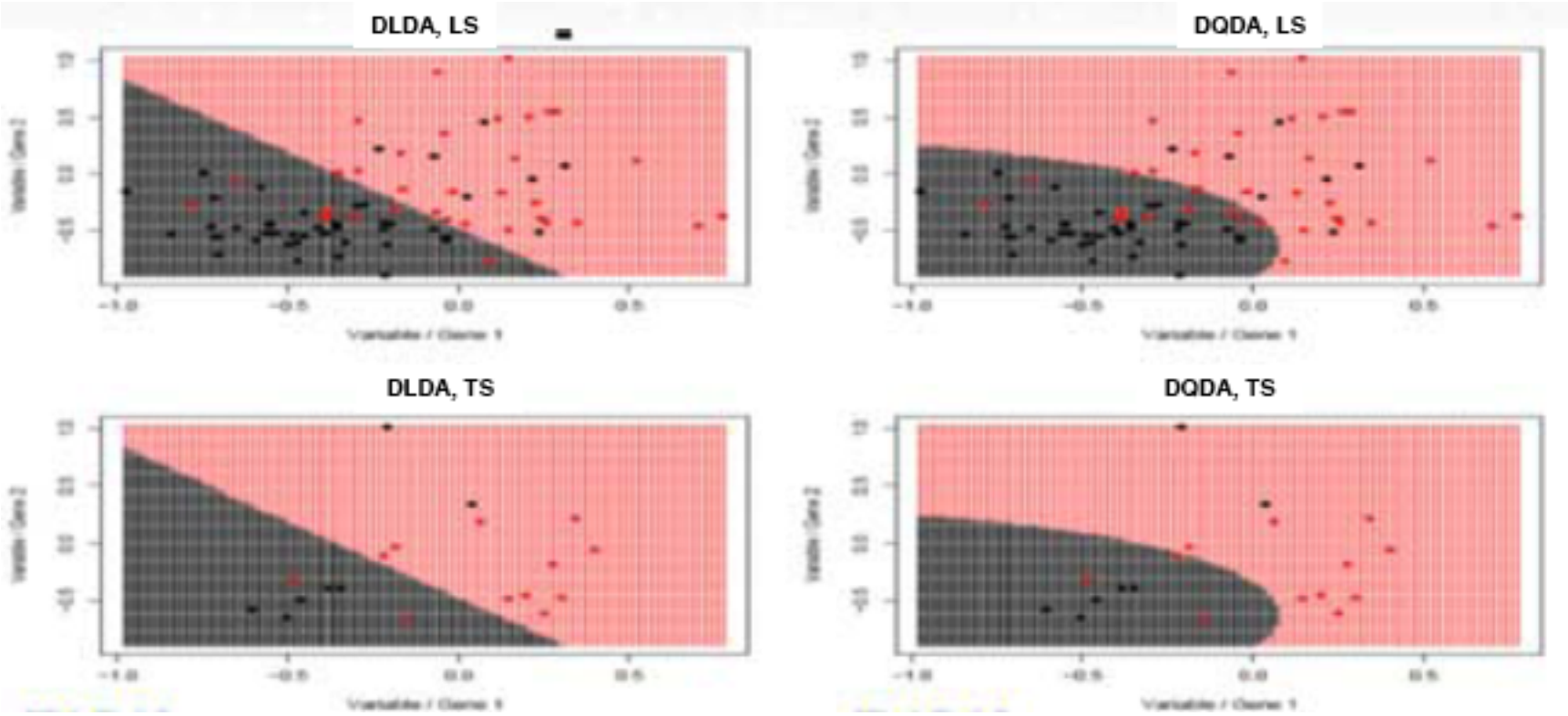
# Discriminant analysis: example(I)




| Linear discriminant analysis (LDA) | Quadratic discriminant analysis (QDA) |
|---|---|
| Same covariance matrix for all groups | Different covariance matrix for all groups |

