UNIVERSITY OF CAMBRIDGE

MPhil in Computational Biology 2016/17
Course: **Functional Genomics**

Clustering and Survival analysis

Maurizio Callari

maurizio.callari@cruk.cam.ac.uk

(Contributions by Oscar Rueda, Matt Ritchie, Christina Curtis, Jean Yang and Stephen Eglen).

Gene expression as a data matrix

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

samples

Genes or features	sample1	sample2	sample3	sample4	sample5	
1	5.46	4.30	7.80	6.51	5.90	...
2	8.10	10.49	9.24	13.06	12.46	...
3	7.15	7.74	8.04	9.10	8.20	...
4	6.45	11.03	10.79	8.56	9.32	...
5	12.06	9.06	11.35	10.09	12.09	...

Downstream analyses

The analyses to answer your biological question will fall in one of three major categories:

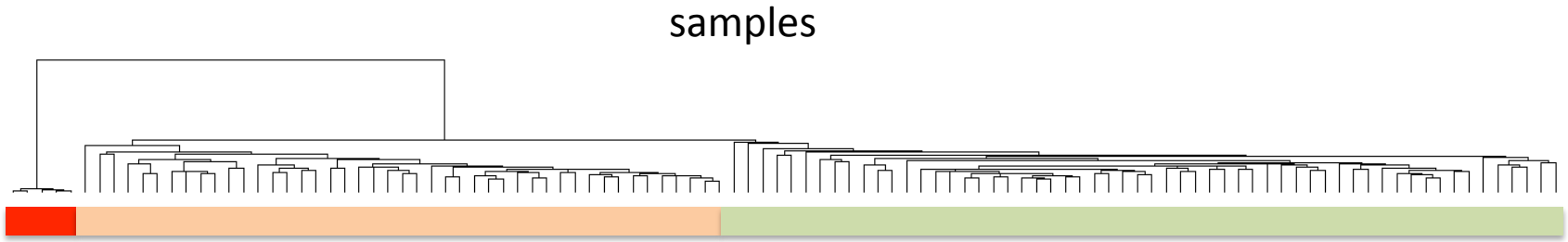
- **Class discovery**
  - When classes are unknown (e.g. discovery of subtypes in a tumour type)
  - Approaches: hierarchical clustering, K-means clustering…
- **Class comparison**
  - When the groups or classes are known and we want to identify genes (or pathways) associated with them (e.g. treated/untreated cells)
  - Approaches: t-test, linear models, pathway analysis…
- **Class prediction**
  - When we want to identify a set of genes able to accurately predict the classes of interest in independent data
  - Approaches: PAM, SVM…
  - Survival analysis…

Clustering analysis

- Leads to readily interpretable figures
- Can be helpful for identifying patterns in the data

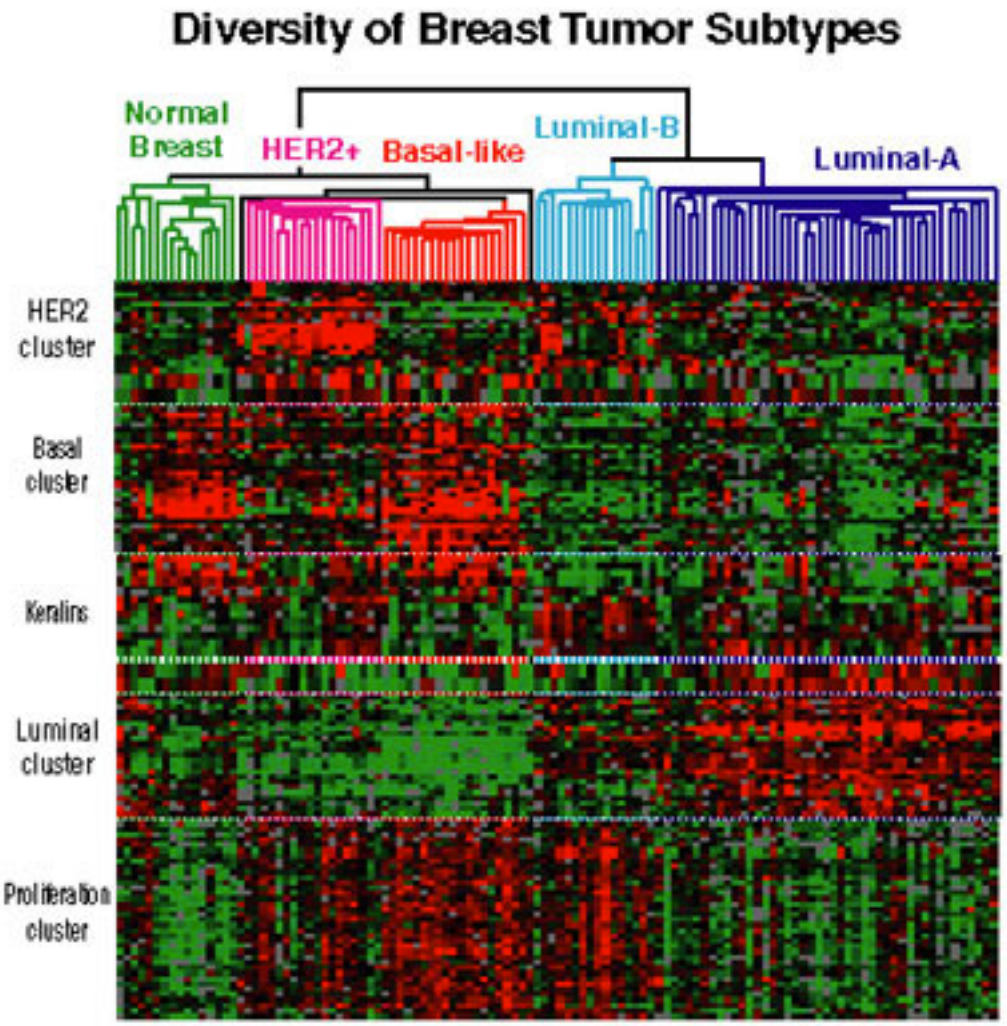
- **Clustering of samples**
  - Quality control purposes
  - Identify new classes
  - Display purposes
- **Clustering of genes**
  - Co-expressed genes => functionally related genes
  - Spatial or temporal expression patterns
  - To reduce redundancy in prediction (feature selection)
  - Display purposes

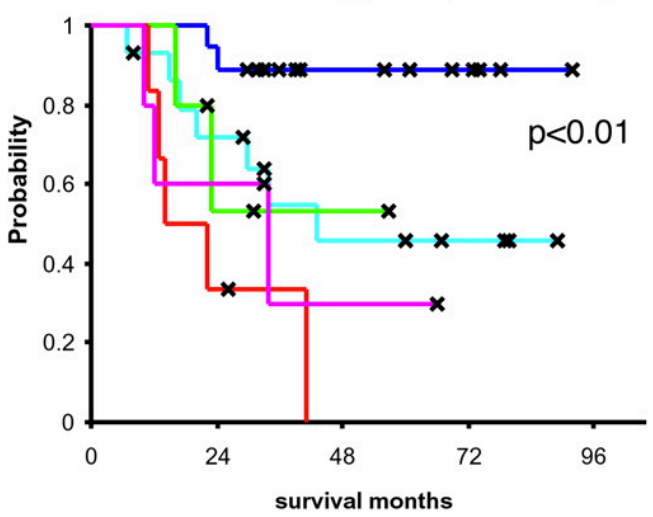
Clustering samples (1)



- miRNA expression profile
- Clusters associated with batch of reagent used
- Importance of tracking all variables able to generate batch effects

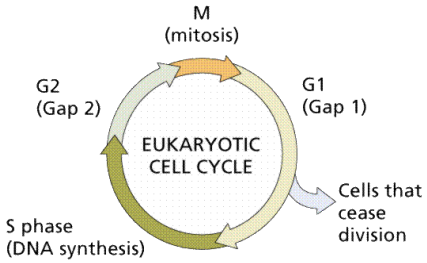
Clustering samples (2)



