Analysis of high dimensionality datasets

Francesca M. Buffa

Computational Biology &
Integrative Genomics

University of Oxford

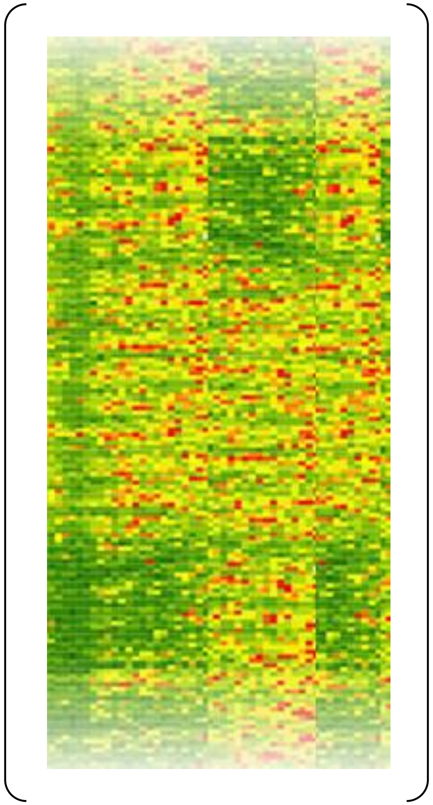
Big (genomic) data

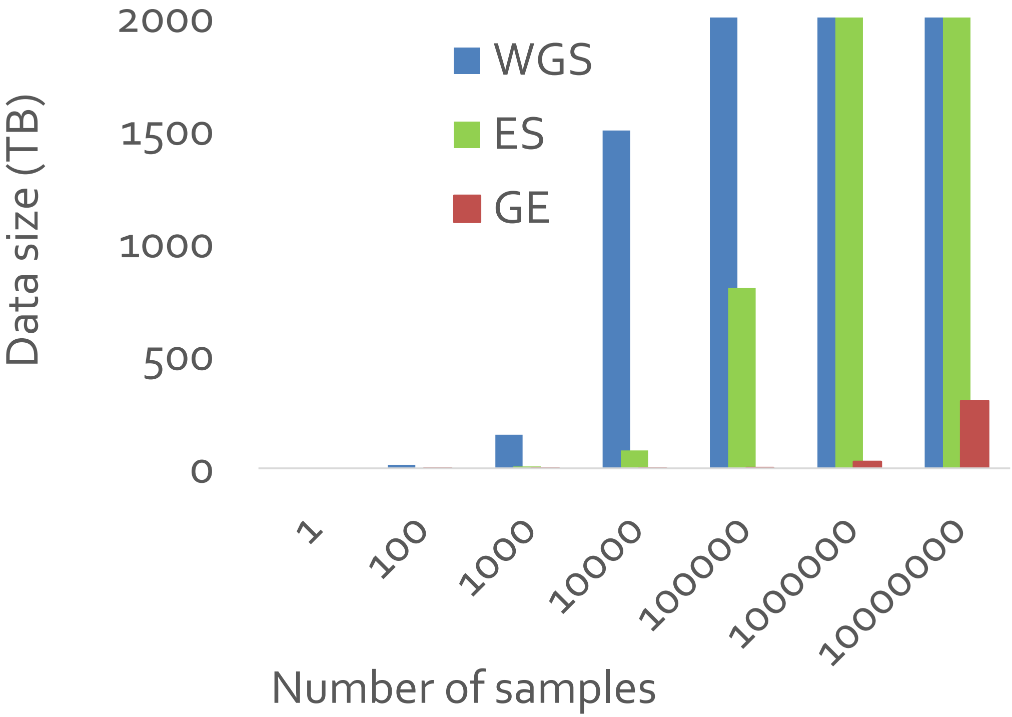
Samples

10-100

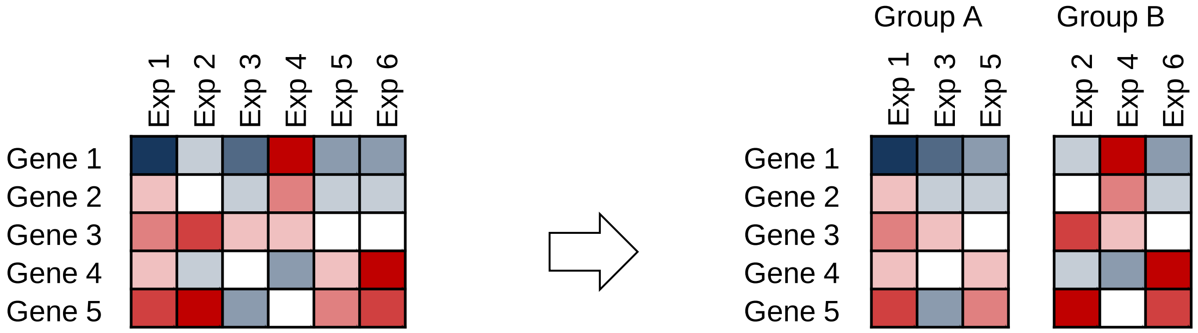
Variables

-> 1,000,000





Question: which genes are differently expressed between group A and group B?



For each gene:

• We state a Null Hypothesis (H0): no difference and assume that H0 is true

• We set a level of significance: decide how much evidence do we want before rejecting H_0

• We calculated the observed test statistics: score which measures how far is our sample from the null

• We obtain a p-value : probability of observing this or higher scores if H0 is true

• We make a decision to accept or reject the null hypothesis based on the p-value

Breakdown of basic assumption: Improbable events do occur if you try enough times!

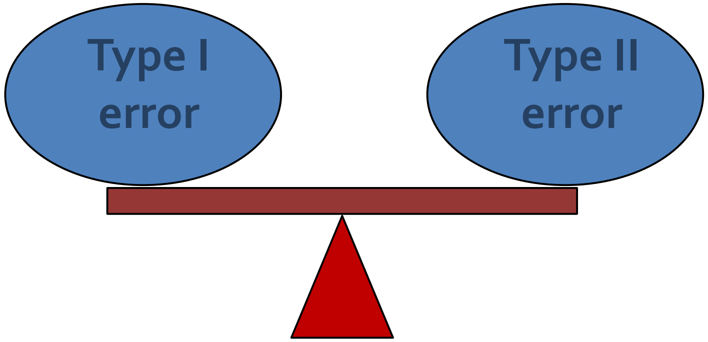
- The confidence level is typically set to 5% ($p=0.05$)
  - The risk of falsely rejecting the null hypothesis (**false positive**) **in one try** is 5%
  - The chance of accepting a true null hypothesis two tries is $0.95^2=90.25\%$
    - The risk of falsely rejecting a true null hypothesis (**false positive**) in at least one of two tries is 1-0.95^2=9.75%
  - In 10 tries: $1-0.95^{10}=40\%$

- The confidence is *eaten up* by multiple testing

- Multiple test correction needed

Example: **Bonferroni correction** divides confidence level by number of trials *(stringent correction)*

Multiple test correction



Balance between
False Positives
and
False Negatives

Several approaches.

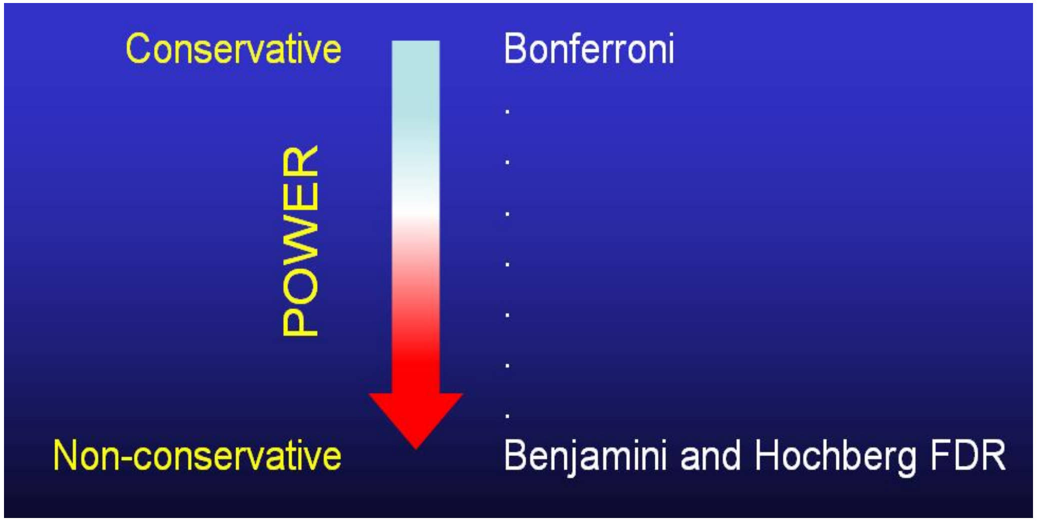
For reviews see e.g.:

Dudoit et al. (2003). *Statistical Science 18, p.71*
Noble (2009). Nature Biotechnology 27, p.1135

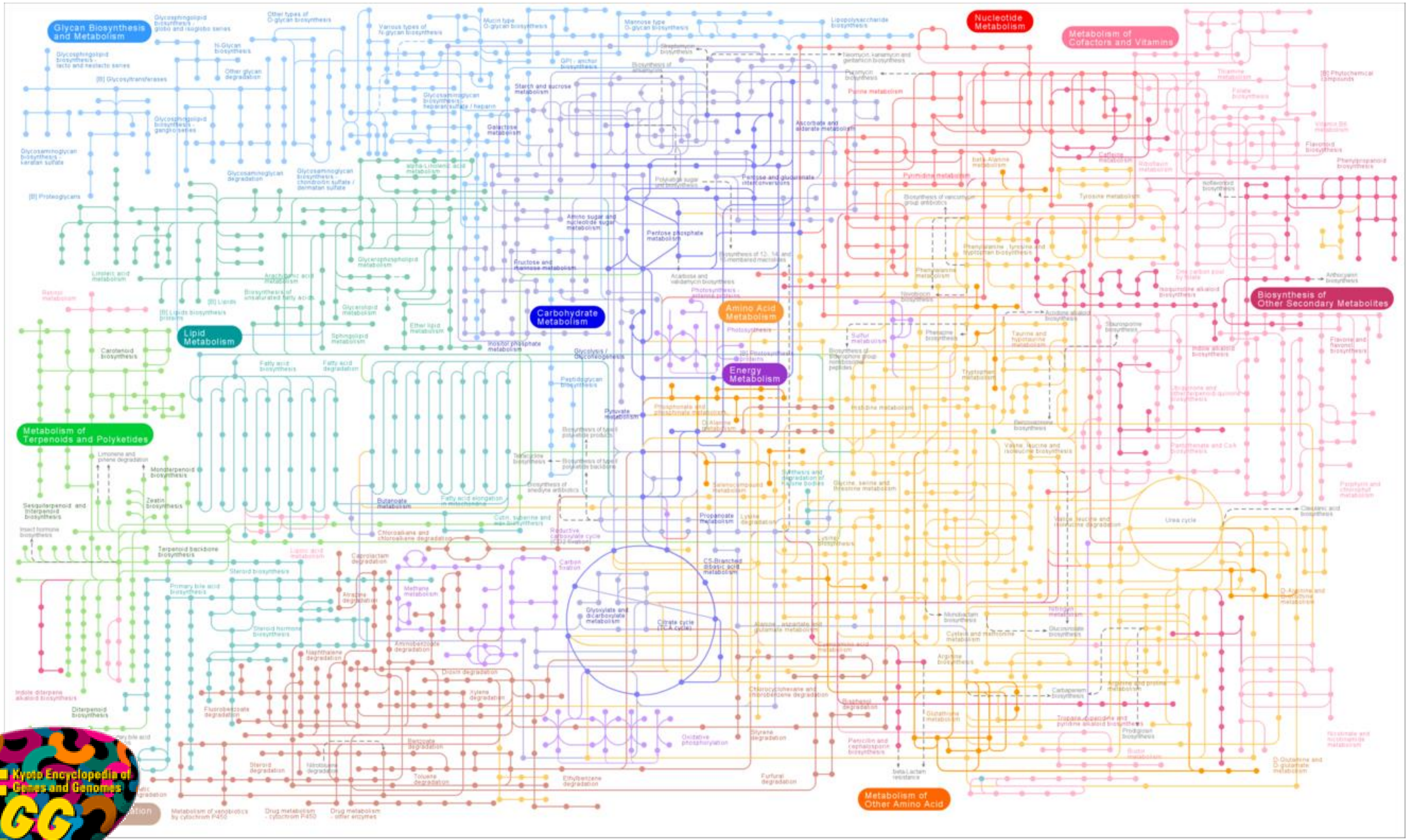
What is the cost of false positives and false negatives ?

e.g. You have done an extremely complicated/long/expensive experiment and further validation is cheap and easy.

In this case the **cost of a false positive** (if validation shows no effect on gene expression) is little but the **cost of a false negative** is large as you could miss out on an important discovery after having already carried out the biggest/most expensive task.



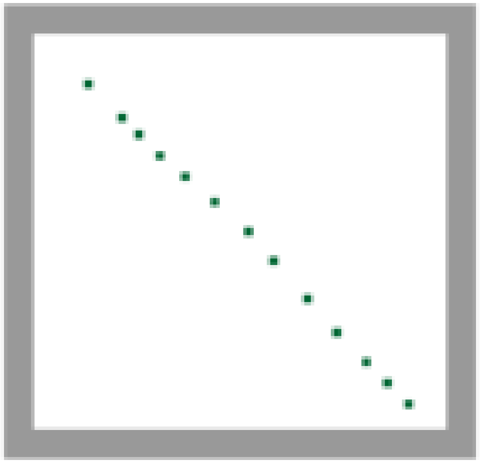
Genes don’t work in isolation


Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Metabolism of Other Amino Acid

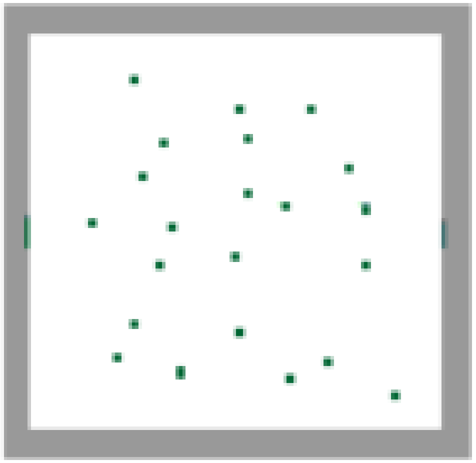
Kyoto Encyclopedia of Genes and Genomes

Do some of these variables (genes) vary together across samples?