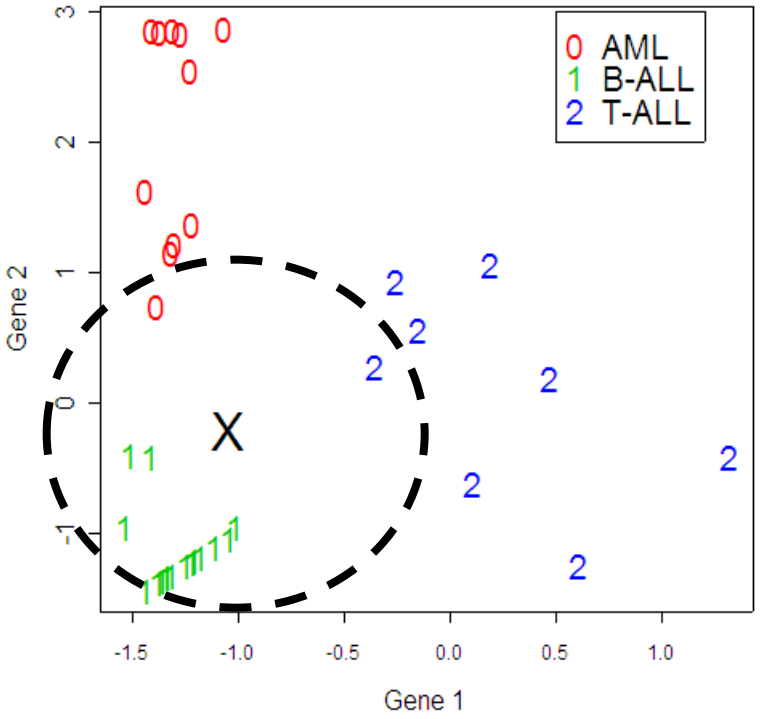
# Discriminant analysis: example(II)




| | |
|---|---|
| Diagonal Linear discriminant analysis (DLDA) | Diagonal Quadratic discriminant analysis (QLDA) |
| Same (diagonal) covariance matrix for all groups | Different (diagonal) covariance matrix for all groups |

# k nearests neighbours

- Based on a measure of distance between observations (e.g. Euclidean distance or one minus correlation).
- k-nearest neighbor rule 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**.