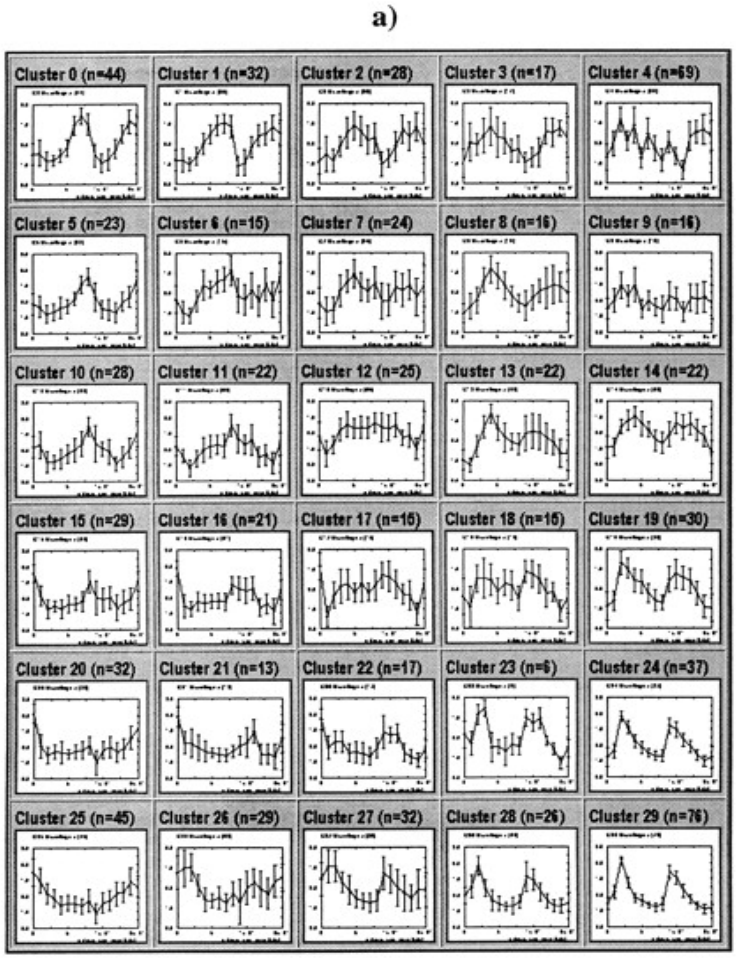


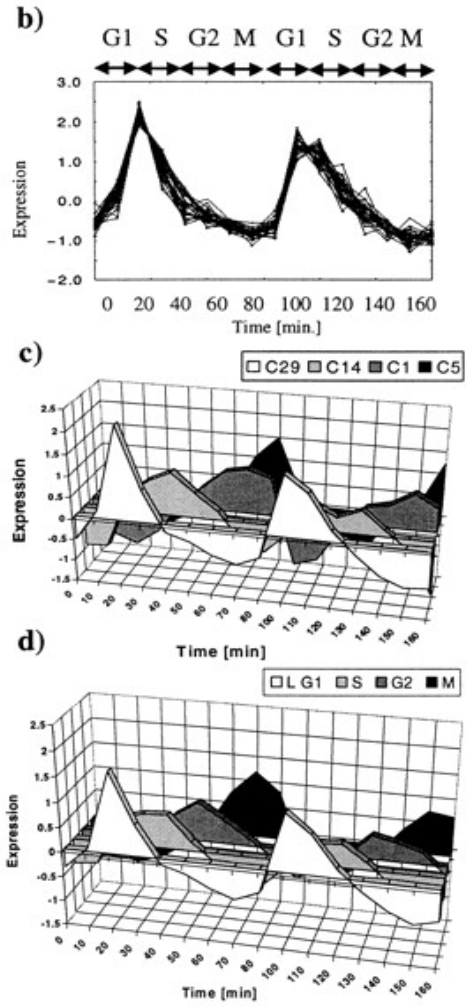
Perou CM et al. Nature. 2000

Clustering genes



Finding different patterns in the data:
Yeast Cell Cycle
6 × 5 SOM
(828 genes)





Tamayo P, et al. *PNAS*, 1999

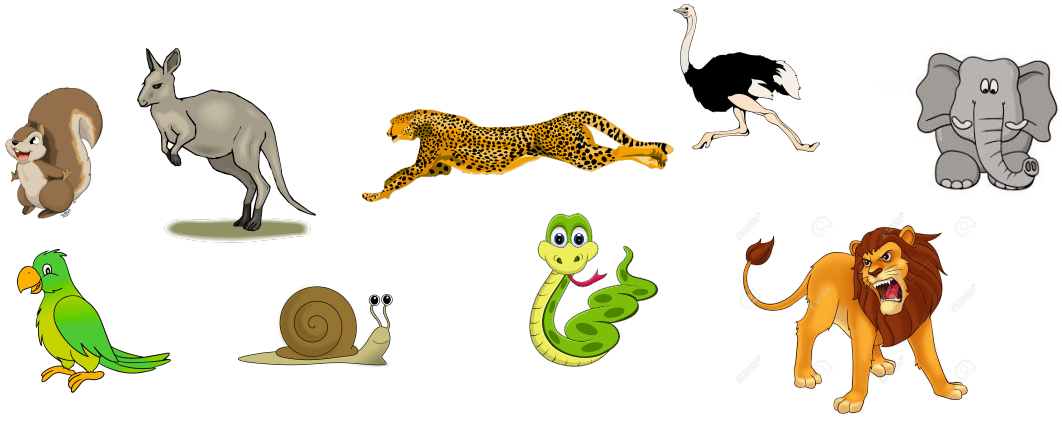
Steps in a Cluster Analysis

1. Preprocess the data.
2. Choose a dissimilarity measure.
3. Choose a cluster algorithm.
4. Select the number of clusters.
5. Validate the procedure.

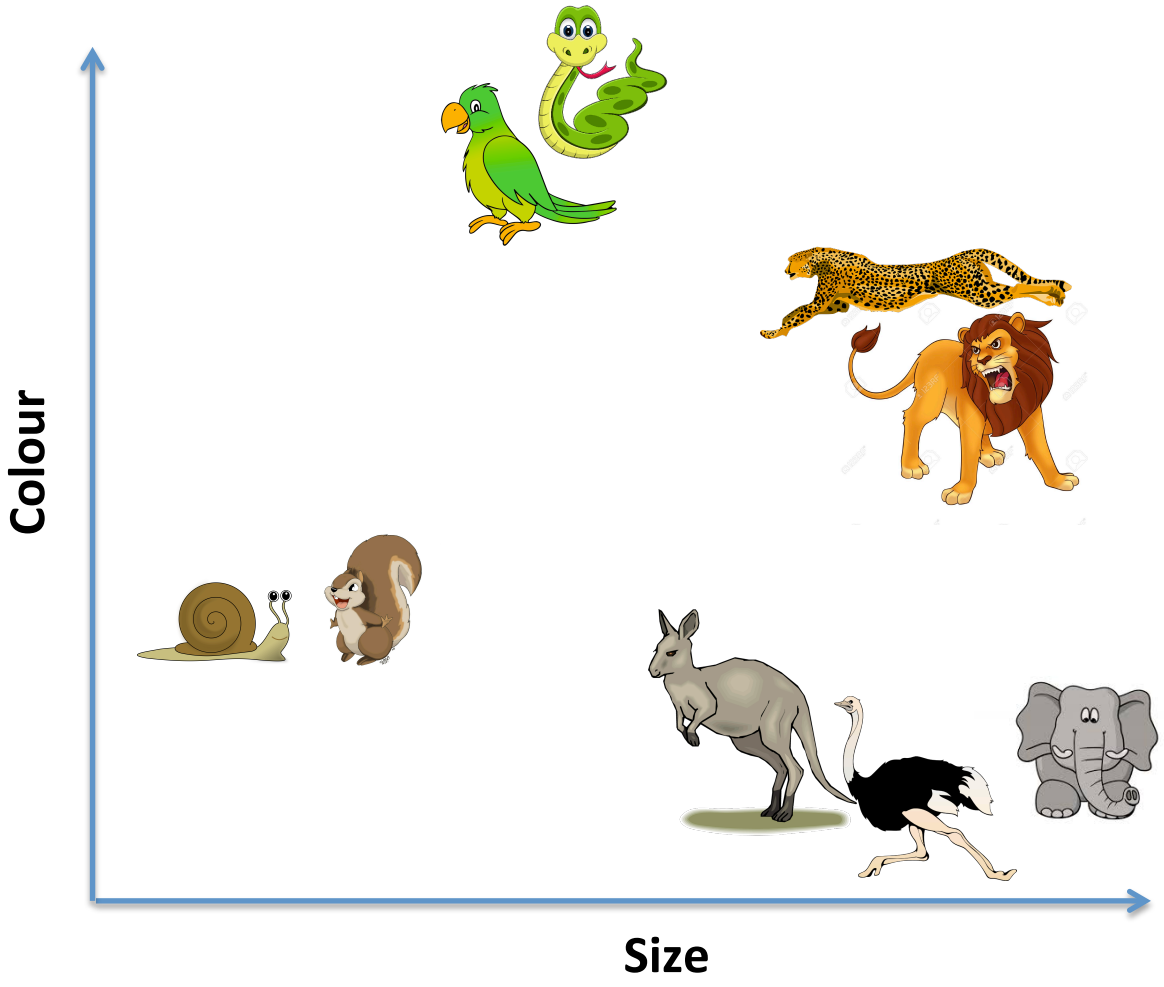
Pre-processing

- Filter
  - Remove genes with low variability across samples
  - In gene clustering only: you might select differentially expressed genes before clustering
  - DO NOT cluster the samples based on the expression levels of differentially expressed genes. Is it remarkable if the samples then cluster into the two groups?