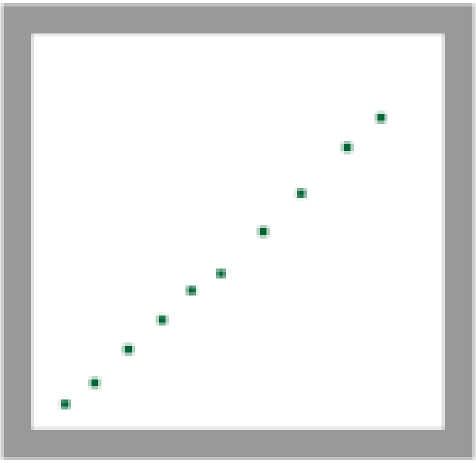
Covariance



Large negative covariance



Near zero covariance



Large positive covariance

Note: Correlation

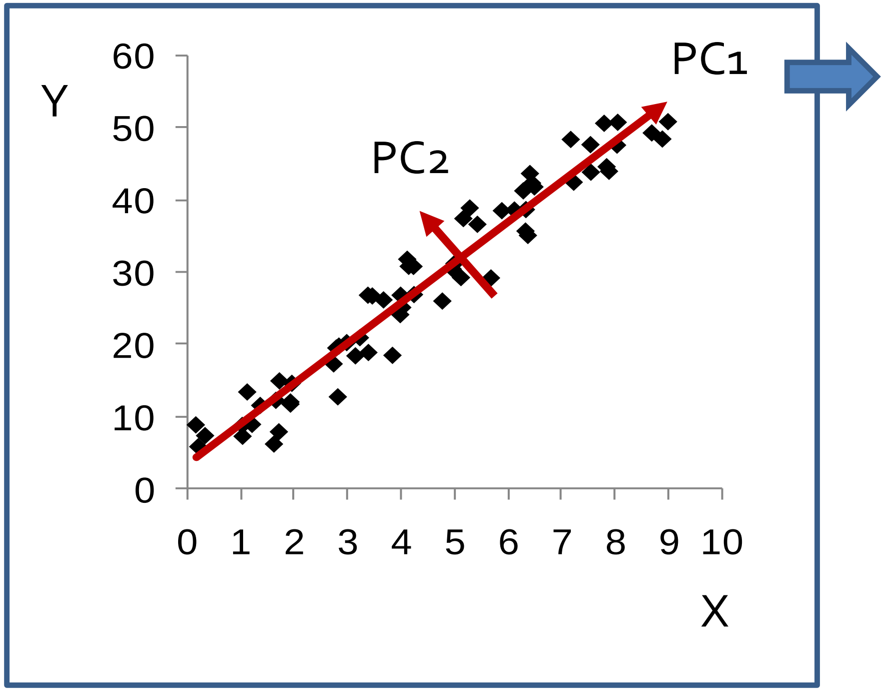
- Large covariance = strong relationship between variables.
- However, we cannot compare covariance over data sets with different scales.
- The larger the X and Y values, the larger the covariance. So a weak covariance in one data set may be a strong one in a different data set with different scales.
- The problem can be fixed by dividing the covariance by the standard deviation to get the correlation coefficient.

$$\mathrm{Corr}(X,Y) = \mathrm{Cov}(X,Y) \,/\, \sigma_X\sigma_Y$$

Principal component analysis

PCs are **linear combinations** of the original variables.

PCs are orthogonal to each other ▶ **no redundant information**.

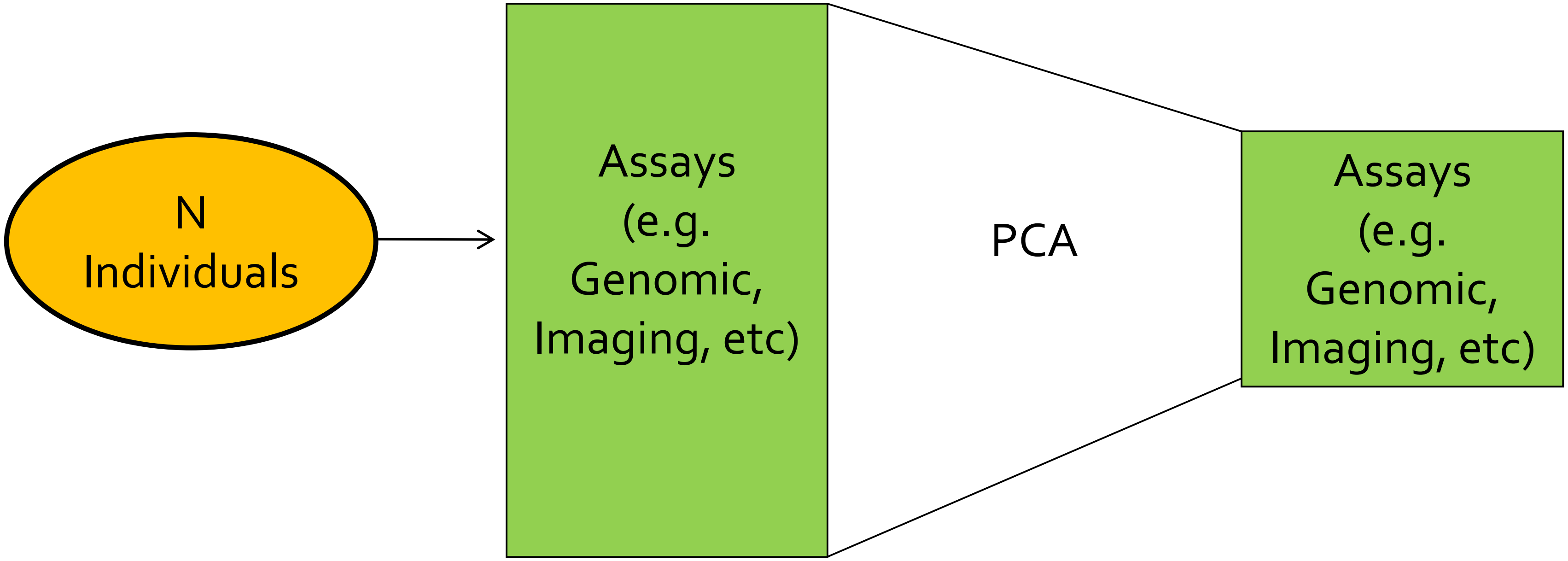


Largest PC maximizes the variance of the projected data.

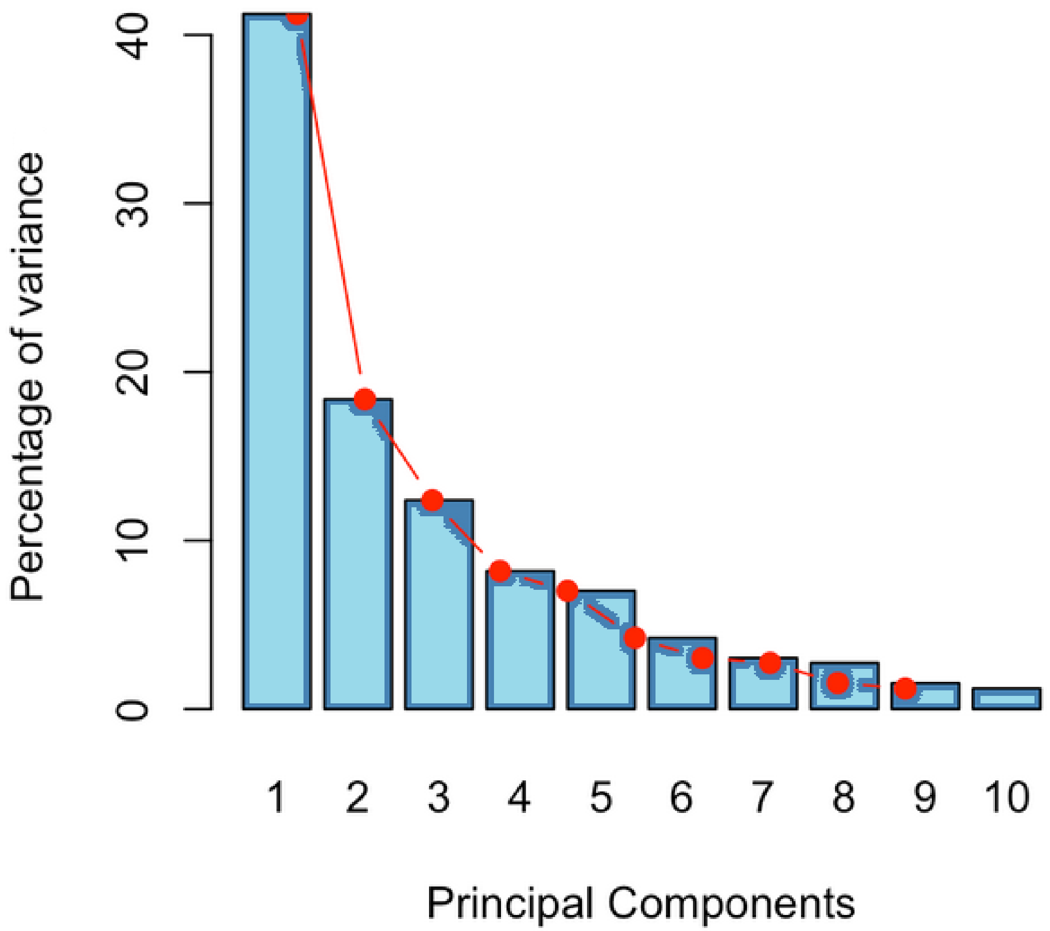
If there are dependencies amongst the measured variables
the variance of the original data can be explained by the first few PCs.

PCA: dimensionality reduction

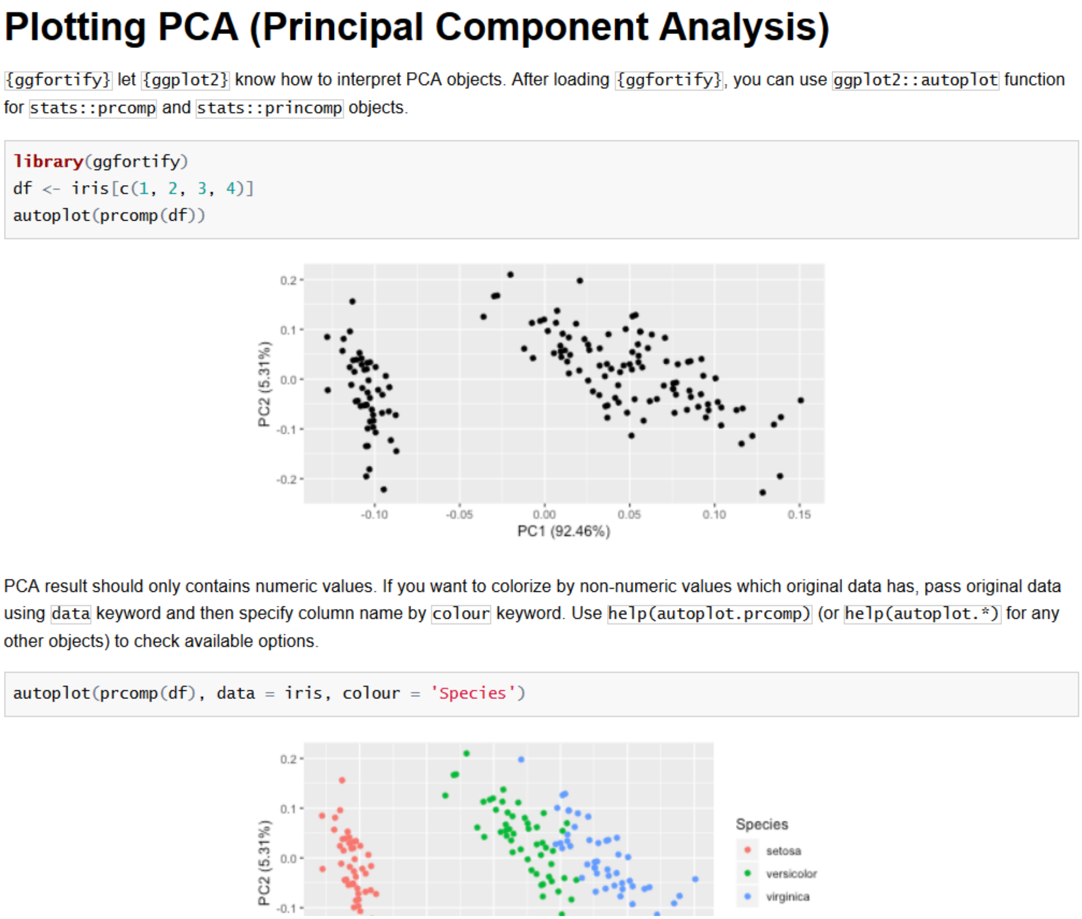
Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions.



PCA: ranked by percentage of variance explained



Plotting PCA



Plotting PCA (Principal Component Analysis)

`{ggfortify}` let `{ggplot2}` know how to interpret PCA objects. After loading `{ggfortify}`, you can use `ggplot2::autoplot` function for `stats::prcomp` and `stats::princomp` objects.

```
library(ggfortify)
df <- iris[c(1, 2, 3, 4)]
autoplot(prcomp(df))
```

PCA result should only contains numeric values. If you want to colorize by non-numeric values which original data has, pass original data using `data` keyword and then specify column name by `colour` keyword. Use `help(autoplot.prcomp)` (or `help(autoplot.*)` for any other objects) to check available options.