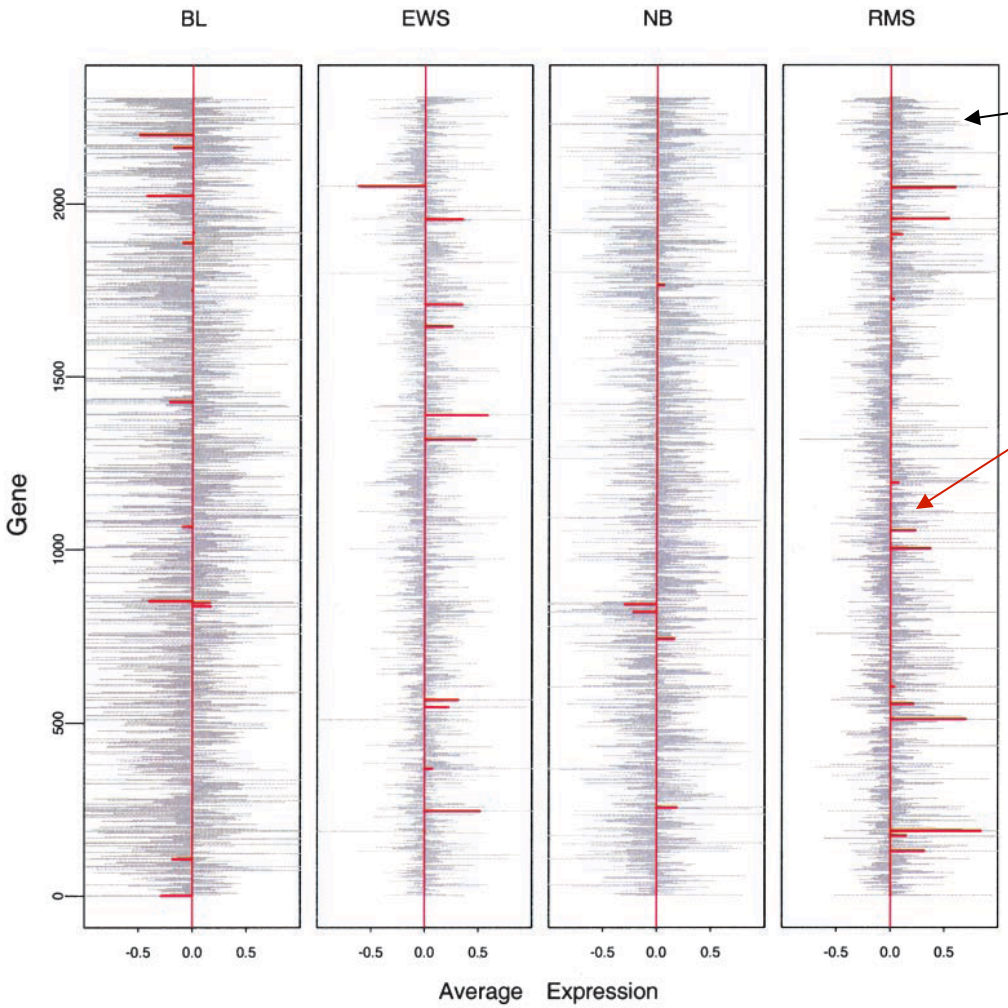


# Classification trees

- A tree is a partition of the feature space.
- We can compare trees with their misclassification rate.
- Interactions between variables and monotonic transformations are handed automatically.
- Trees are "pruned" to avoid overfitting.
- A set of trees can be assembled into **random forests**.

# PAM

- Tibshirani et al., 2002.
- Uses **nearest shrunken centroids**.




Standardized centroid (average expression of the gene divided by the intra-class s.d.)

Shrunken centroid: each centroid is shrunken towards de overall centroid for all classes.

Each sample is assigned to the class whose centroid is closest to it.

It is easy to interpret and incorporates automatic feature selection.

# Common problems

- over-fitting

  *(with enough genes, you can perfectly classify random data)*
- bias

  *(observational/confounding: sample handling, background differences between classes - sex, age)*
- results can be sensitive to tuning parameters, standardization methods, feature selection
- interpretability of classifier (no black box)

  *(how to make sense, biologically)*

Important to assess performance of classifier using independent data set

# 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

# Estimation of error rates (I)

- **Apparent error rate:** misclassification error on the samples of the dataset used to build the classification rule. It is downward biased.
- **Estimation based on a test sample:** the sample is divided (randomly) in two subsets: training sample, to build the classifier, and test sample, to estimate the error rate in classifying those samples. We lose sample size.
- **K-fold cross-validation:** the sample is divided in K groups of roughly the same size. Sequentially, the rule is obtained leaving one set out and the error is estimated on this subset. The error rate is averaged on the K estimates.

# Estimation of error rates (II)

- **Leave-one-out cross-validation:** special case with K=n. Cross-validation methods are computationally intensive.
- **Bootstrap:** A bootstrap sample (sample with replacement from the original dataset) is generated as the training sample. The test sample is formed by the observations not selected for the bootstrap sample. This procedure is repeated P times and the error is averaged.
- **0.632 Estimator:** Efron, 1986.

$$\hat{e} = 0.632\varepsilon_B + 0.368\varepsilon_{App}$$

# Selection bias

- Filter approach: select the genes that are relevant for the prediction (F-ratio, Wilcoxon test, …) and use these genes to build the classifier. Then, estimate the error with cross-validation.
- This approach leads to a **downward bias**: the genes were selected using all samples, including the ones used to test the rule.
- Solution: use **cross-validation on the whole process** (gene selection and prediction).
- We can even add another layer in the cross-validation: selecting the number of genes that leads to lower error rate (finding the best subset among subsets).
- Web application **tnasas**: http://tnasas.bioinfo.cnio.es

# Classification - Summary

- Many methods available.
- No Free Lunch Theorem: No classifier is superior to the others in all scenarios.
- Some methods are black boxes.
- It is crucial to obtain unbiased estimations of the error rate.