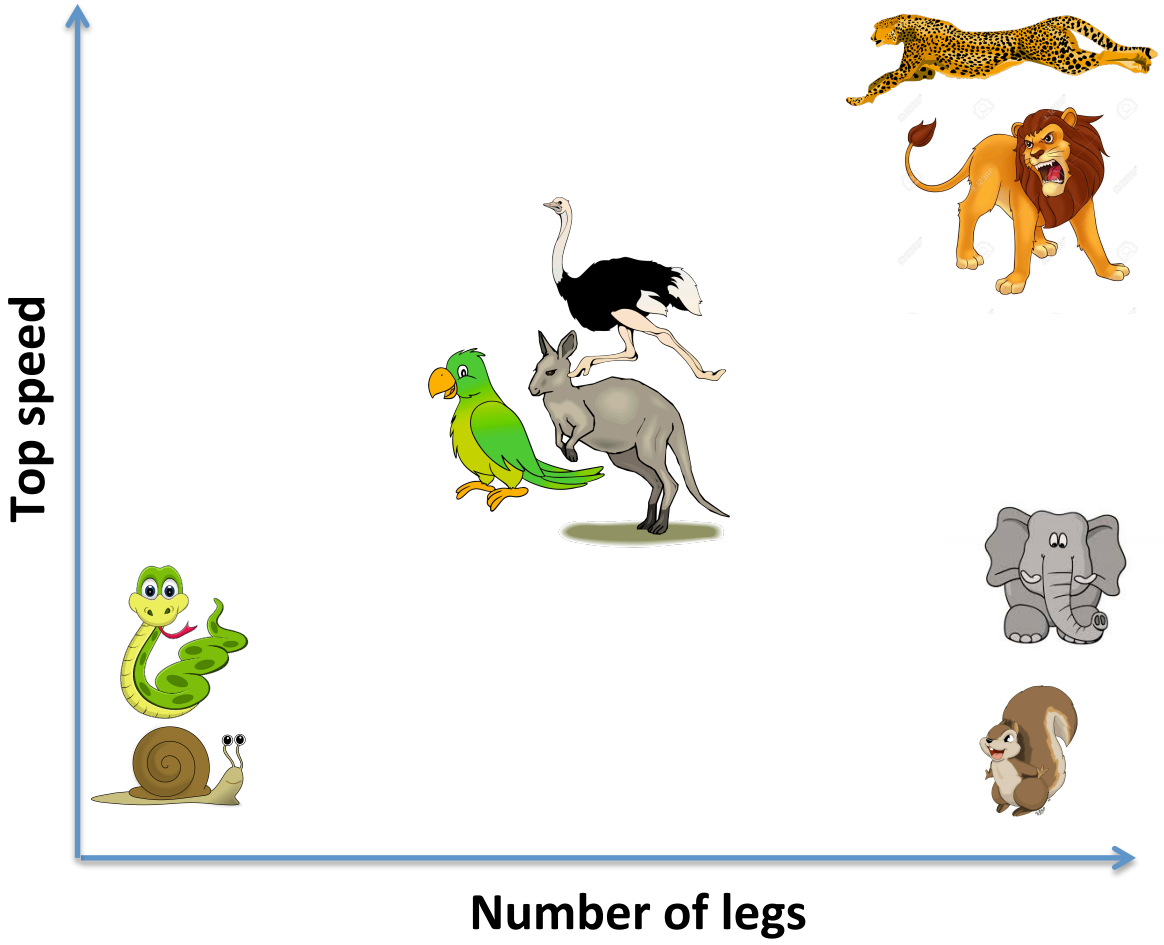
- Clustering depends on which features you select…



Effect of feature selection



Effect of feature selection



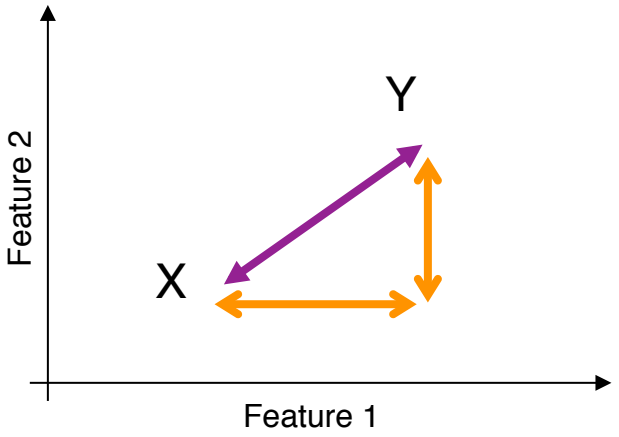
Similarity measures

The feature data are often transformed to an *n x n* distance or similarity matrix, D=(d_{ij}), for the *n* objects to be clustered

- **Correlation coefficient**: *scale invariant*
  - Pearson's correlation: $$d_C(x,y) = 1 - \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}}.$$
  - Spearman ρ: Pearson's correlation of ranks
- **Distance**: *scale dependent*
  - **Euclidean** distance: $$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$
  - City block (**Manhattan**) distance: $$d_M(x,y) = \sum_{i=1}^{n}|x_i - y_i|.$$
- Many others

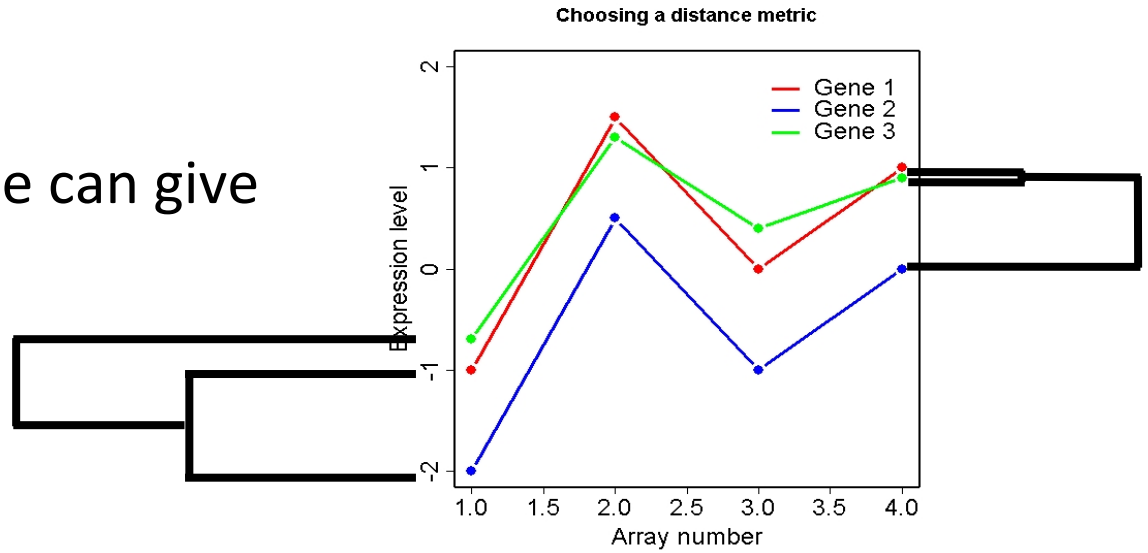
Similarity measures

- Euclidean and Manhattan distance both measure absolute differences between vectors



- Manhattan distance is more robust against outliers

- Correlation and Distance can give very different results



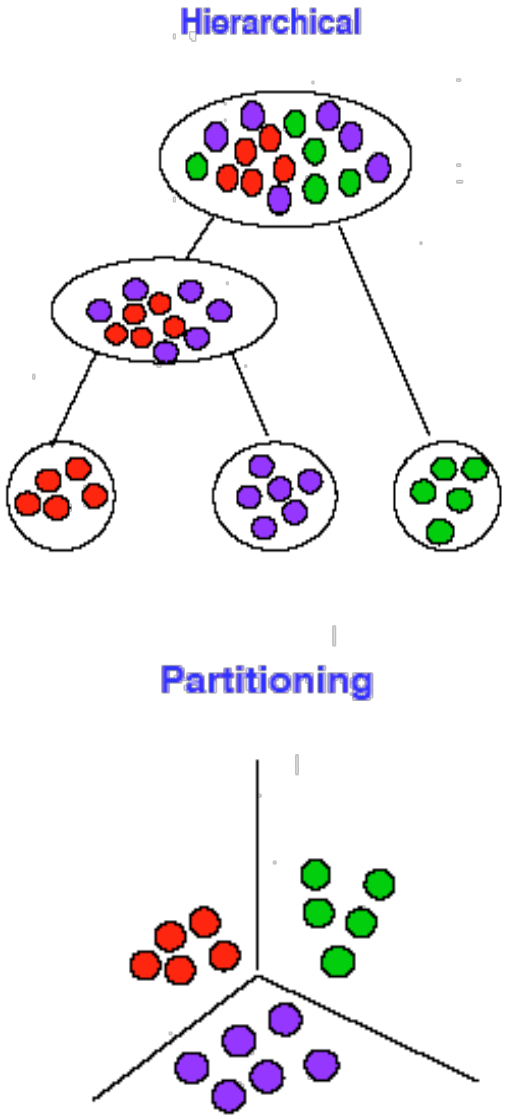
Clustering algorithms

- **Hierarchical (agglomerative) methods**: provide a hierarchy of clusters
  - Hierarchical clustering

- **Partitioning methods**: require the specification of the number of clusters
  - K-means
  - Partitioning around medoids (PAM)
  - Self-organizing maps (SOM)

- **Model-based clustering**
  - e.g. Gaussian mixtures

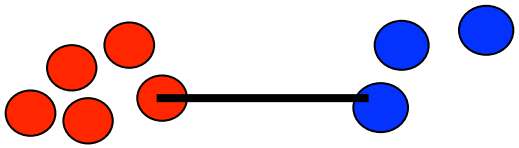
- **Biclustering**



Hierarchical clustering

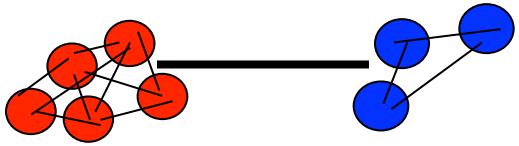
- Start with **n** samples (or **p** gene) clusters
- At each step, merge the two closest clusters using a measure of between-cluster dissimilarity
- The distance between clusters is defined by the method used (**linkage**)
- Similarity of objects is represented in a tree structure (**dendrogram**)