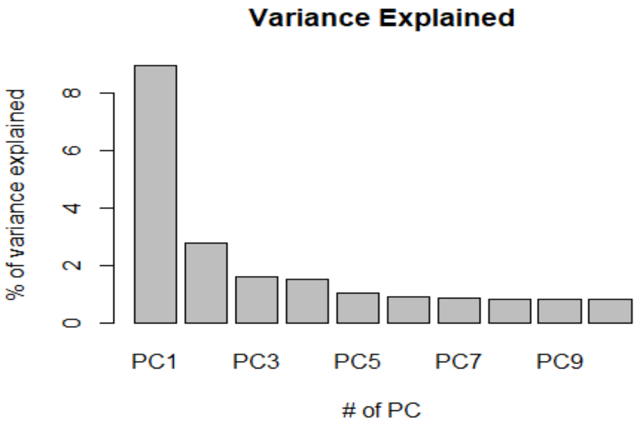
```
autoplot(prcomp(df), data = iris, colour = 'Species')
```

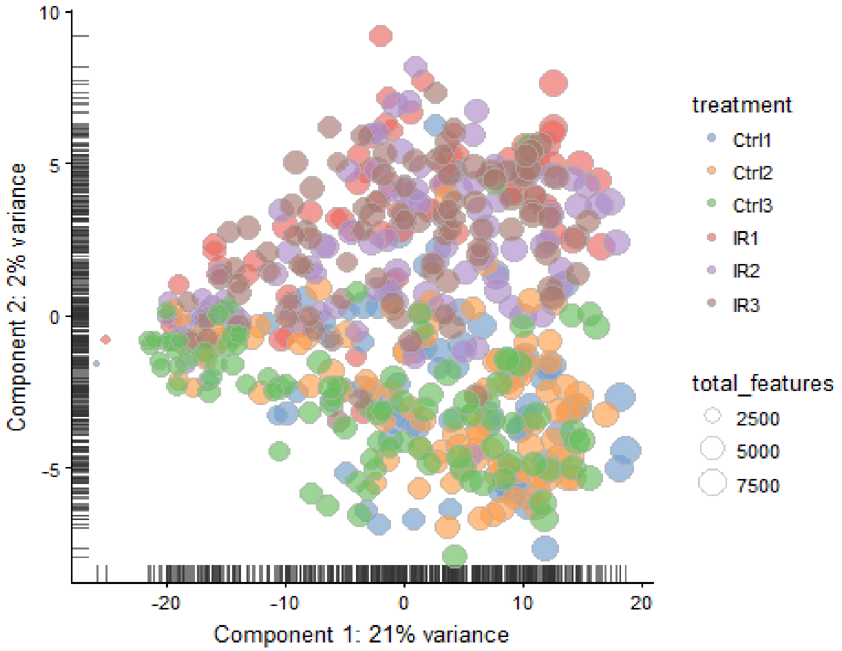
https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

Example: Single cell sequencing experiment

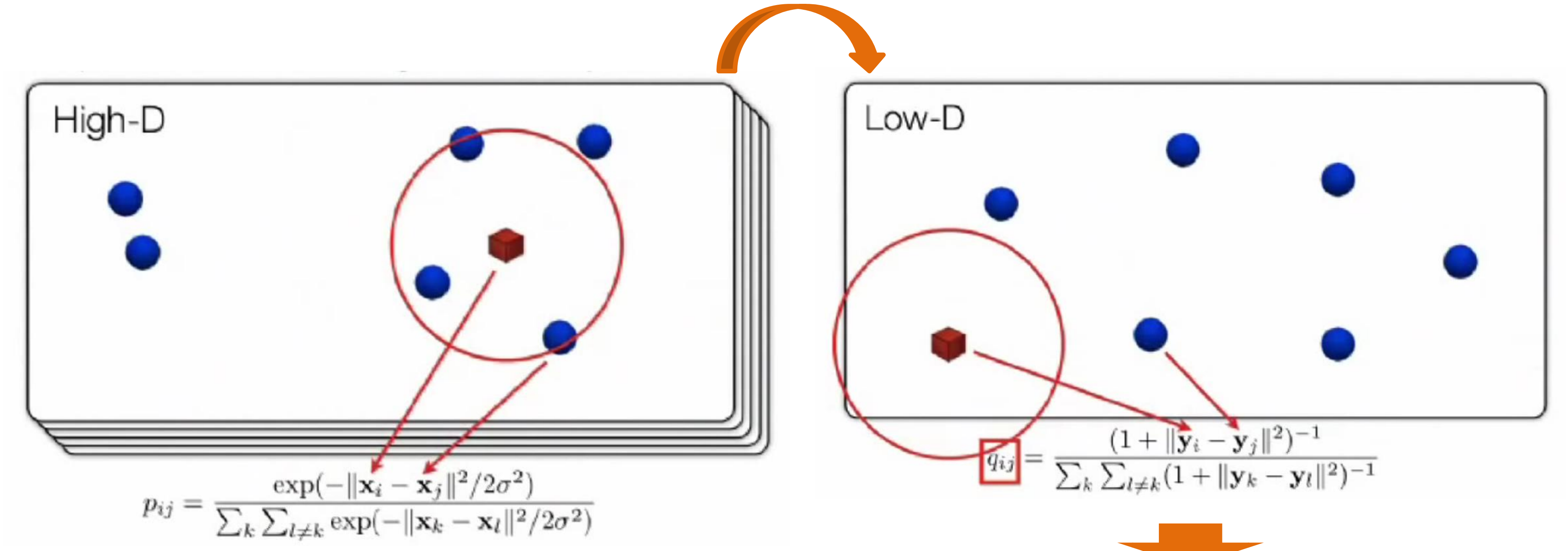
#Samples (IR vs no IR) X #Cells
528

Variables
26376





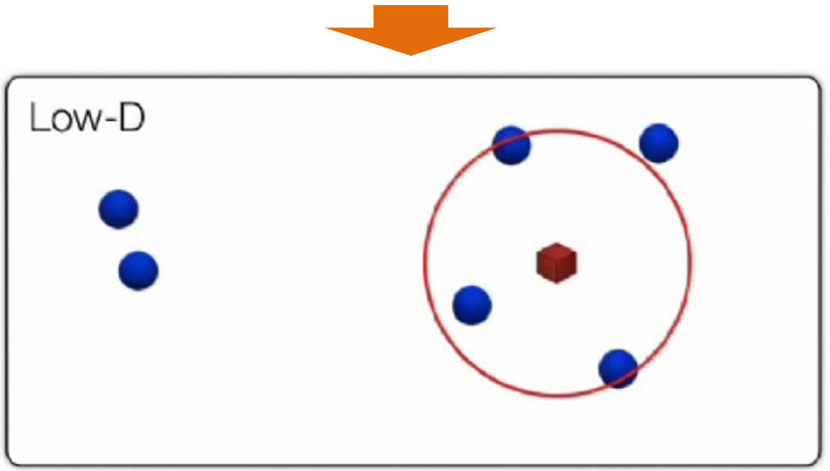
t-distributed Stochastic Neighbor Embedding



$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)}{\sum_k \sum_{l \neq k} \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2 / 2\sigma^2)}$$

$$\boxed{q_{ij}} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

- Measure pairwise similarity between features
- Compute a probability of similarity (accounts for non uniform density)
- Symmetrize the data



Keeps local structure of the data

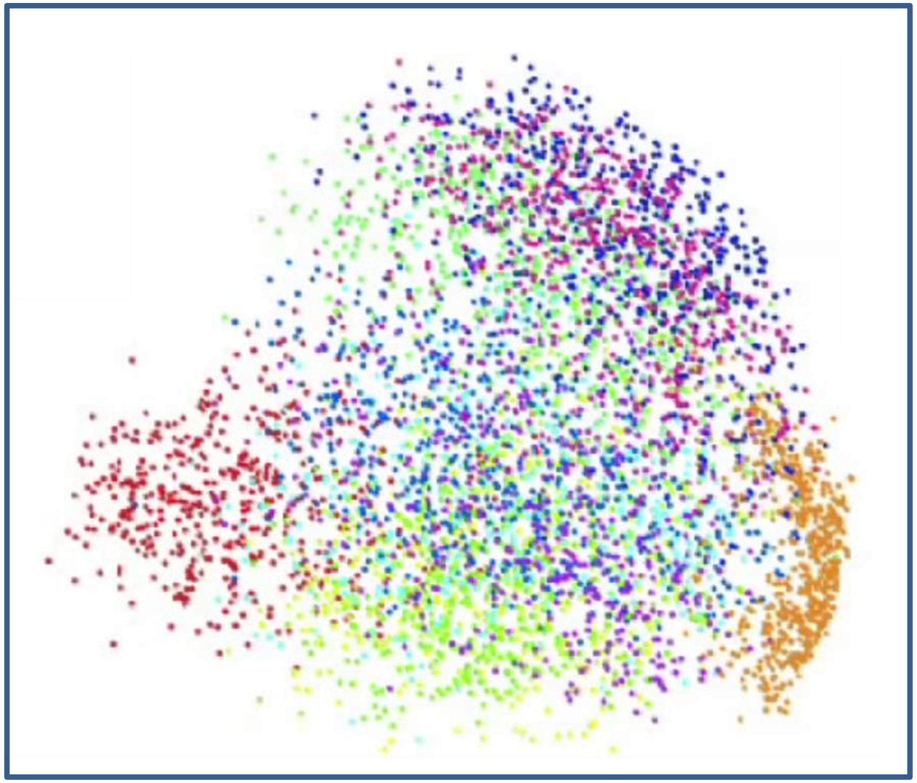
$$KL(P\|Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

t-distributed Stochastic Neighbor Embedding

- t-SNE is not a linear projection. It uses the **local relationships** between points to create a low-dimensional mapping.
- t-SNE creates a **probability distribution** to define relationships between the points in high-dimensional space.
- It then **recreates** the probability distribution in low-dimensional space.