# Classification software in R/ Bioconductor

| Package | Function | What |
|---|---|---|
| MASS | lda | linear discriminant analysis |
| | qda | quadratic discriminant analysis |
| class | knn | K-nearest neighbour |
| e1071 | svm | support vector machines |
| rpart | | classification & regression trees |
| tree | | |
| nnet | nnet | neural networks |
| ipred | slda, cv, bagging | |
| pamr | pamr.train, pamr.cv | |
| randomForest | | |
| MLInterfaces | function names as above, add 'B' to end | |
| | xval | cross validation |

# Survival Analysis

# Survival Analysis

- Analysis of **failure times** (events).
- The response variable is **time until the event**.
- Examples of events: death, metastasis, relapse…
- In **microarray studies**, we are usually interested in finding **signatures** (sets) of genes that are related to **prognosis.**

# Censoring

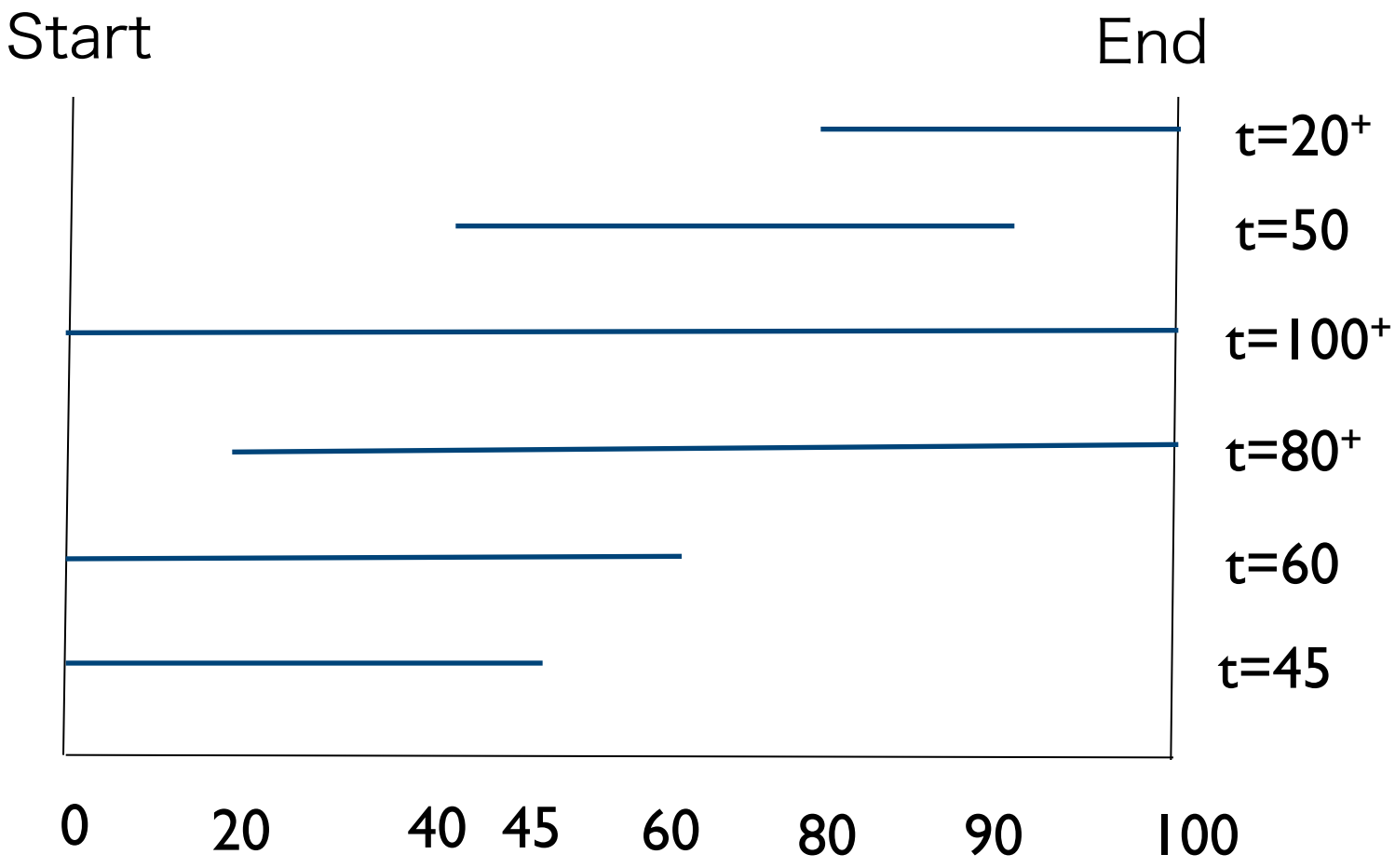
- Times are usually **censored**: we are not able to observe the failure times for all individuals.
- **Interval-censoring**: the event has occurred within an interval of time.
- **Left censoring**: the event has occurred before a certain time.
- **Left truncation**: an unknown number of subjects failed before a certain time, but they never got into the study.