Linkage

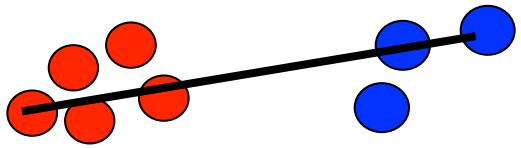


Single
(min. of pairwise distances)
Elongated clusters;
Sensitive to outliers

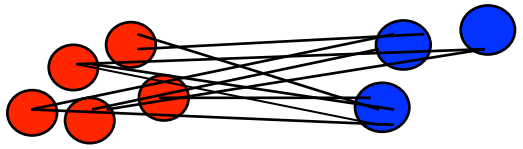
Distance between centroids



Ward method
(min. increase in within-clusters variance)

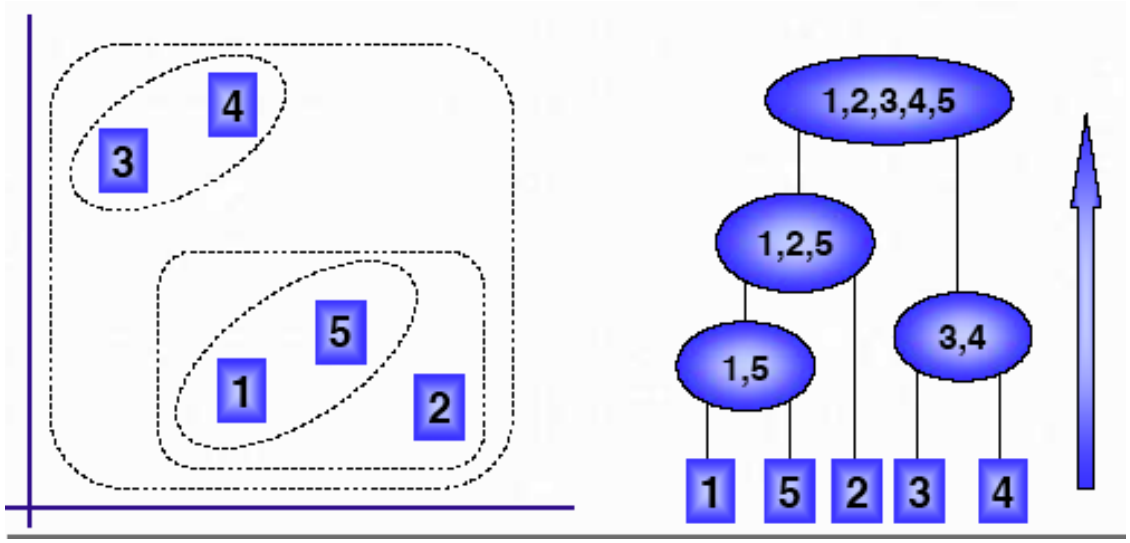


Complete
(max. of pairwise distances)
Compact clusters;
Sensitive to outliers



Average linkage
(mean of all pairwise distances)

Dendrograms



Dendrograms are good visual guides but arbitrary!
Nodes can be reordered. Closer on dendogram ≠ more similar.

K-means

Ask user how many clusters they'd like.

(e.g. k=2)

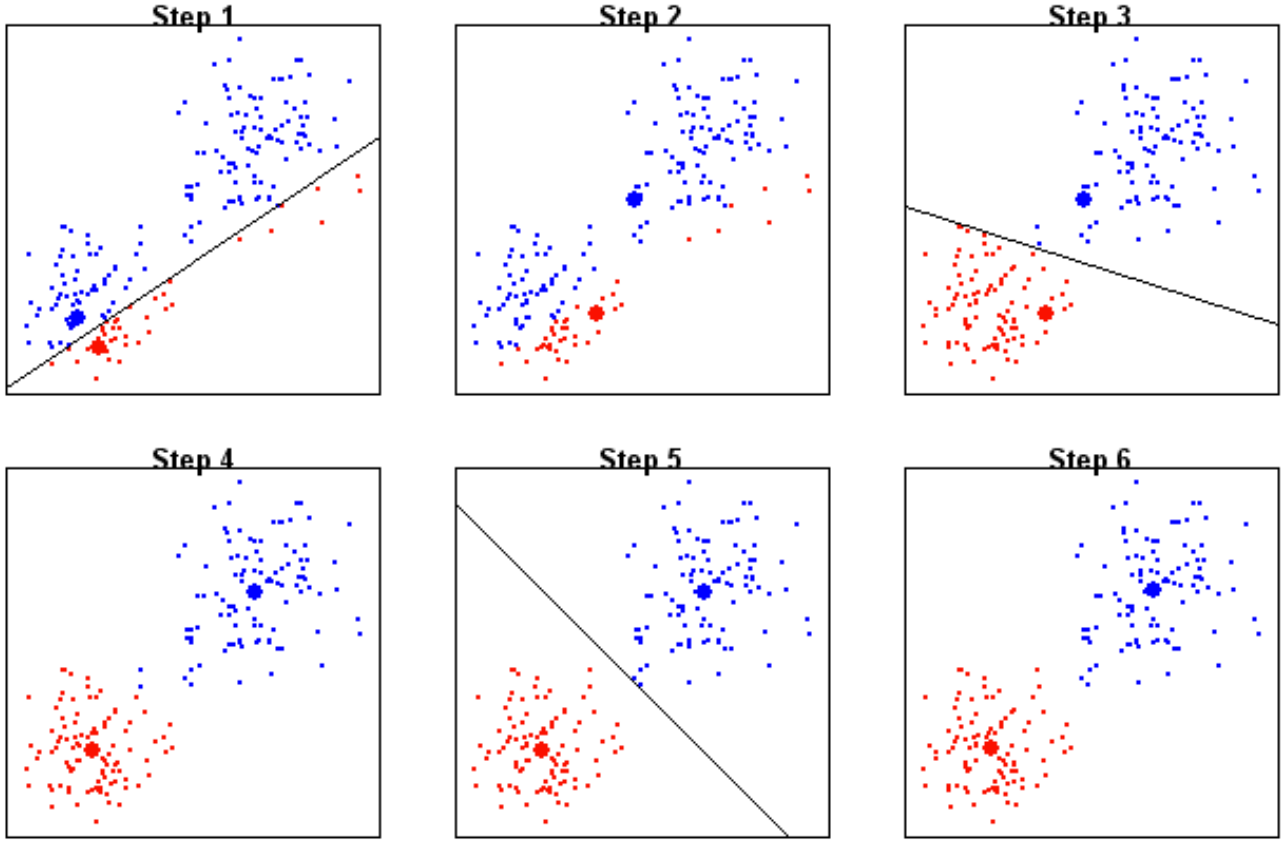
Randomly guess k cluster Center locations

Each datapoint finds out which Center it's closest to.

Each Center finds the centroid of the points it owns...

...and jumps there

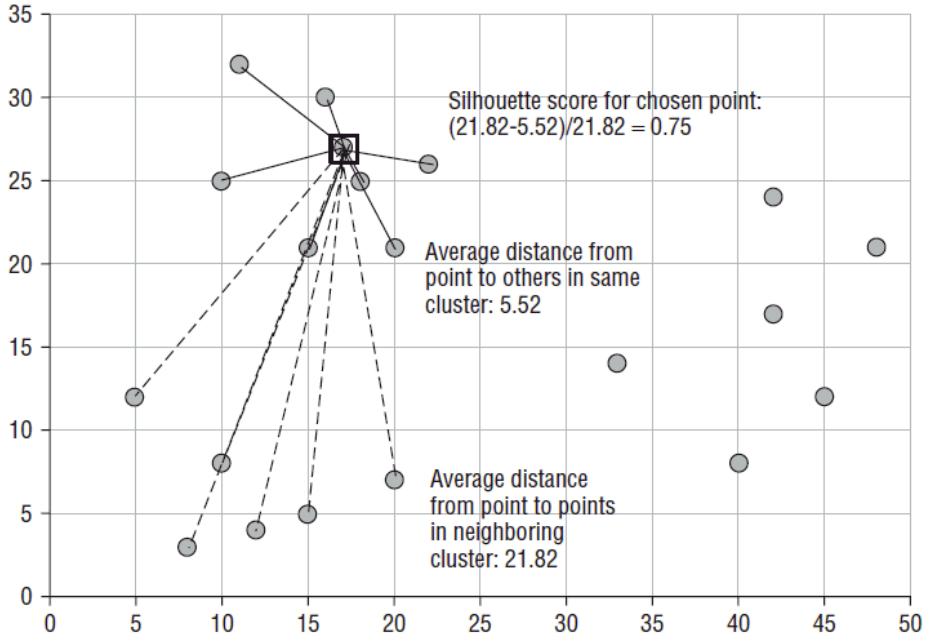
...Repeat until terminated!