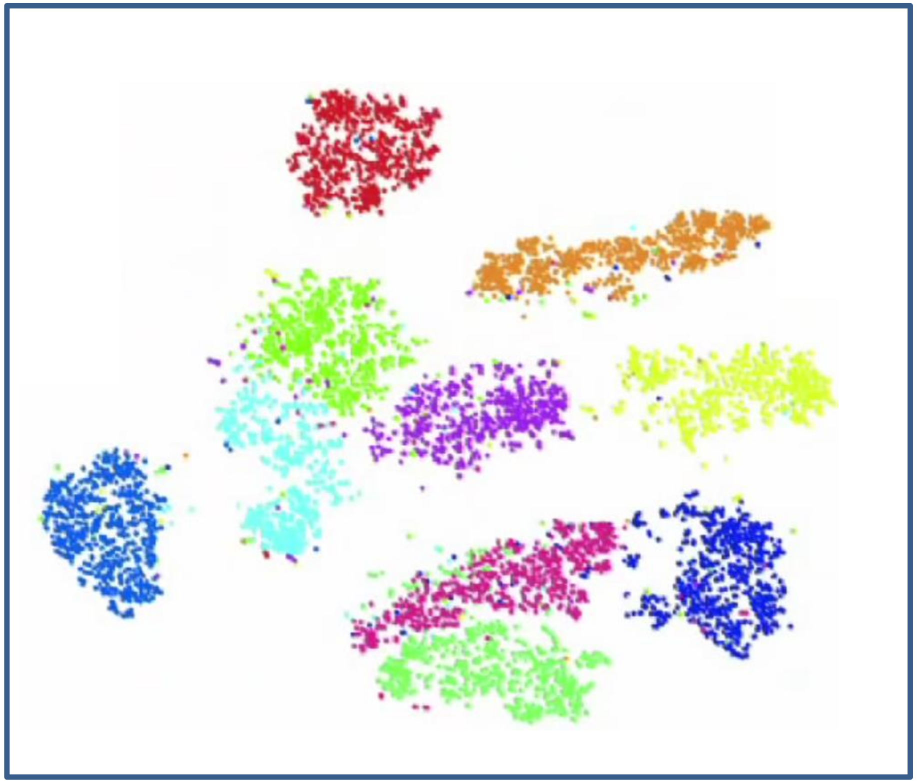
PCA vs t-SNE

PCA



Overall variance

T-SNE



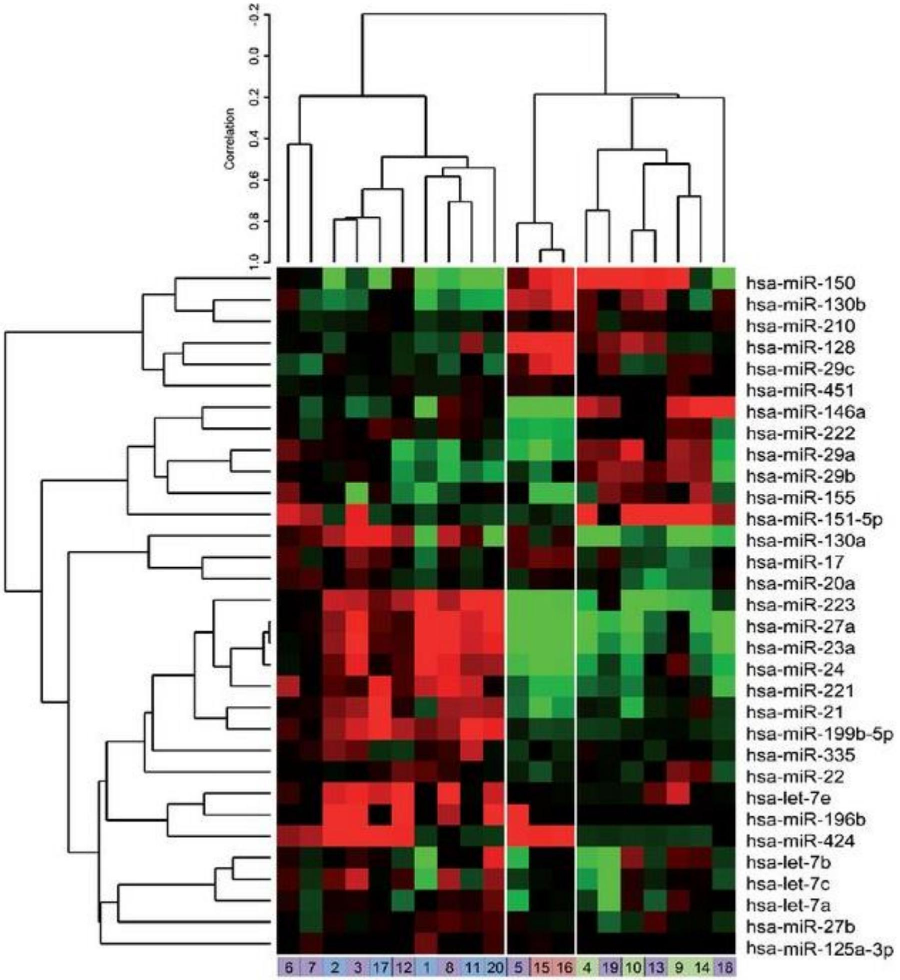
Local structure

https://lvdmaaten.github.io/tsne/

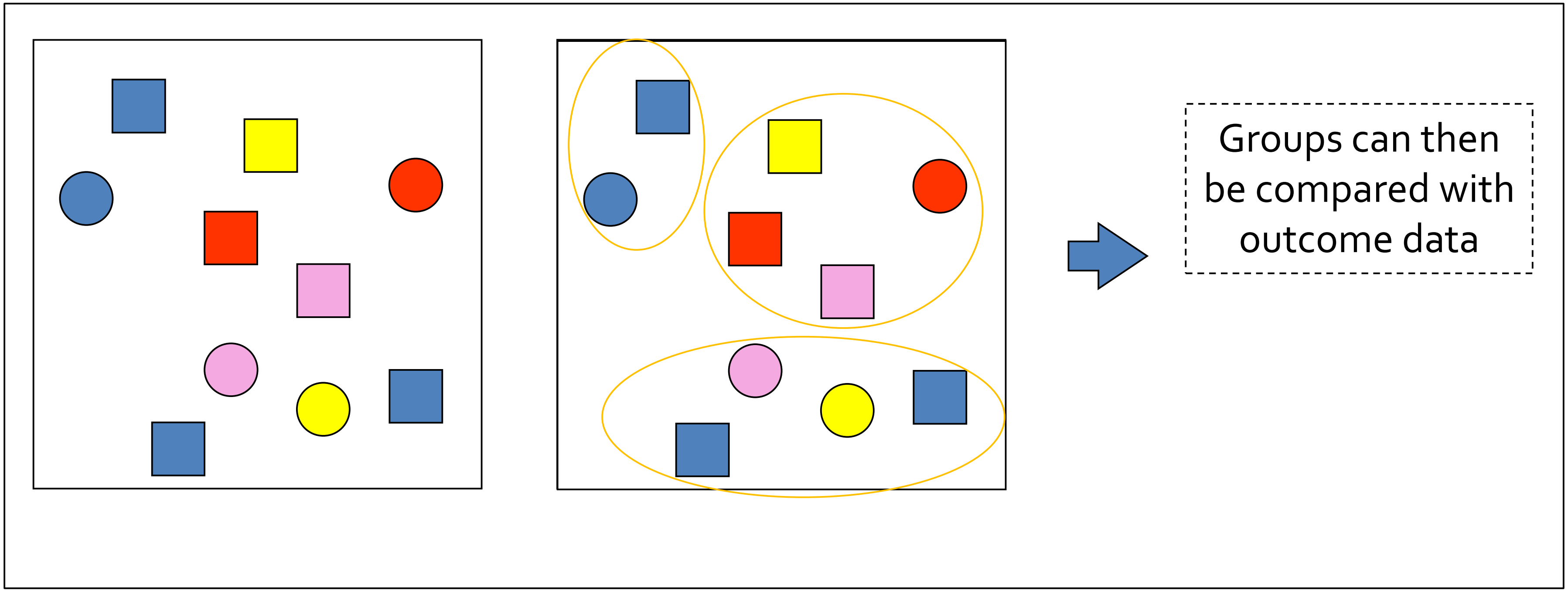
Unsupervised approach: guilty-by-association

Are my sample similar?
Samples with similar genomic profile might for example have a similar prognosis or response to treatment. Or in single cell might come from the same cell population.

Are my gene similar?
Suppose genes A and B are grouped in the same cluster. This mean they are expressed under the same conditions. Then we can hypothesize that genes A and B are involved in similar pathways/share function.



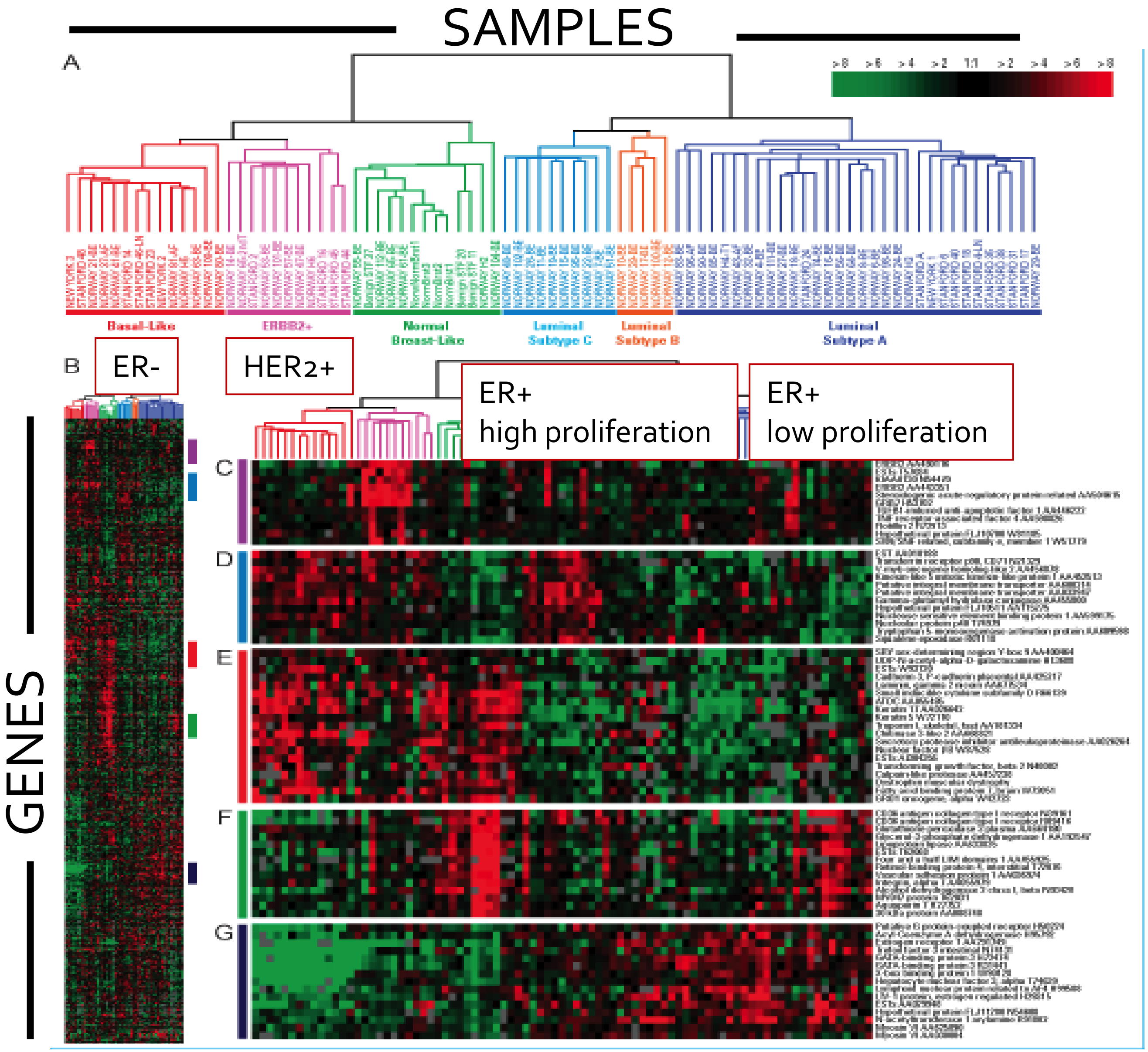
Clustering:
how to define similarity between objects?



Groups can then be compared with outcome data

Choice on distance metric will prioritize different properties of the data
=> Try to focus on what the aim of the clustering is and ask for advise

Example: Breast Cancer Classifier

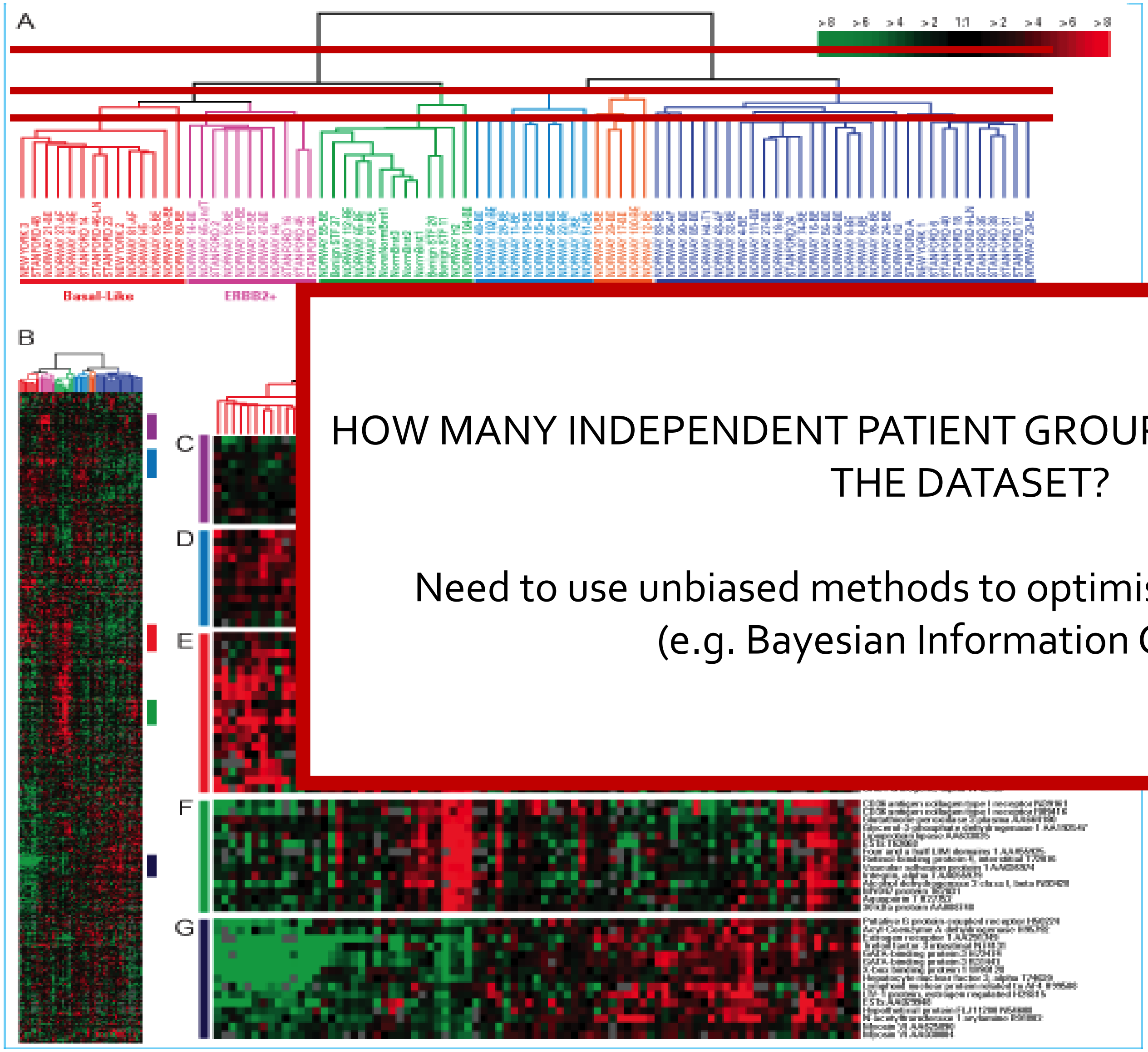


Classification based on Intrinsic Genes

Hierarchical clustering used to classify breast cancer patients based on expression of intrinsic genes in tumour samples

Perou et al, Nature 2000
Sorlie et al, PNAS 2001

How many clusters?



HOW MANY INDEPENDENT PATIENT GROUPS ARE REPRESENTED IN THE DATASET?

Need to use unbiased methods to optimise number of clusters (e.g. Bayesian Information Criteria)

Reproducibility?

Weigelt, B et al, Lancet Oncology, 2010

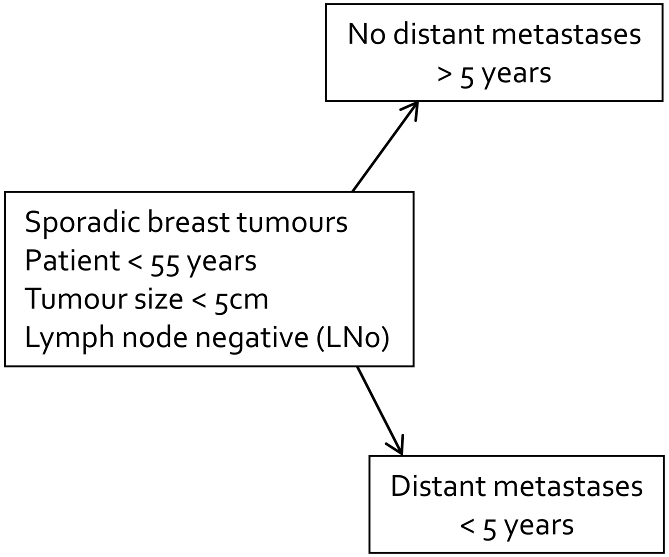
When using different methods to assign patients to each cluster:

- Basal-like cancers were consistently classified.
- Assignment of individual cases to luminal A, luminal B, HER2, and normal breast-like subtypes was dependent on the method used.
- The significance of associations with outcome of each molecular subtype, other than basal-like and luminal A, varied depending on the method used.