# Right censoring

- **Right censoring**: the event has not occurred up to a certain time.
  - Type I censoring: the study finishes at a pre-specified time (but the censoring can vary between subjects).
  - Type II censoring: the study finishes after a fixed number of events.

- **Assumption**: censoring is **non informative** about the event (for example, patients are not removed from the study because of a worsening condition).

# Type I Censoring



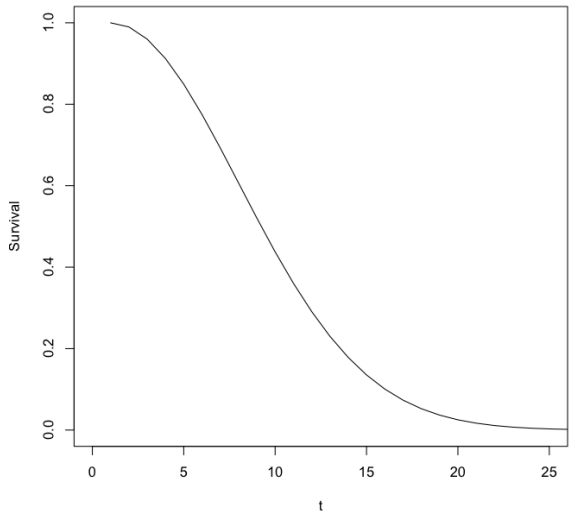
# Functions of interest

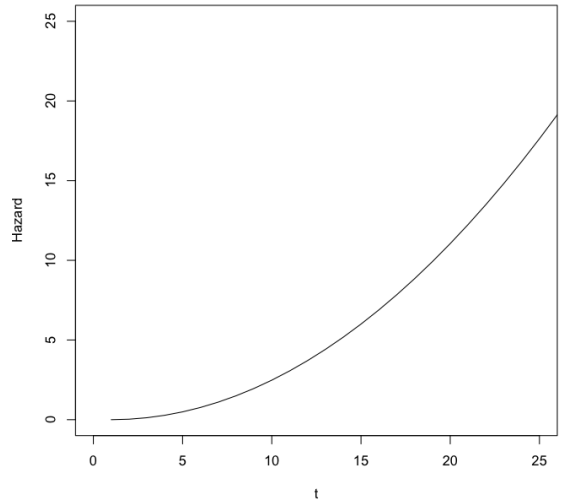
- **Survival function:**

$$S(t) = P(T > t) = 1 - F(t)$$

- **Hazard function:**

$$\lambda(t) = \lim_{u \to 0} \frac{P(t < T \leq t + u \mid T > t)}{u} = \frac{f(t)}{S(t)}$$