http://sherrytowers.com/

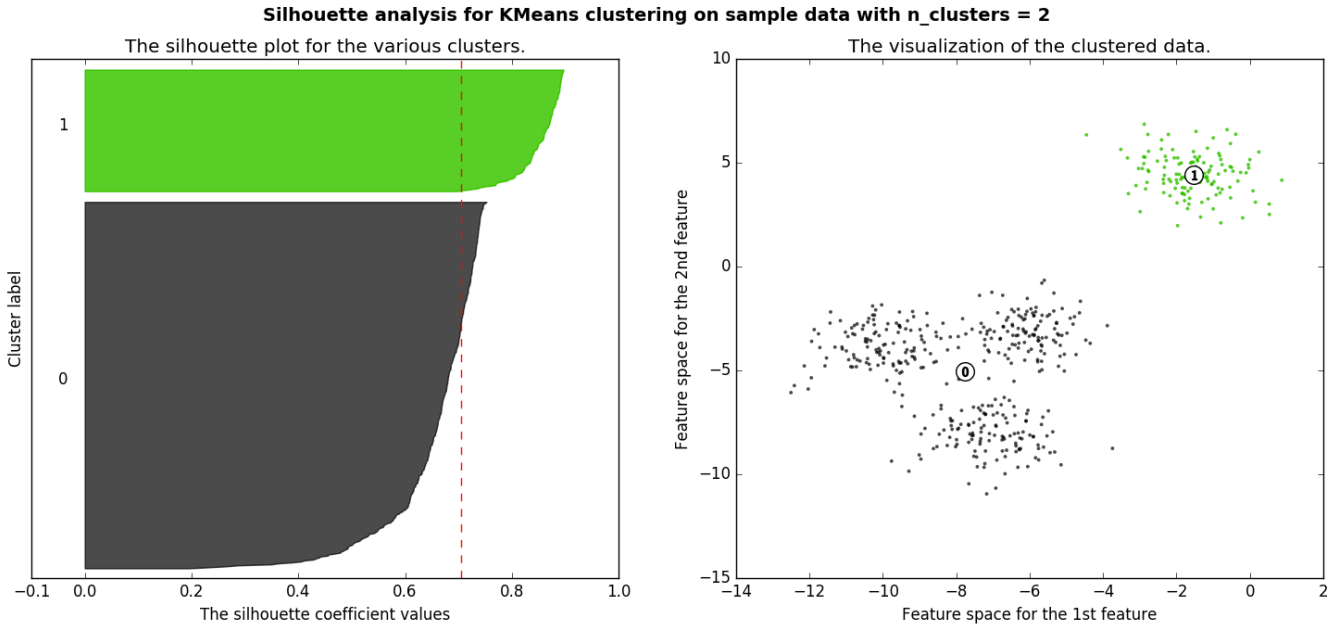
K-means

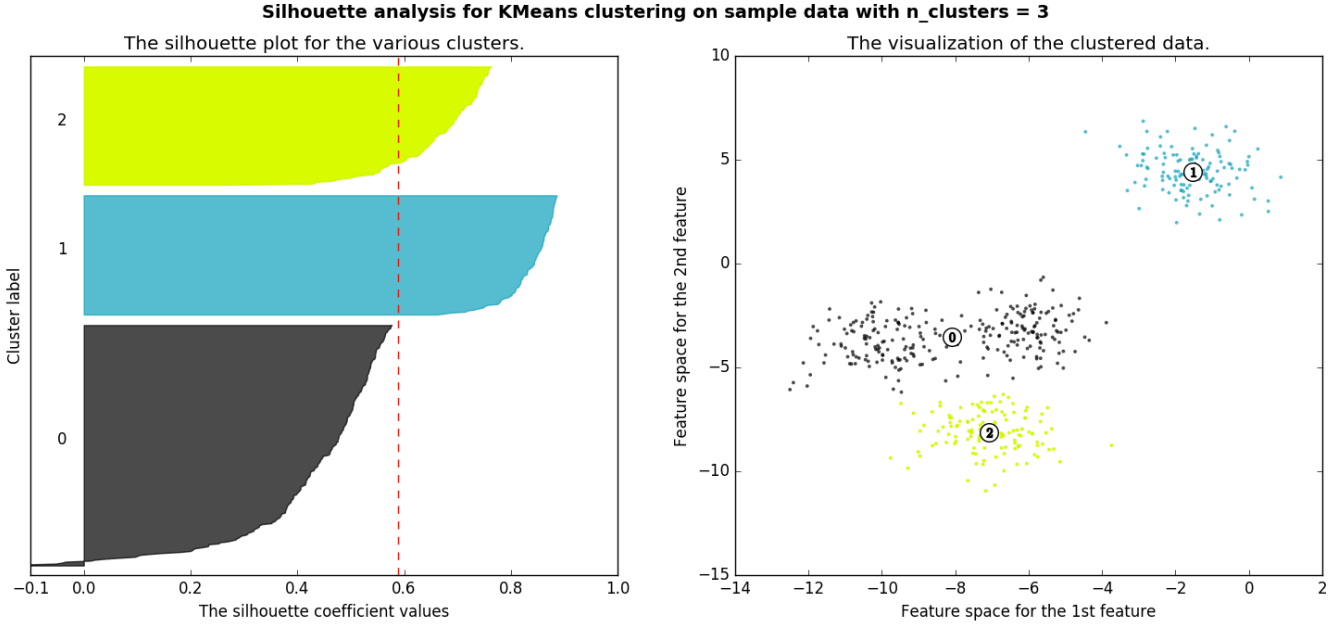
- How many clusters?
- No easy answer
- Compare the quality of clustering results for different values of *K* (e.g. Dudoit et al. Genome Biol 2002)
- Silhouette analysis (Rousseeuw PJ. J Comput Appl Math. 1987) can be used to:
  - select the number of clusters
  - assess how well individual observations are clustered