Differential expression approaches

- Univariate analysis: testing one gene at the time (e.g. Dseq)
- PCA + univariate analysis
- Cluster of genes + univariate analysis
- Multiple regression: all genes are tested together in the same model

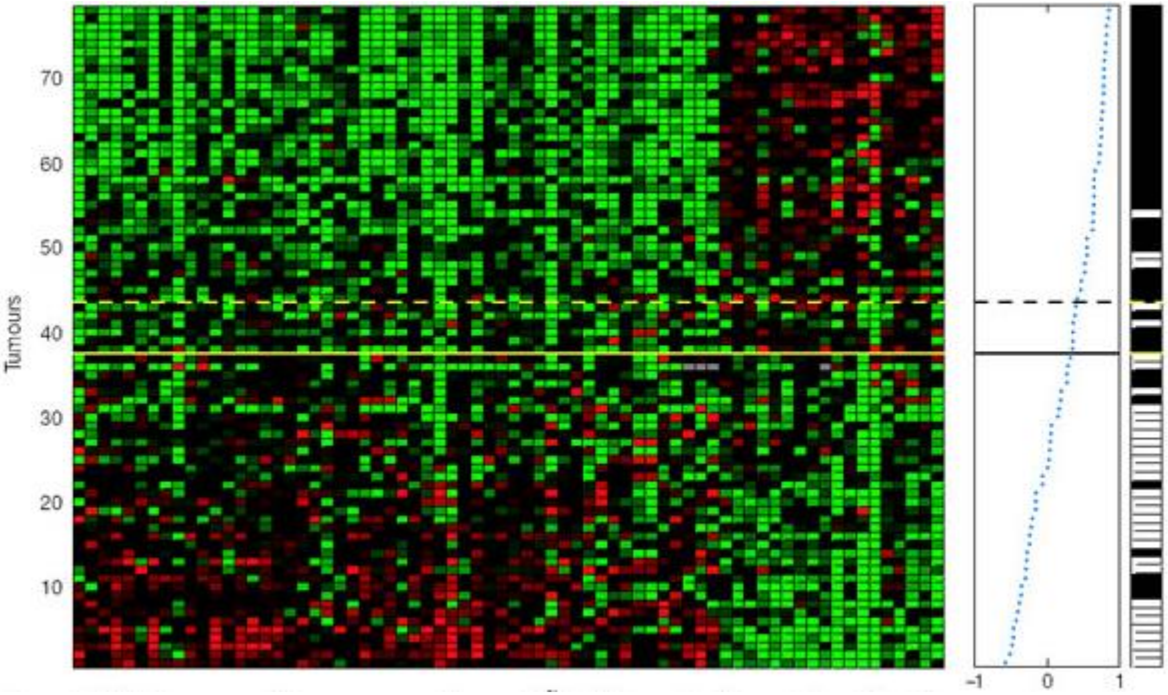
A 70-gene signature for risk of metastasis in breast cancer [Van ‘t Veer et al, Nature, 2000]



Sporadic breast tumours
Patient < 55 years
Tumour size < 5cm
Lymph node negative (LN0)

No distant metastases
> 5 years

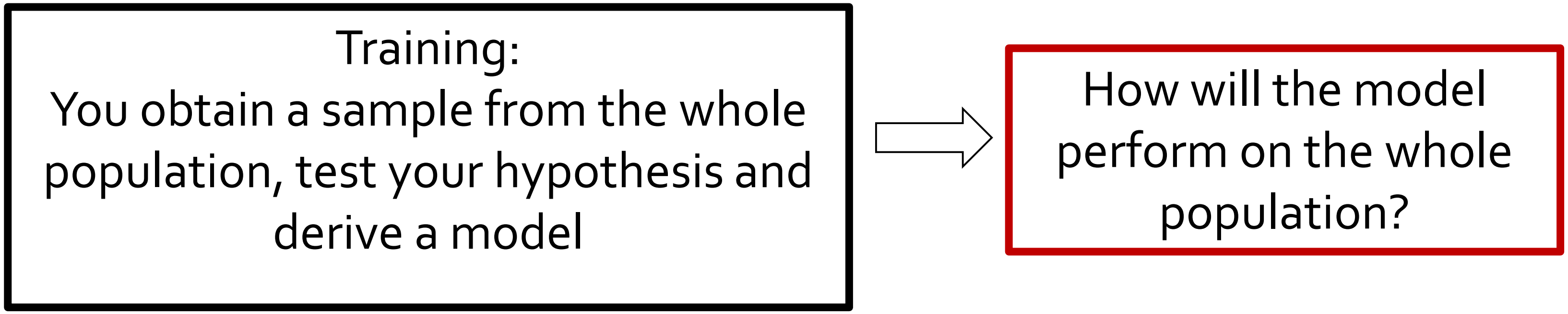
Distant metastases
< 5 years



Correlation to average good prognosis profile

Metastases

Class comparison:
How general are my results?



Training:
You obtain a sample from the whole population, test your hypothesis and derive a model

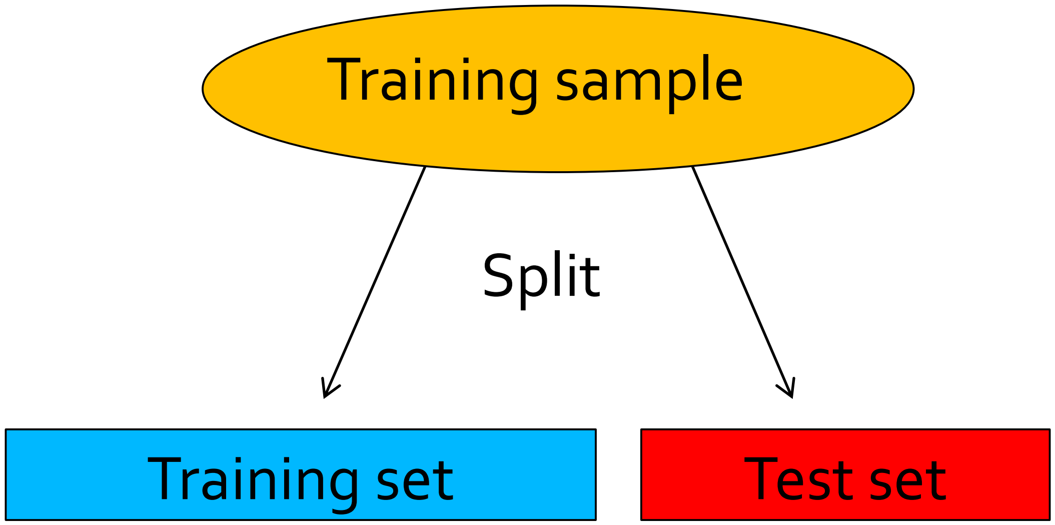
How will the model perform on the whole population?

- How do we estimate the model true error rate? (i.e. the model prediction error rate when tested on the ENTIRE POPULATION)

- If only we had access to an unlimited number of examples this would be easy to estimate….

- But we don’t have unlimited number of examples. In fact we have very limited number of data as data collection is a very expensive process!

Testing the model on an independent set



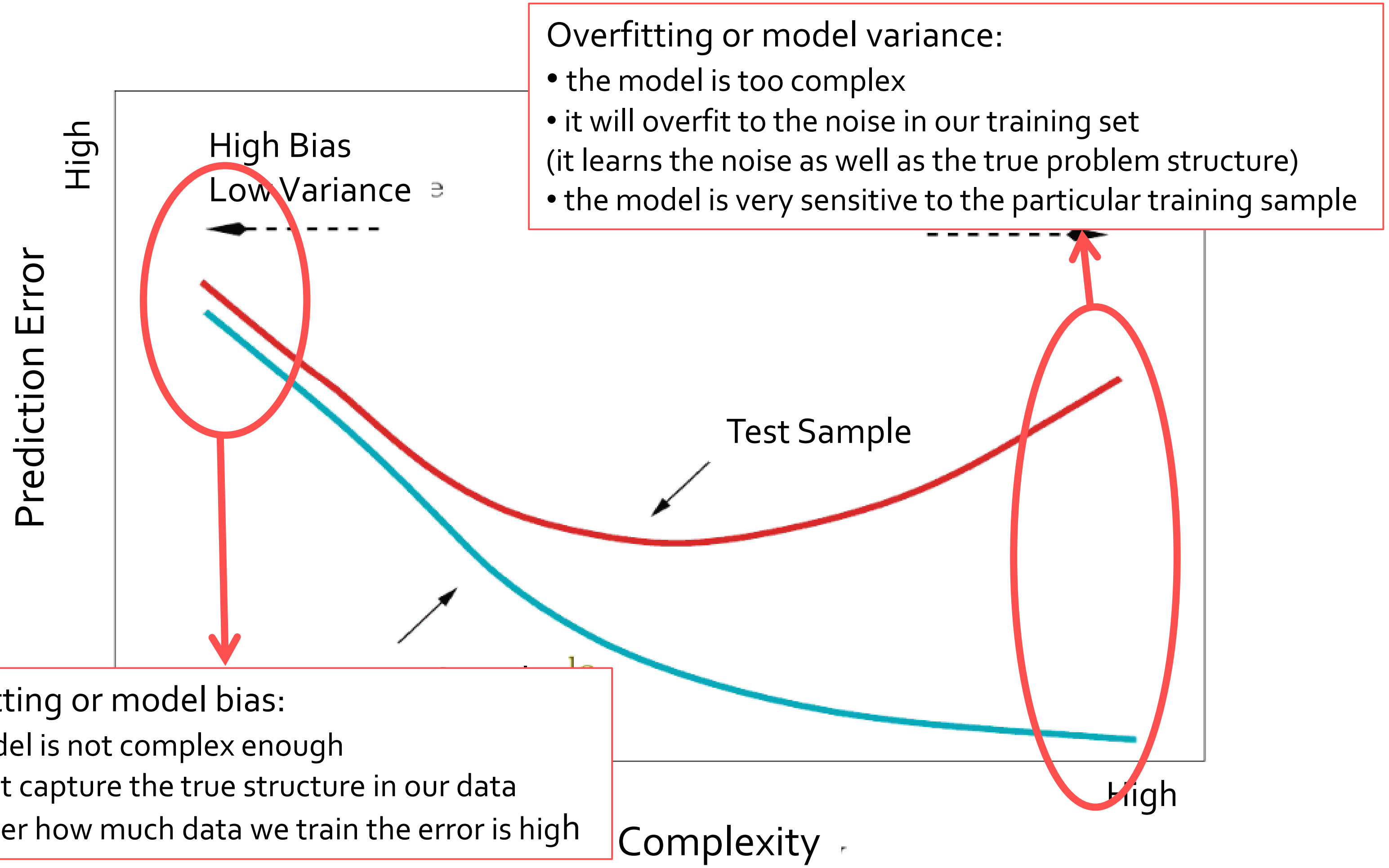
Training sample

Split

Training set

Test set

How general is my model?



Overfitting or model variance:

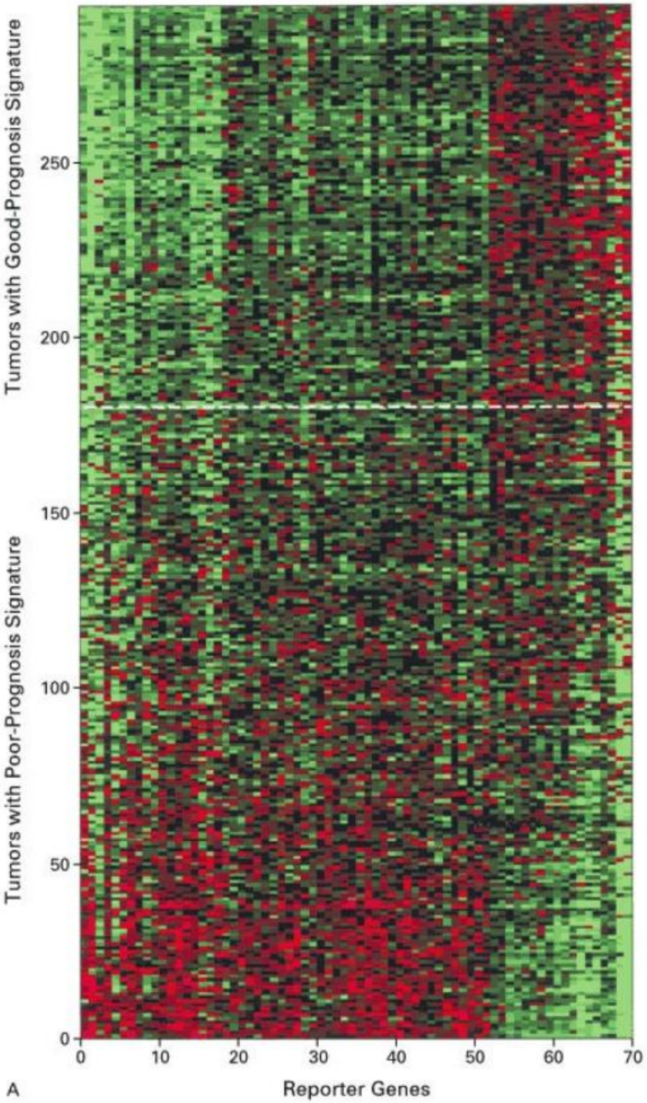
- the model is too complex
- it will overfit to the noise in our training set
(it learns the noise as well as the true problem structure)
- the model is very sensitive to the particular training sample