Distributions of interest: exponential, Weibull, lognormal…

# Kaplan-Meier Estimator

- **Empirical survival function** when censoring is present.

$$S_{KM}(t) = \prod_{i:t_i<t} (1 - d_i / n_i)$$

$d_i$ is the number of failures at $t_i$

$n_i$ is the number of subjects at risk at $t_i$

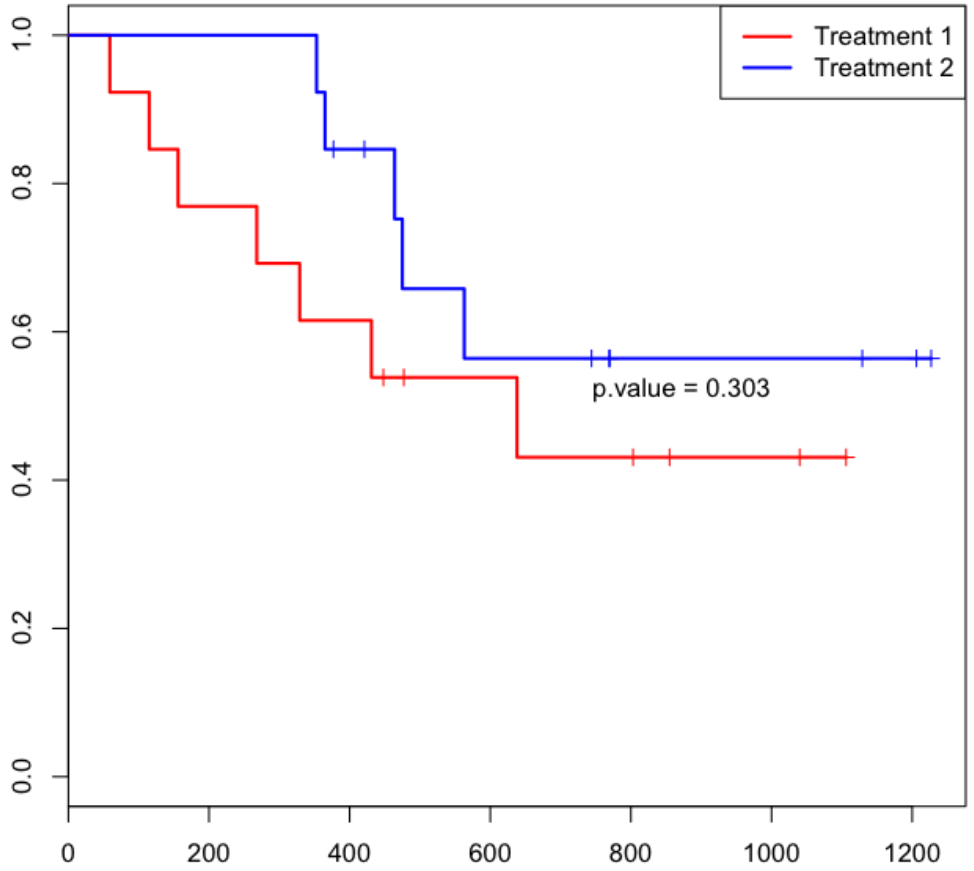
| Day | Subjects at risk | Deaths | Censored | Cumulative Survival |
|---|---|---|---|---|
| 12 | 100 | 1 | 0 | 99/100=0.99 |
| 30 | 99 | 2 | 1 | 97/99 x 0.99=0.97 |
| 60 | 96 | 0 | 3 | 96/96 x 0.97 = 0.97 |
| 72 | 93 | 3 | 0 | |

Table 16.1, Harrell.

# Log rank test

- Tests **differences between the survival** functions for two or more groups.

Compares observed and expected events in each group



# Cox model

- **Semiparametric proportional hazards model.**

$$\lambda(t \mid X) = \lambda(t)\exp(X\beta)$$

• Uses a partial likelihood to estimate $\beta$

• No assumptions about the shape of the underlying hazard, but the relative hazard function must be constant through time. The predictors have the same effect on the hazard function at all values of t.