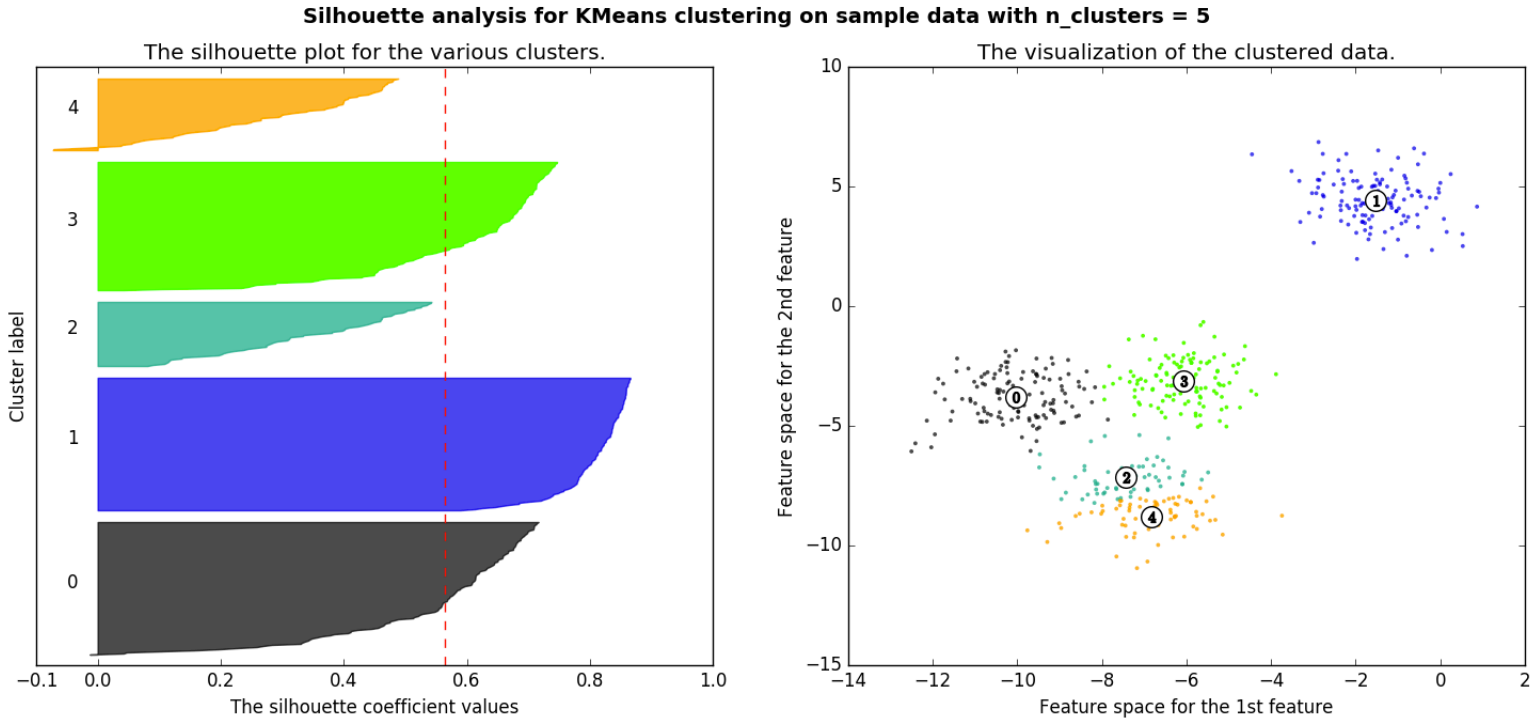
http://blog.data-miners.com/

Silhouette analysis





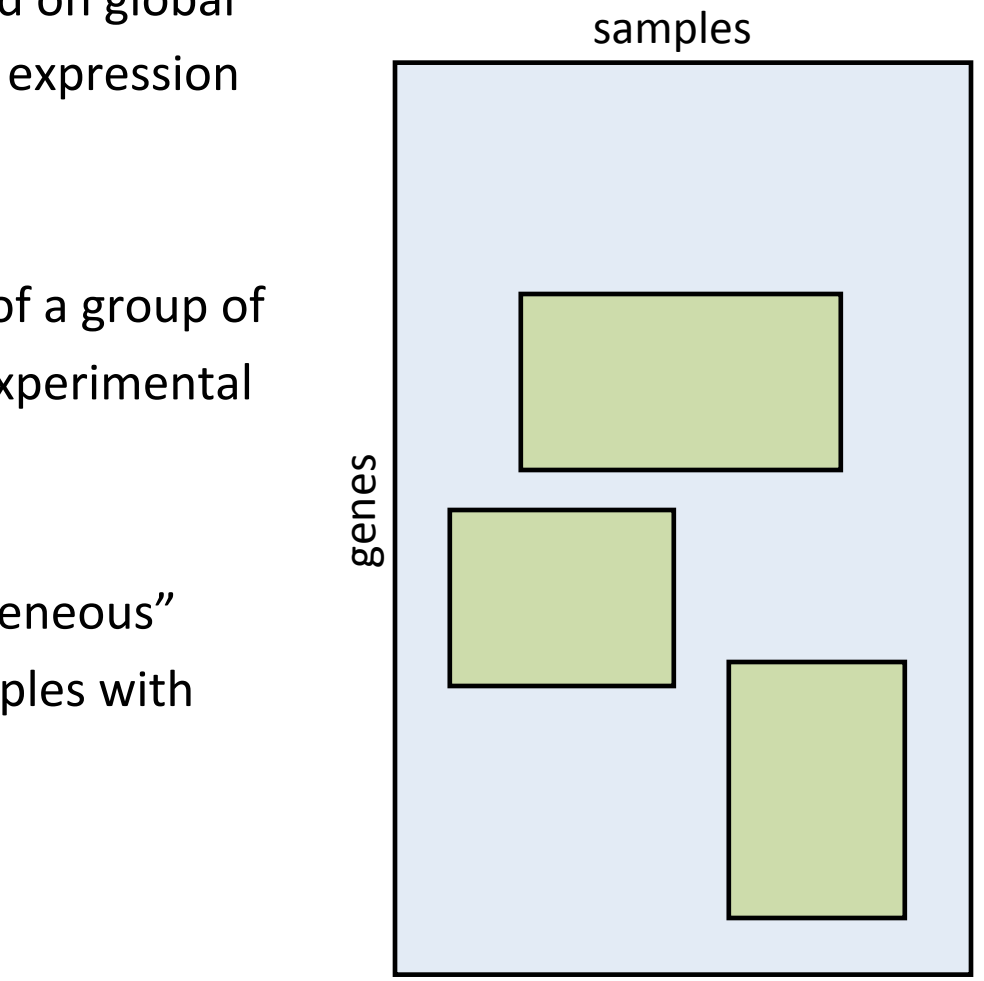
Silhouette analysis



n_clusters = 2 The average silhouette_score is : 0.70
n_clusters = 3 The average silhouette_score is : 0.59
n_clusters = 4 The average silhouette_score is : 0.65
n_clusters = 5 The average silhouette_score is : 0.56
n_clusters = 6 The average silhouette_score is : 0.45

Biclustering

- Usual clustering algorithms are based on global similarities of rows or columns of an expression data matrix
- Similarity of the expression profiles of a group of genes may be restricted to certain experimental conditions
- Aim of biclustering: identify “homogeneous” submatrices (i.e. find subsets of samples with similar values of genes)
- Computational complexity
- Example: Tanay A et al. Bioinformatics. 2002.



Cluster validation

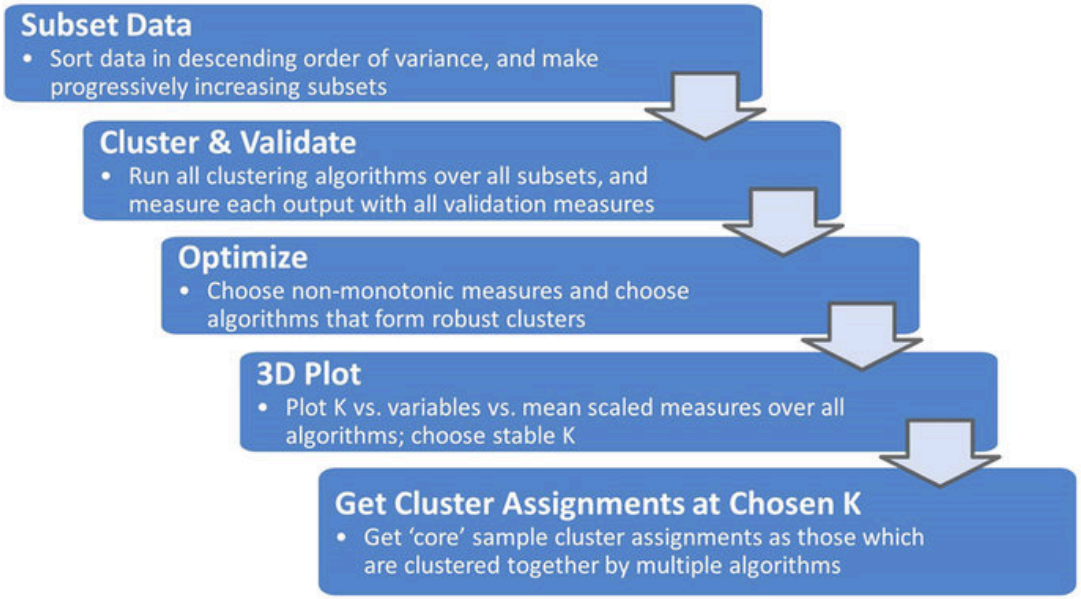
- There are three different techniques for evaluating the result of the clustering Algorithms:
  - **External Criteria**: used to measure the extent of which cluster labels match externally supplied class labels
  - **Internal Criteria**: used to measure the goodness of a clustering structure without respect to external information
  - **Relative Criteria**: used to compare two different clustering or clusters

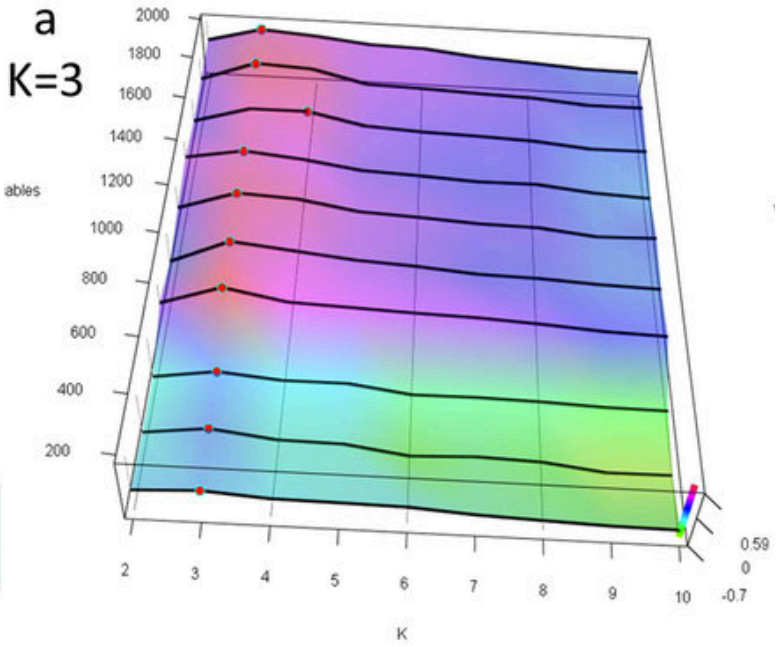
- Two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme (Berry and Linoff, 1996):
  - **Compactness**: The member of each cluster should be as close to each other as possible. A common measure of compactness is the variance
  - **Separation**: The clusters themselves should be widely separated

Cluster validation

- **COMMUNAL**: Combined Mapping of Multiple clUsteriNg ALgorithms (Sweeney TE et al. Sci Rep. 2015)

- Uses multiple clustering algorithms, multiple validity metrics, and progressively bigger subsets of the data to produce an intuitive 3D map of cluster stability that can help determine the optimal number of clusters in a data set





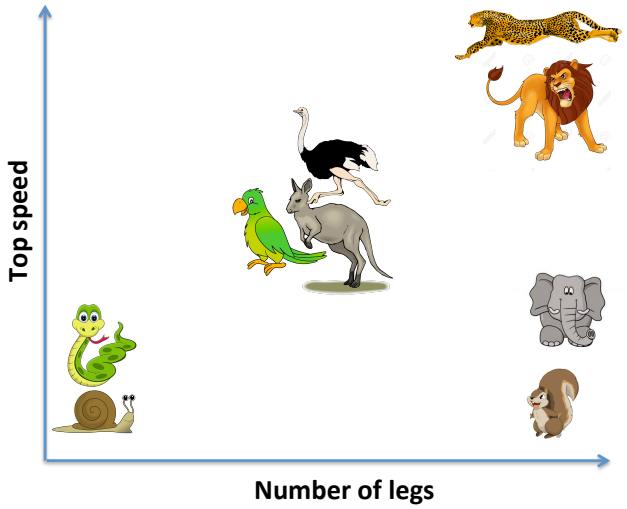
Cluster validation

- **Biological validation** (gene clustering): enrichment of functional categories within clusters