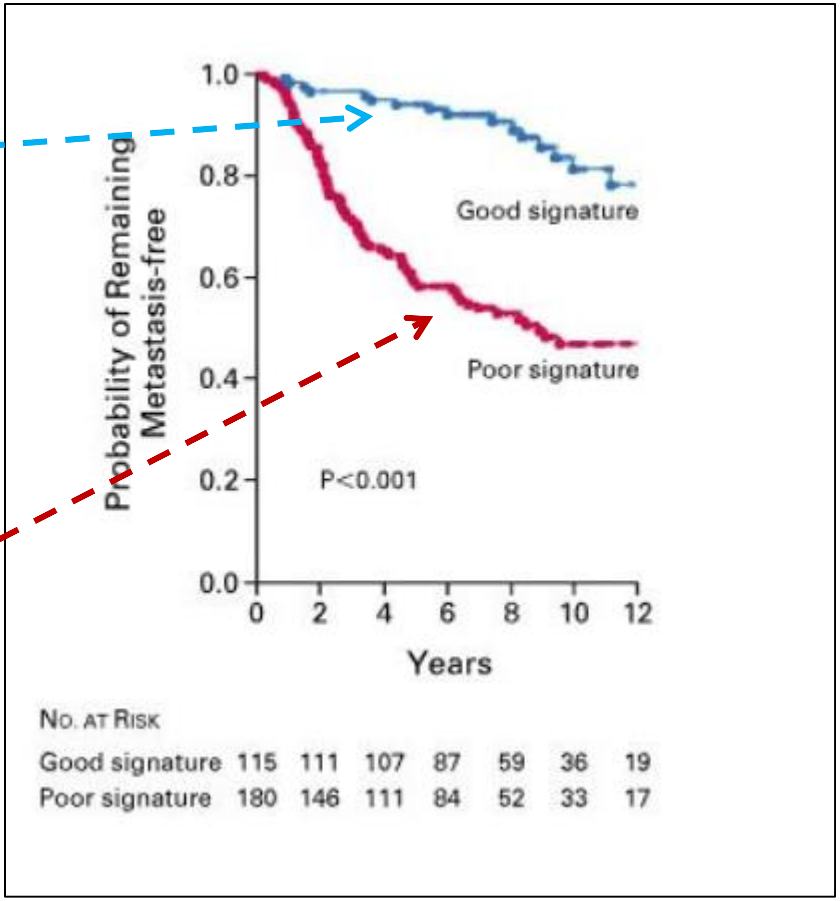
Underfitting or model bias:

- the model is not complex enough
- it cannot capture the true structure in our data
- no matter how much data we train the error is high

Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, 2001

70-gene signature is predictive of survival and risk of metastases [Van de Vijver et al, N Engl J Med, 2002]

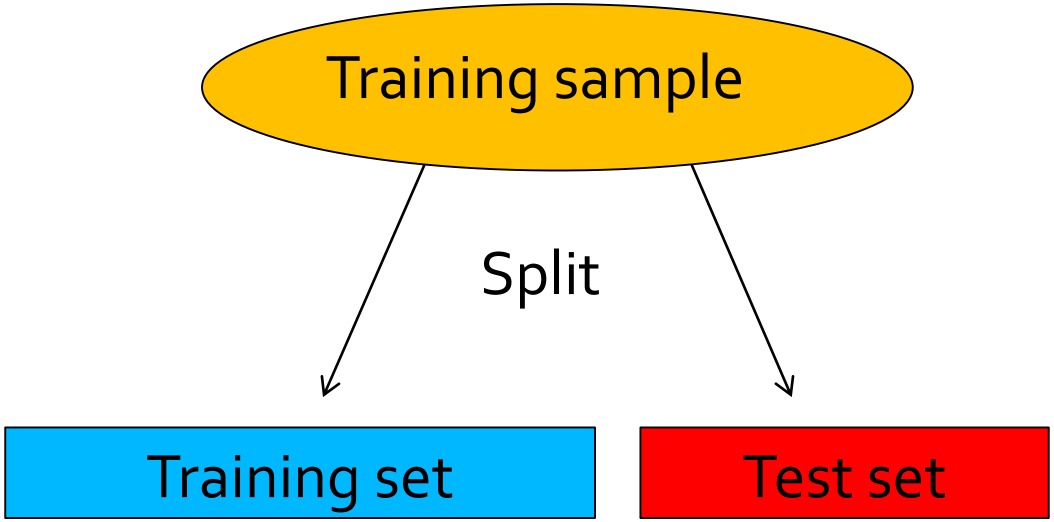




No. at Risk

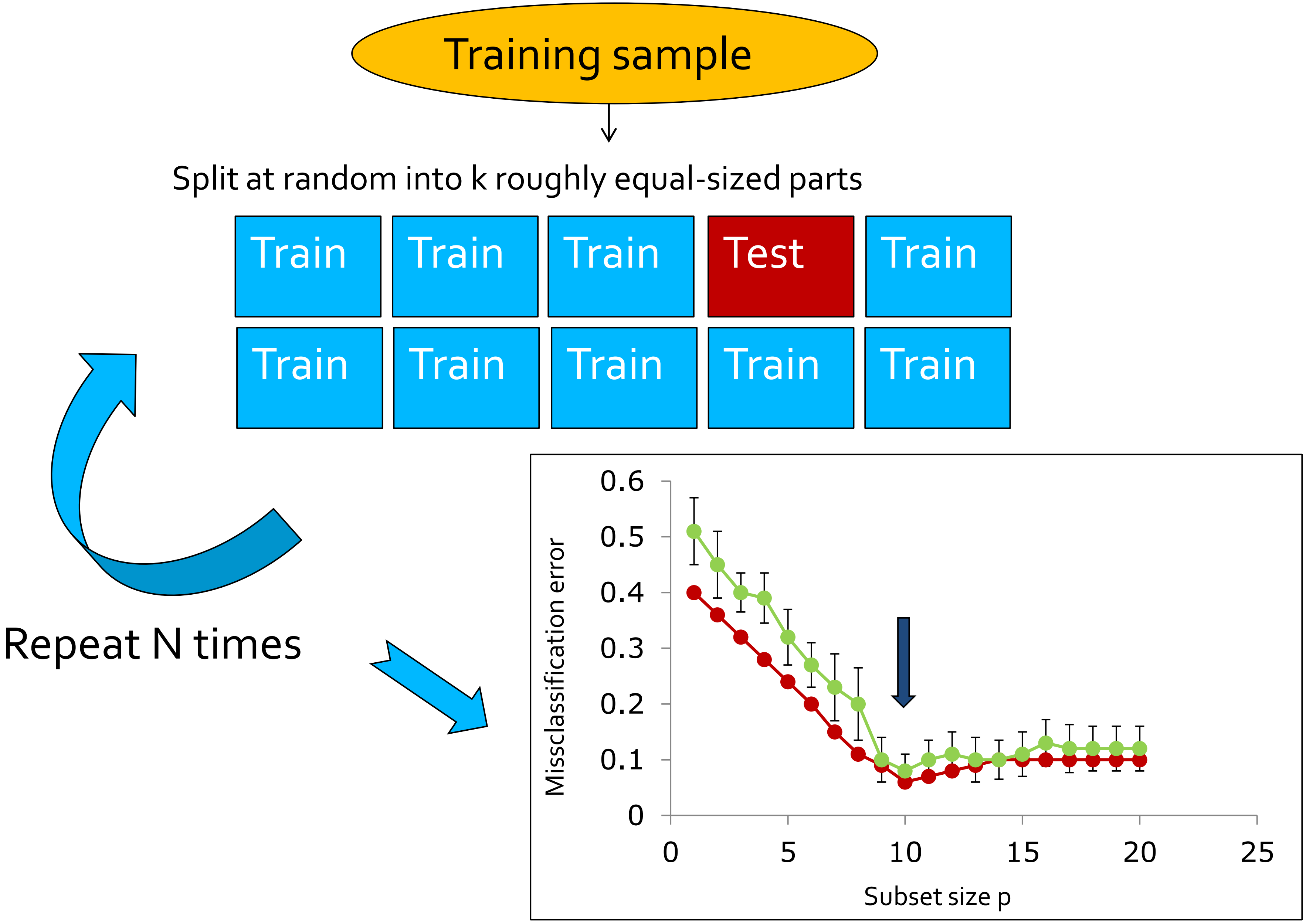
Good signature	115	111	107	87	59	36	19
Poor signature	180	146	111	84	52	33	17

Testing the model on an independent set



- NOT ALWAYS IDEAL
- Dataset might be too small
- Single train-test experiment: the estimate of the error could be misleading

Cross-validation (k-fold)



Training sample

Split at random into k roughly equal-sized parts

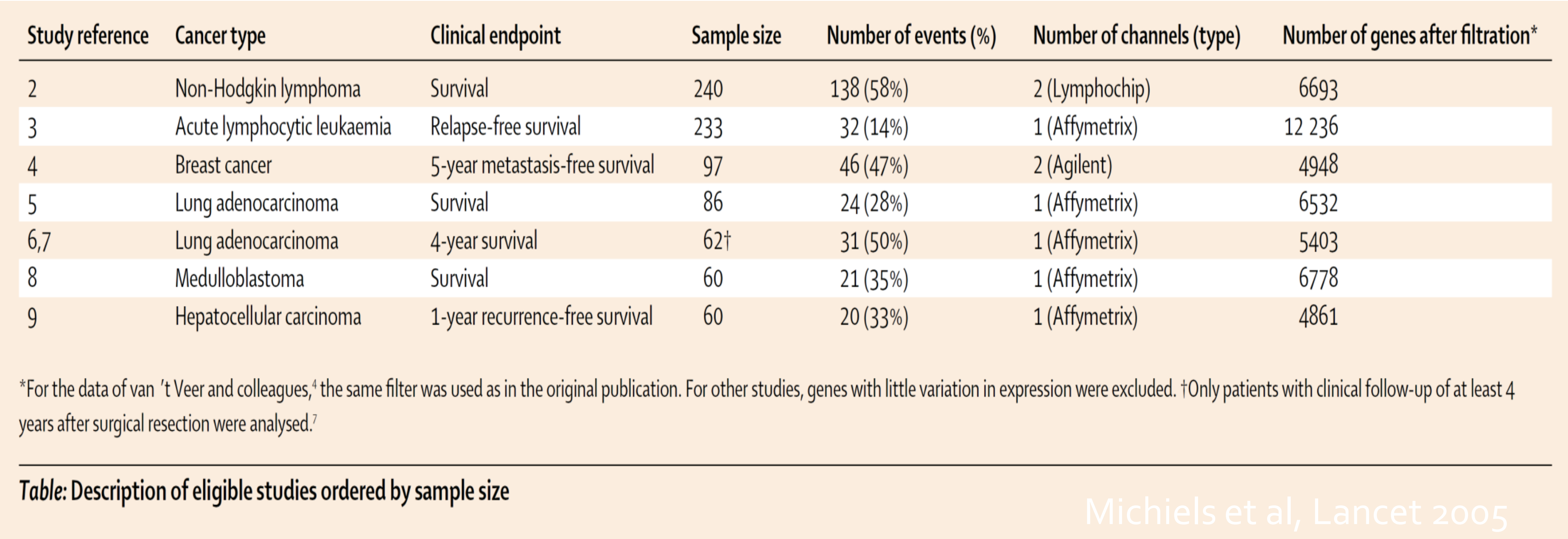
Train	Train	Train	Test	Train
Train	Train	Train	Train	Train

Repeat N times

Initial validation based on resampling helps selecting stronger biomarkers for prospective studies

Resampling strategy:

1) identify a GES in a training set of patients
2) estimate the proportion of misclassifications on an independent validation set
3) iterate on multiple random sets to study stability and performance of the GES



Study reference	Cancer type	Clinical endpoint	Sample size	Number of events (%)	Number of channels (type)	Number of genes after filtration*
2	Non-Hodgkin lymphoma	Survival	240	138 (58%)	2 (Lymphochip)	6693
3	Acute lymphocytic leukaemia	Relapse-free survival	233	32 (14%)	1 (Affymetrix)	12 236
4	Breast cancer	5-year metastasis-free survival	97	46 (47%)	2 (Agilent)	4948
5	Lung adenocarcinoma	Survival	86	24 (28%)	1 (Affymetrix)	6532
6,7	Lung adenocarcinoma	4-year survival	62†	31 (50%)	1 (Affymetrix)	5403
8	Medulloblastoma	Survival	60	21 (35%)	1 (Affymetrix)	6778
9	Hepatocellular carcinoma	1-year recurrence-free survival	60	20 (33%)	1 (Affymetrix)	4861