• The model can be extended to include strata and time-dependent covariates.

# R functions and packages

- Package survival:

| | |
|---|---|
| Surv(time,status) | Define survival (time, censoring) |
| survfit() | Kaplan-Meier estimator |
| survdiff() | Log-rank test |
| coxph() | Cox model |

- Package Design *(Harrell)*
- Signs web application (http://signs.bioinfo.cnio.es/))

# Acknowledgements

**UCSF**

- Ru-Fang Yeh
- Agnes Paquet
- David Erle
- Andrea Barczac

**UCB**

- Terry Speed
- Jane Fridlyand
- Sandrine Dudoit

**University of Sydney**

- Jean Yang

**U. Cambridge**

- Stephen Eglen

**UCL**

- Andrew Teschendorff

# References (I)


- Chapter 12,13,16 - Bioinformatics and Computational Biology Solutions using R and Bioconductor
- Dov Stekel (Microarray bioinformatics), chapter 8
- Chapters 3,4 - Speed book
- **Frank Harrell, Regression model strategies**
- Chapter 10,12 - Data Analysis and Visualization in Genomics and Proteomics
- **Venables and Ripley, Modern Applied Statistics with S**
- Breiman, Classification and regression trees, 1984
- SVD and PCA for microarrays - http://public.lanl.gov/mewall/kluwer2002.html
- Eisen et al. *PNAS*, 95(25):14863-8, 1998
- Tamayo et al. *PNAS*, 96(6):2907-12, 1999
- Tavazoie et al. *Nature Genetics*, 22(3):281-5, 1999
- Alizadeh et al. *Nature*, 403(6769):503-11, 2000
- Cho et al. *Mol Cell*, 2(1):65-73, 1998
- Golub et al, *Science*, 286(5439):531-7, 1999
- van' t Veer et al, *Nature*, 415(6871):530-6, 2002
- van de Vijver et al, *N Engl J Med*, 347(25):1999-2009, 2002
- http://videolectures.net/bootcamp07_guyon_ifs/ - lecture on feature selection
- Ambroise and McLachlan, *PNAS*, 99(10):6562-6566, 2002
- http://www.clopinet.com/isabelle/Projects/ETH/Questions_lecture_8.html

# References (II)


- Amy V. Kapp and Robert Tibshirani, 2007. ***Are clusters found in one dataset present in another dataset?.*** Biostatistics 8(1), 9-31
- Ronglai Shen, Adam B. Olshen and Marc Ladanyi, 2009.***Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.*** Bioinformatics 25(22), 2906–2912
- van Wieringen et al., ***Survival prediction using gene expression data: A review and comparison***. Computational Statistics & Data Analysis, 2009: 53(5),1590-1603
- Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. ***Prediction by supervised principal components***. Journal of the American Statistical Association 101, 119-137.
- Park, M.Y., Hastie, T., 2006. L1 ***Regularization Path Algorithm for GeneralizedLinear Models***. Technical Report, Stanford University.
- Pawitan, Y., Bjohle, J., Wedren, S., Humphreys, K., Skoog, L., Huang, F., Amler, L., Shaw, P., Hall, P., Bergh, J., 2004. ***Gene expression profiling for prognosis using Cox regression***. Statistics in Medicine 23, 1767-1780.
- Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan MJ. (2006a). ***Survival Ensembles.*** Biostatistics 2006,7:355-373.
- Dave, SS, Wrigth, G, Tan B, Rosenwald A, Gascoyne RD, et al. (2004). ***Prediction of survival in follicular lymphoma based on molecualr features of tumor-inflitrating immune cells***. New England Journal of Medicine 351: 2159-219.
- Efron, B. 1983. ***Estimating the Error Rate of a Prediction Rule: Improvement on cross-validation.*** Journal of the American Statistical Association, 78 (382).
- Tibshirani, Hastie, Narasimhan and Chu, (2002) ***Diagnosis of multiple cancer types by shrunken centroids of gene.*** PNAS 99:6567-6572.