- **Validation in independent datasets** (gene or sample clustering)
  - "in-group proportion" (**IGP**): validation procedure for clusters found in datasets independent of the one in which they were characterized (Kapp AV et al. Biostatistics. 2007)
  - **SubMap** (Hoshida Y et al. PLoS One 2007): reveals common subtypes between independent data sets
    - Define a measure of correspondence for subtypes using gene set enrichment analysis

Clustering - summary

- Useful as **exploratory/visualisation** tools
- Useful to identify **biologically/clinically relevant** subtypes
- Choice of metric, methods and parameters usually guided by prior knowledge about the question…
- The result is guided by what you are looking for
- **Clustering cannot NOT work. Always produce some clusters!**



Clustering software in R/Bioconductor

Package	Function	What
stats	hclust	hierarchical clustering
	heatmap	color image with dendrogram
	kmeans	k-means
	dist	distance
mclust		model-based clustering
Class	SOM	self-organizing maps
Cluster	pam	partition around medoids
	clara, fanny,	
	diana, agnes,	hierarchical clust. (divisive, agglomerative)
	mona,	
	silhouette	silhouette (choose number of clusters)
	daisy	distance
hopach		hierarchical ordered partitioning and collapsing hybrid
pvclust		hierarchical with cluster reliability assessment
e1071	bclust, cmeans	
bioDist		additional distance functions
clusterRepro	clusterRepro	IGP measure for cluster validation

Downstream analyses

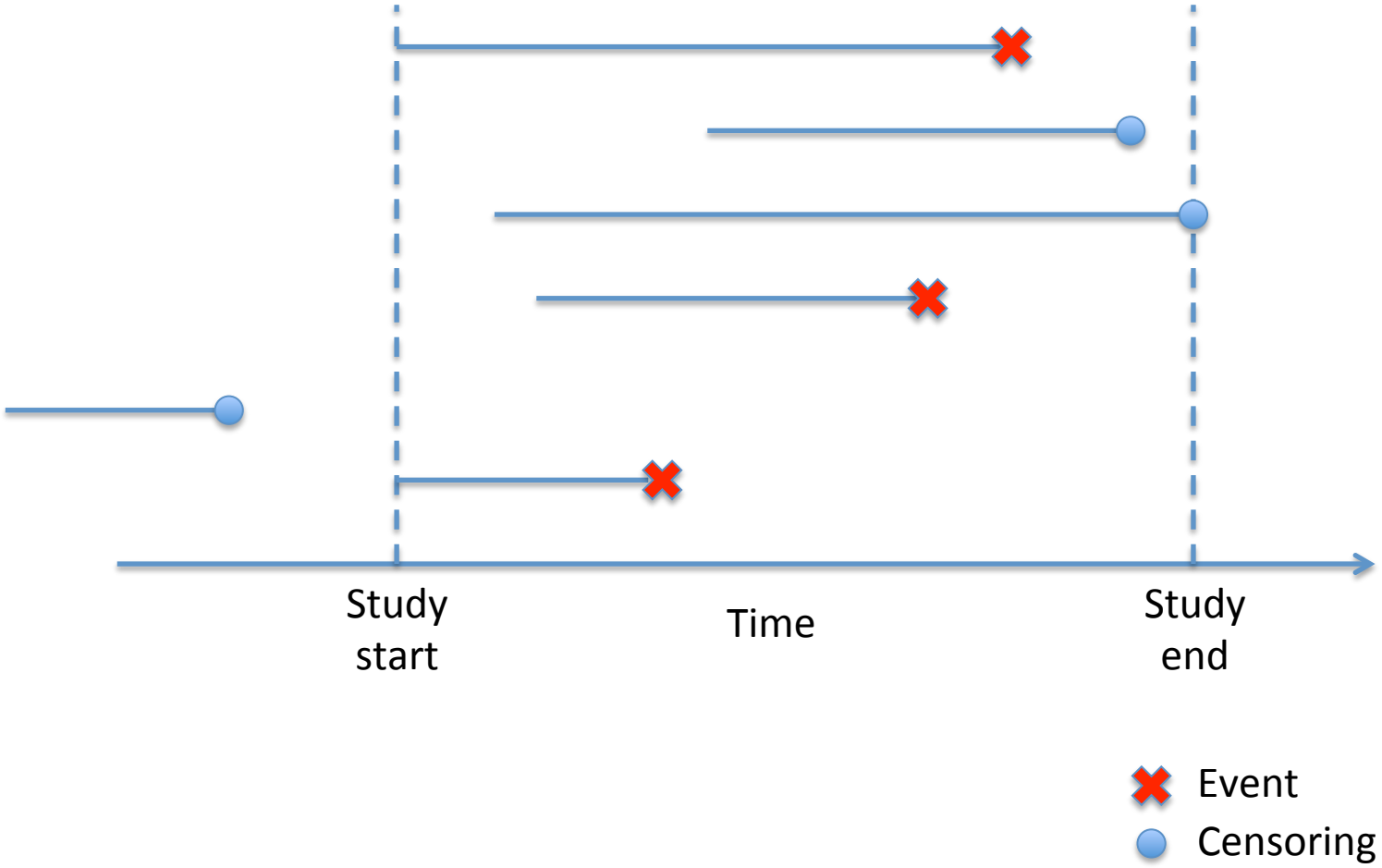
The analyses to answer your biological question will fall in one of three major categories:

- **Class discovery**
  - When classes are unknown (e.g. discovery of subtypes in a tumour type)
  - Approaches: hierarchical clustering, K-means clustering…
- **Class comparison**
  - When the groups or classes are known and we want to identify genes (or pathways) associated with them (e.g. treated/untreated cells)
  - Approaches: t-test, linear models, pathway analysis…
- **Class prediction**
  - When we want to identify a set of genes able to accurately predict the classes of interest in independent data
  - Approaches: PAM, SVM…
  - 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 genes or set of genes (**signatures**) that are related with **prognosis**

Censoring



Censoring

Times are often **censored**: we are not able to observe the failure times for all individuals

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).

Censoring

- **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.

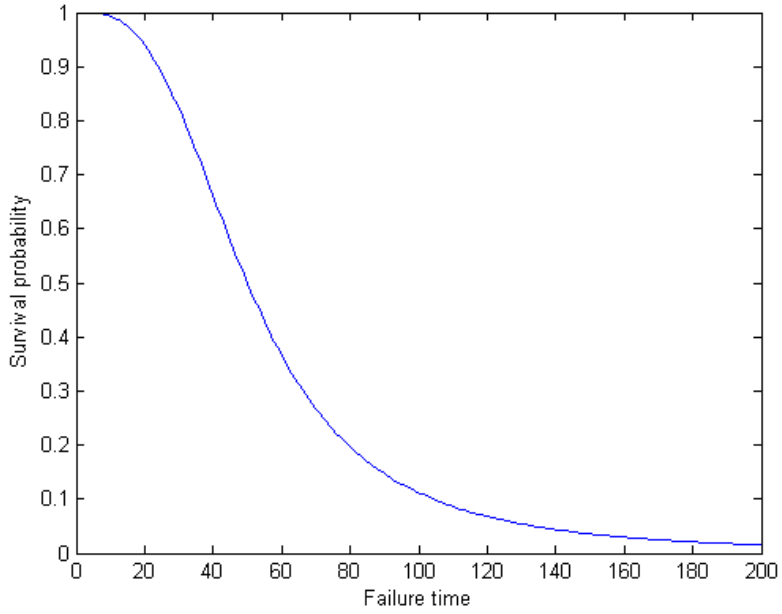
Functions of interest

Survival function:

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

T=time until event

F*(t)*=cumulative distribution function for T



Functions of interest

Hazard function (or instantaneous failure rate)

Related to the probability that the event will occur in a small interval around t, given that the event has not occurred before time t

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

f*(t)* = probability density function of T at time t

Kaplan-Meier estimator

Empirical survival function

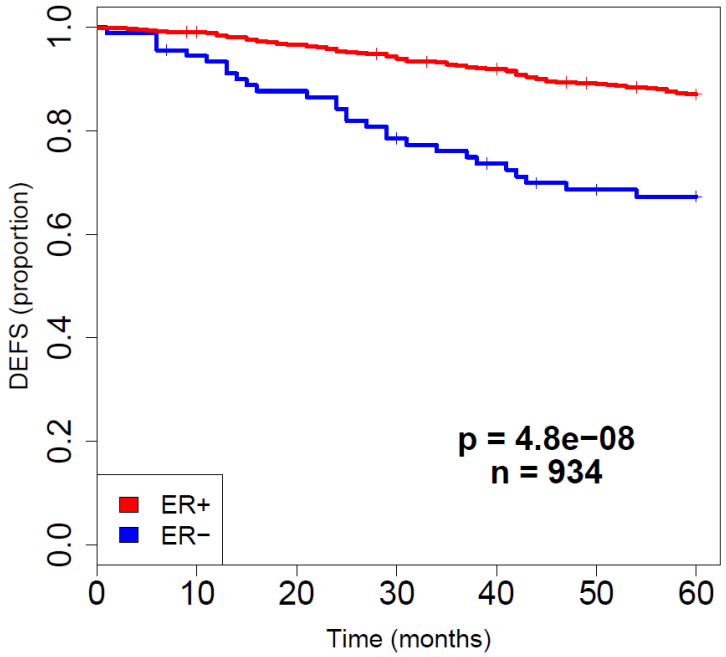
$$S_{KM}(t) = \prod_{i:t_i < t} (1 - d_i / n_i)$$