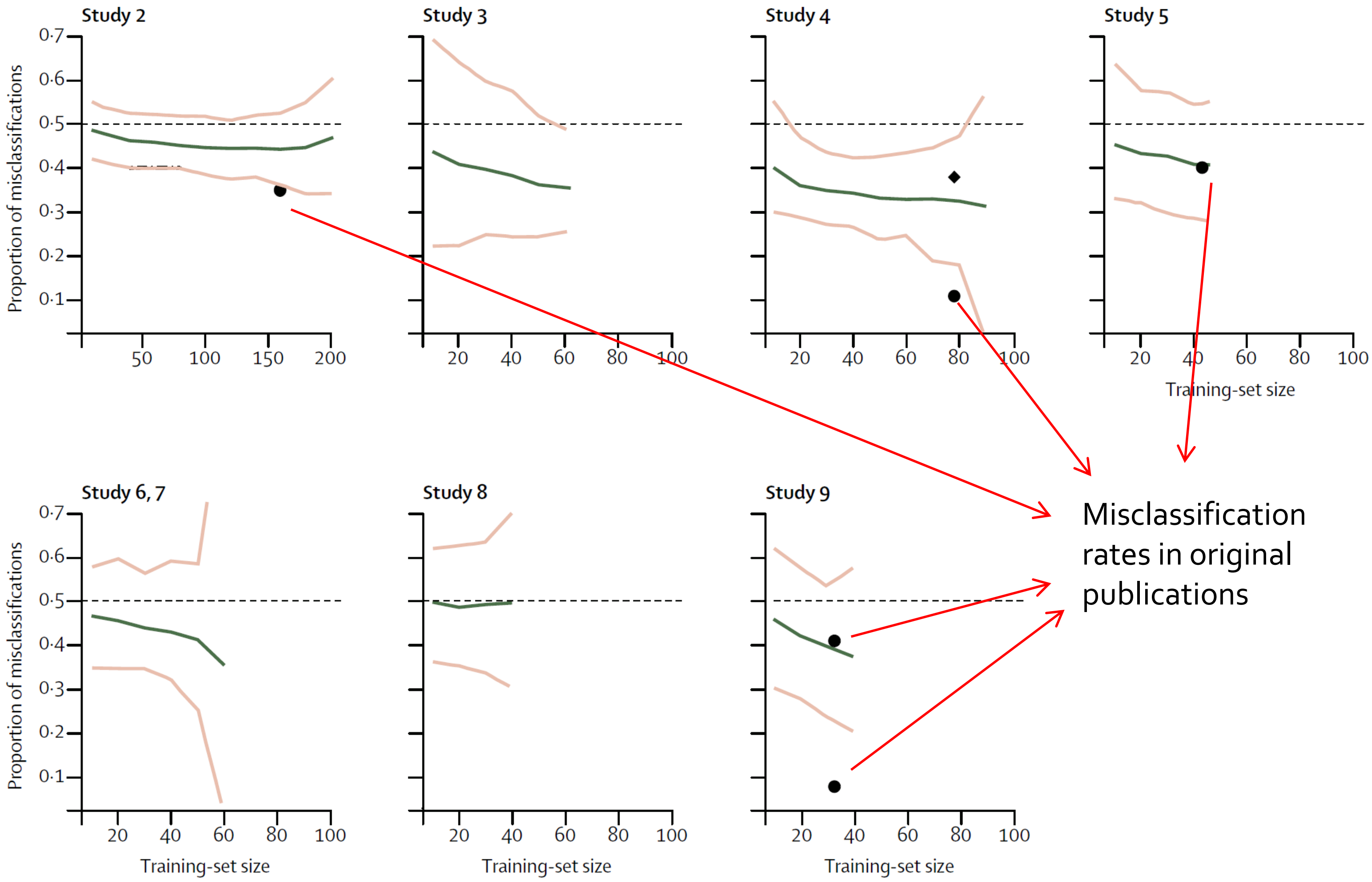
*For the data of van 't Veer and colleagues,[4] the same filter was used as in the original publication. For other studies, genes with little variation in expression were excluded. †Only patients with clinical follow-up of at least 4 years after surgical resection were analysed.[7]

Table: **Description of eligible studies ordered by sample size**

Michiels et al, Lancet 2005

Gene Expression Signatures misclassification

Misclassification rate from 500 random training-validation sets vs. training-set size (mean and 95% Cis)

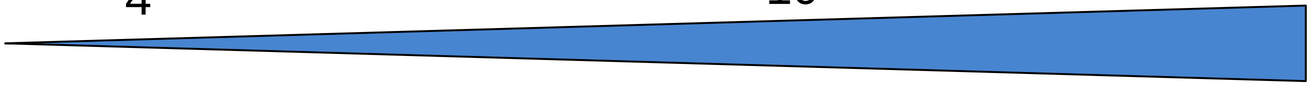


Misclassification rates in original publications

How many K folds in cross-validation?

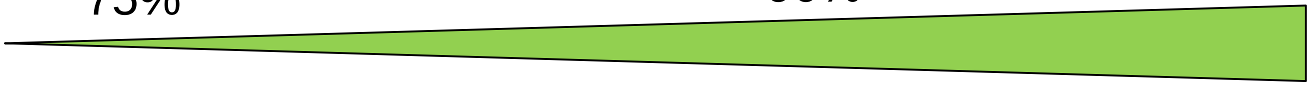
Number of folds, K

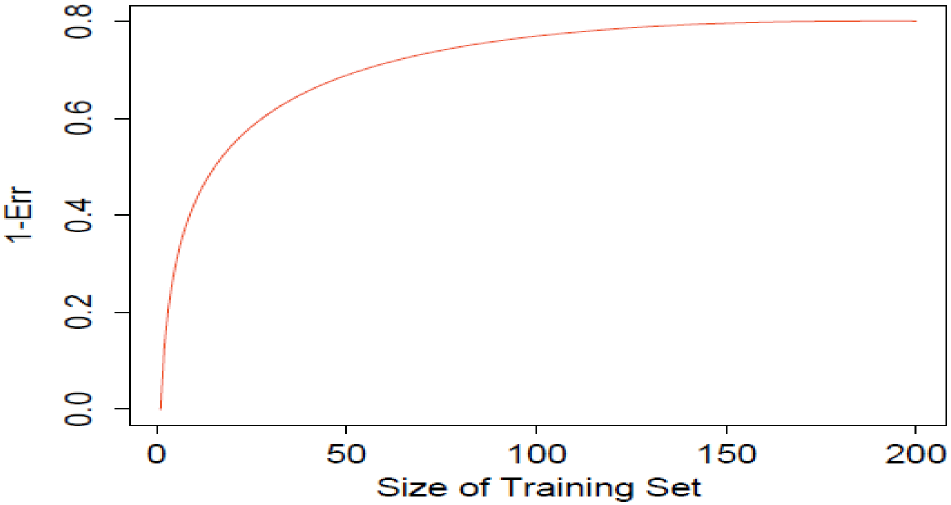
4 10 N (leave-one-out)



Size of training set

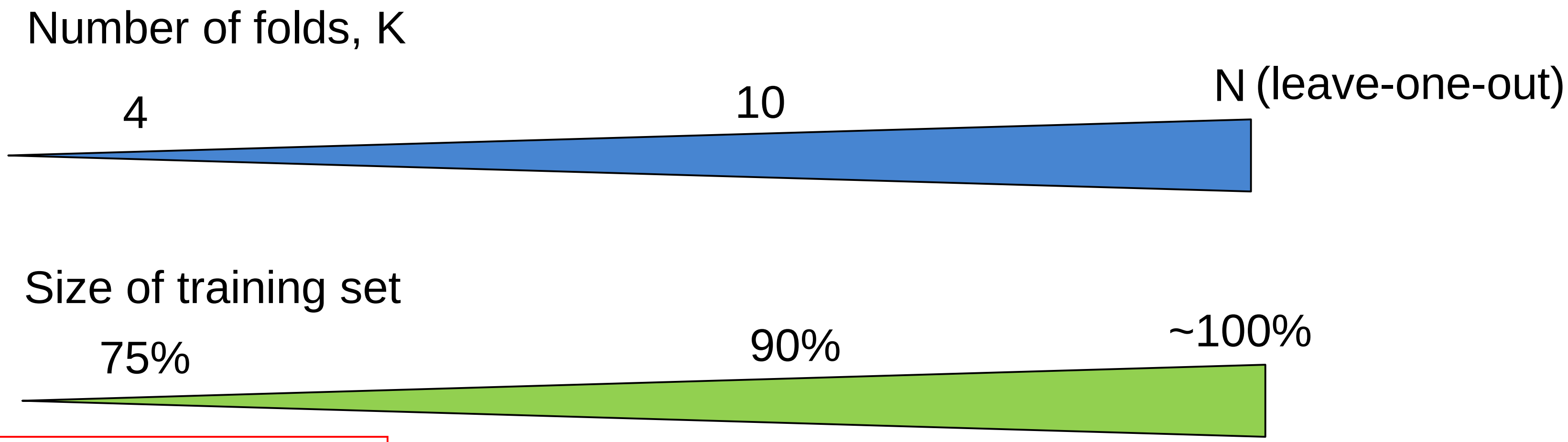
75% 90% ~100%



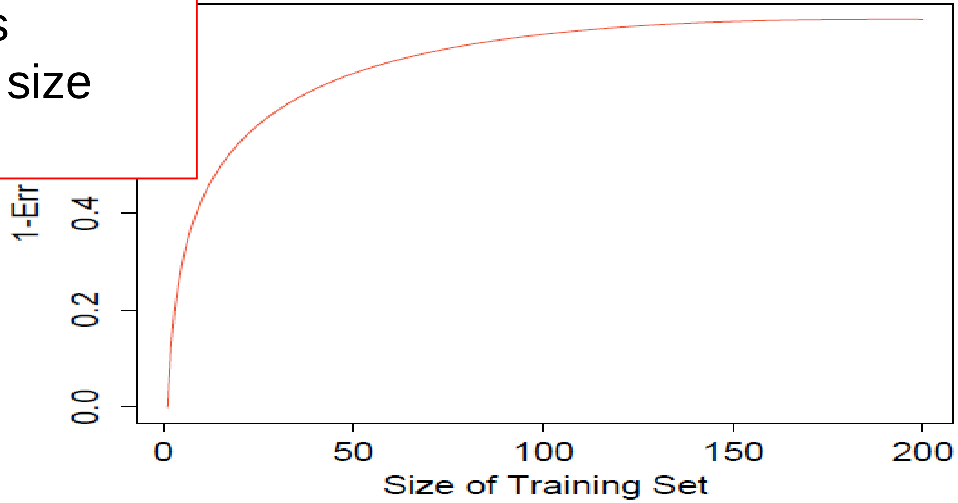


Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, 2001

How many K folds in cross-validation?

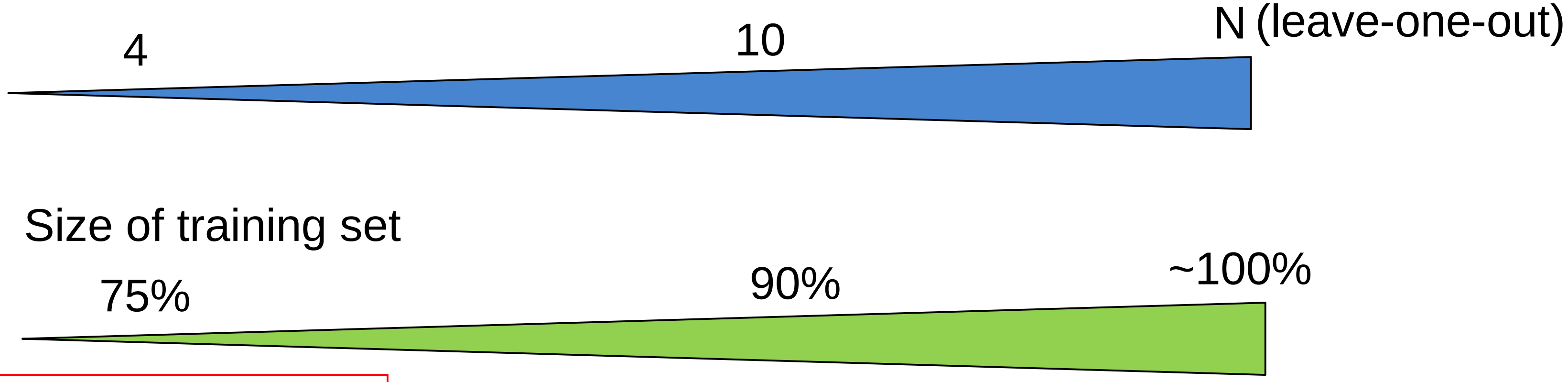


The performance of the predictor improves as the training set size increases



Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, 2001

How many K folds in cross-validation?



Number of folds, K

4

10

N (leave-one-out)