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	90/93 x 0.97 = 0.94

Log-rank test

- Tests **differences between the survival** functions for two or more groups
- H_0 = no difference between (true) survival curves
- Compares observed and expected events in each group



Cox model

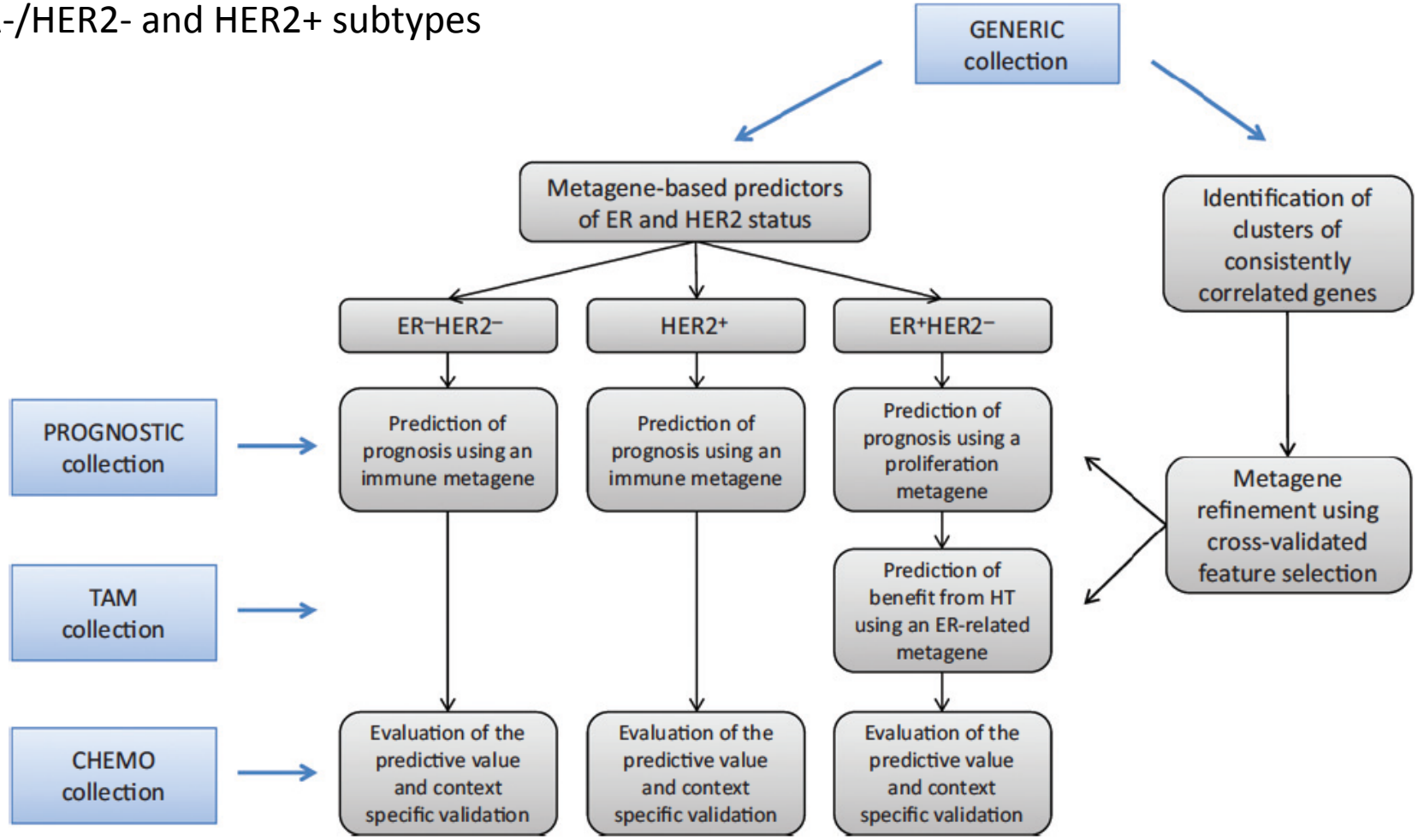
Semiparametric proportional hazards model

$$\lambda(t \mid X) = \lambda_0(t)\exp(X\beta)$$

- Uses a partial likelihood to estimate β
- 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

Example study

- Prediction of outcome in early breast cancer patients
- Identification of a cluster of immune related genes
- ER-/HER2- and HER2+ subtypes



Callari et al. CCR 2016

Example study

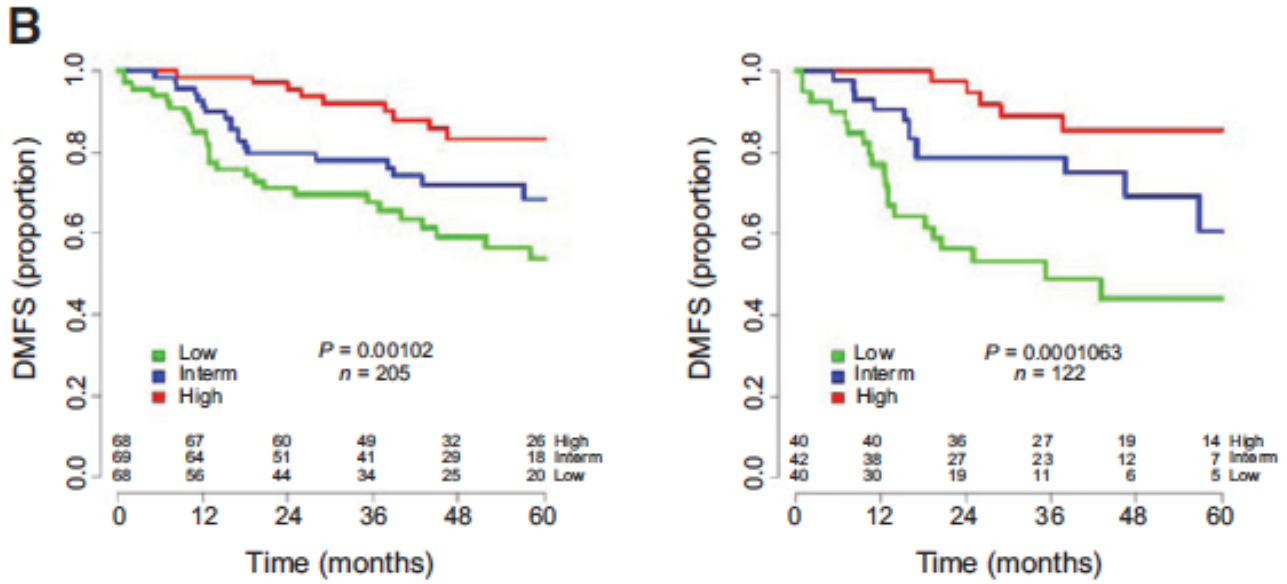


Table 3. Multivariable Cox analysis in 371 HER2$^-$ patients treated with adjuvant chemotherapy

	All HER2$^-$		ER$^-$HER2$^-$	
Variables	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***
MBRP				
High vs. low	3.53 (1.57–7.92)	0.0022	6.39 (1.37–29.86)	0.0183
Interm vs. low	2.54 (1.19–5.43)	0.0165	8.25 (1.60–42.57)	0.0117
Node pos (vs. neg)	2.38 (1.17–4.83)	0.0169	3.20 (0.90–11.42)	0.0730
Grade III (vs. I and II)	1.55 (0.90–2.68)	0.1172	>100 (0.00–inf)	0.9980
Age >50 (vs. ≤50)	0.73 (0.44–1.22)	0.2301	1.34 (0.49–3.62)	0.5660

R functions and packages

- Package *survival*:

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

References

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.

Tamayo P et al. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation". Proc Natl Acad Sci U S A, 1999 Mar 16;96(6):2907-12.

Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. 2002 Jun 25;3(7)

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987 Nov 20;53-65

Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. 2002;18 Suppl 1:S136-44

Hoshida Y, Brunet JP, Tamayo P, Golub TR, Mesirov JP. Subclass mapping: identifying common subtypes in independent disease data sets. PLoS One. 2007 Nov 21;2(11):e1195

Sweeney TE, Chen AC, Gevaert O. Combined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number, K. Sci Rep. 2015 Nov 19;5:16971

Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? Biostatistics. 2007 Jan;8(1):9-31

Callari M, Cappelletti V, D'Aiuto F, Musella V, Lembo A, Petel F, Karn T, Iwamoto T, Provero P, Daidone MG, Gianni L, Bianchini G. Subtype-Specific Metagene-Based Prediction of Outcome after Neoadjuvant and Adjuvant Treatment in Breast Cancer. Clin Cancer Res. 2016 Jan 15;22(2):337-45