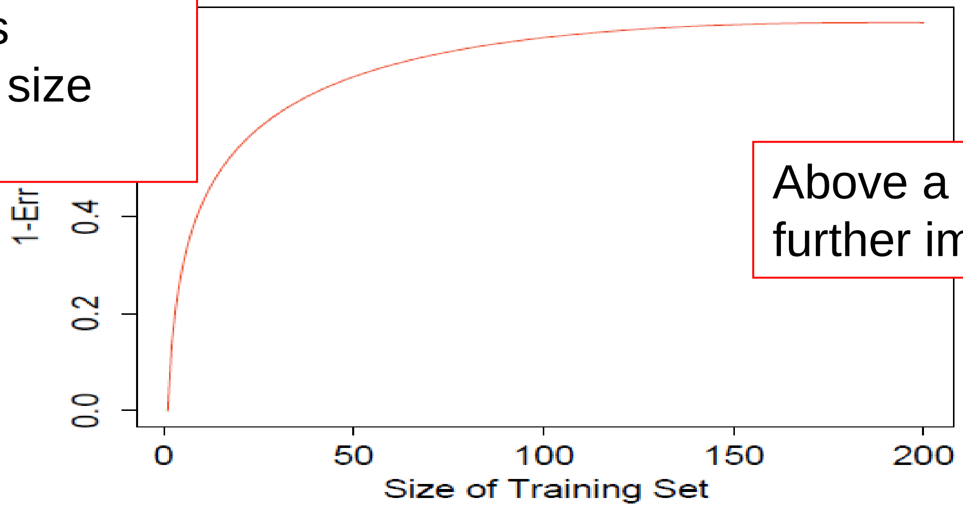
Size of training set

75%

90%

~100%

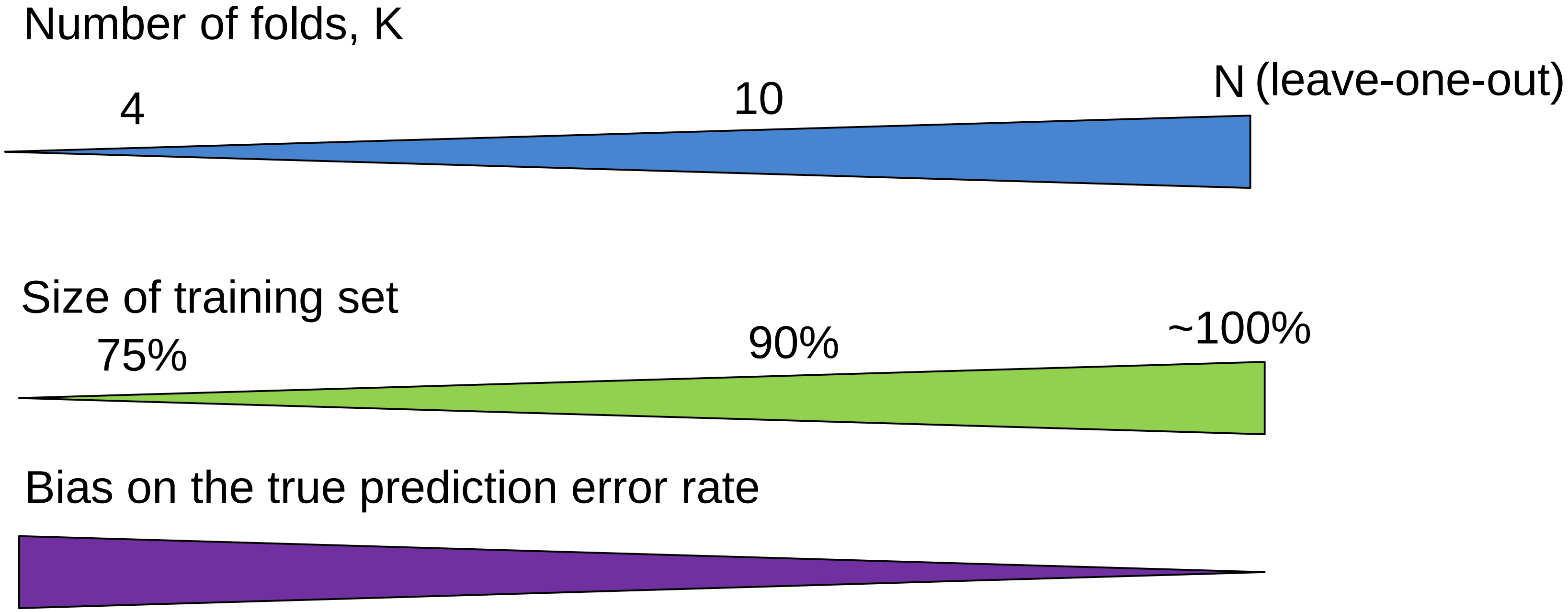
The performance of the predictor improves
as the training set size increases



Above a certain size no further improvement

Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, 2001

How many K folds in cross-validation?



Number of folds, K

4

10

N (leave-one-out)

Size of training set

75%

90%

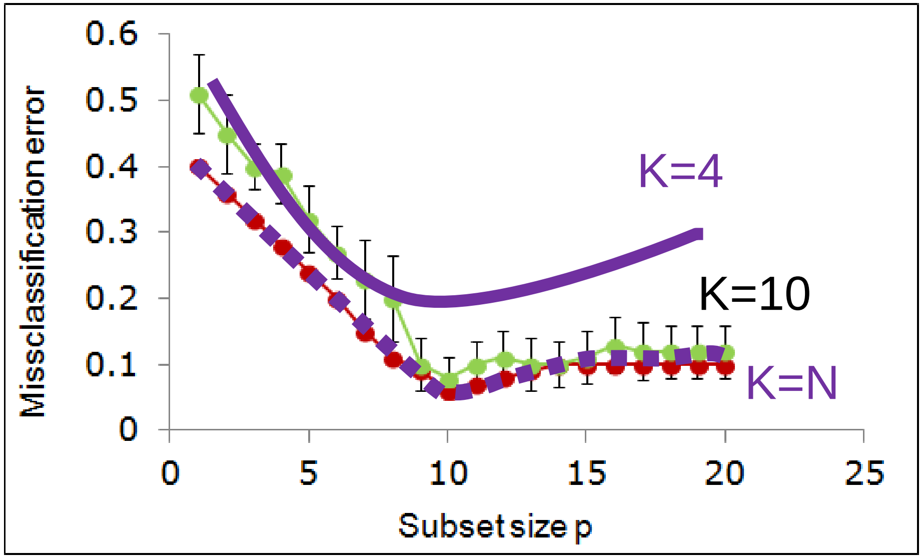
~100%

Bias on the true prediction error rate

➢ The training set is much smaller than the original set

➢ True error rate can be overestimated

➢ Bias upward (conservative)



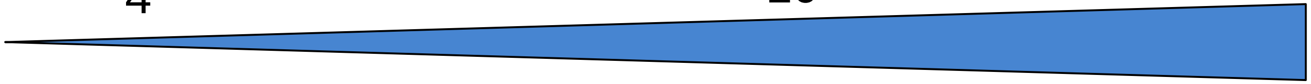
How many K folds in cross-validation?

Number of folds, K

4

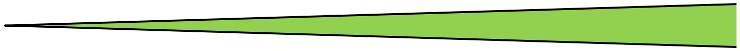
10

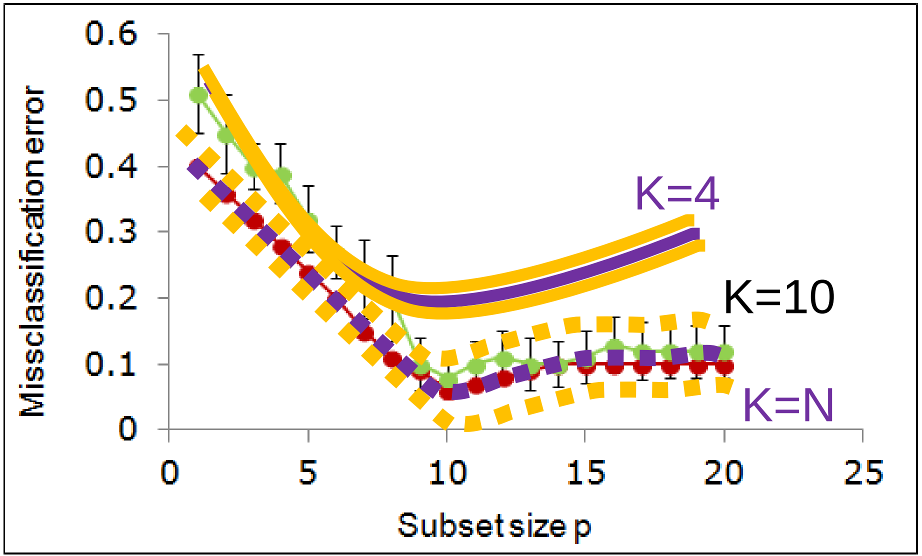
N (leave-one-out)



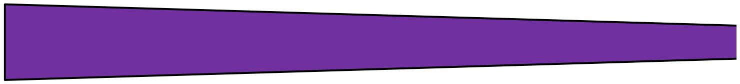
Size of training set

75%





Bias on the true prediction error ra

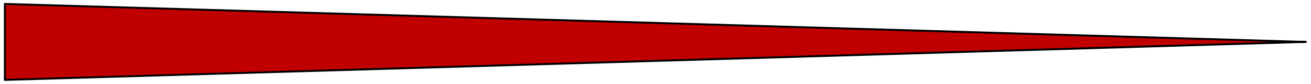


Size of testing set

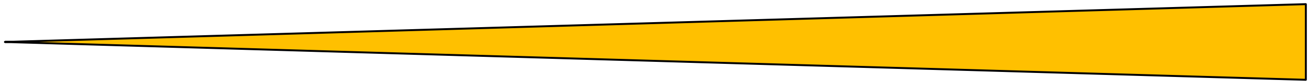
25%

10%

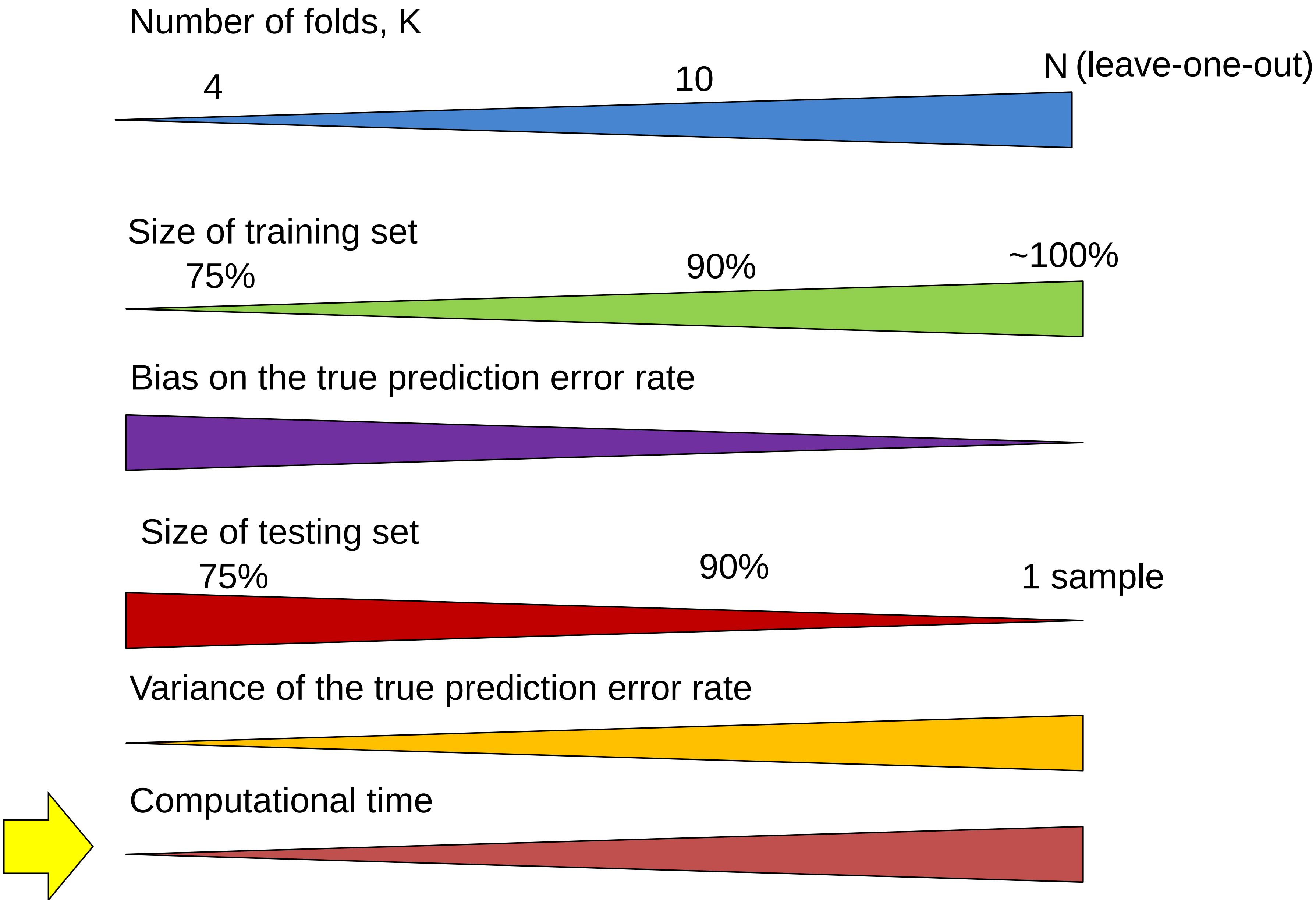
1 sample



Variance of the true prediction error rate



How many K folds in cross-validation?



Number of folds, K

4

10

N (leave-one-out)

Size of training set

75%

90%

~100%

Bias on the true prediction error rate

Size of testing set

75%

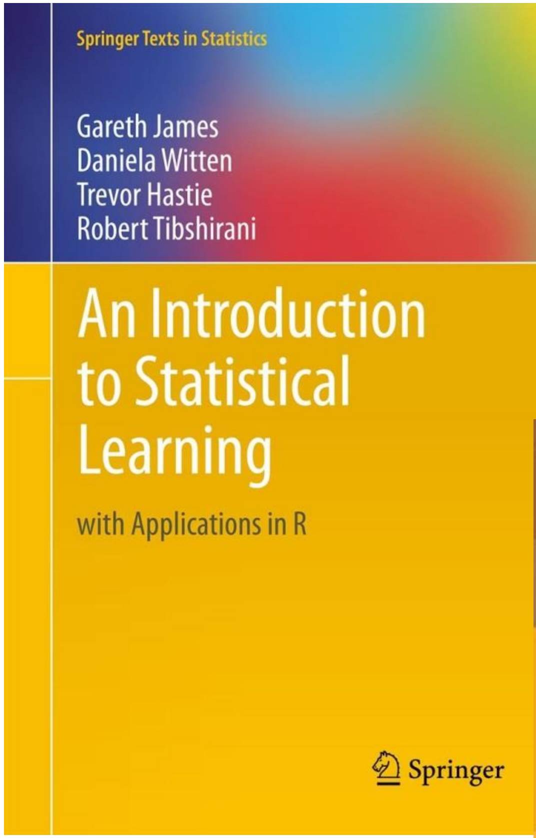
90%

1 sample

Variance of the true prediction error rate

Computational time

A couple of good books



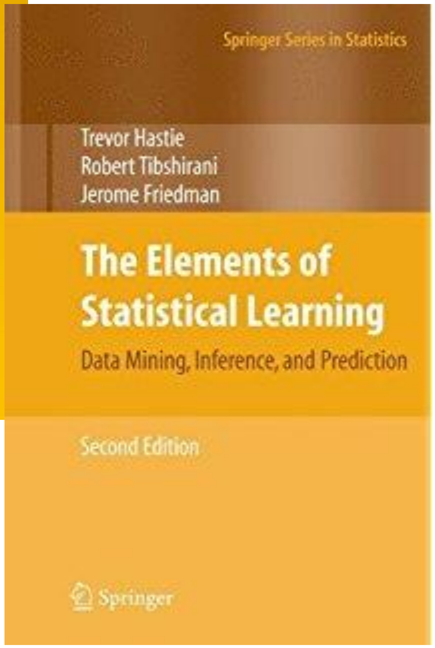
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer



Springer Series in Statistics

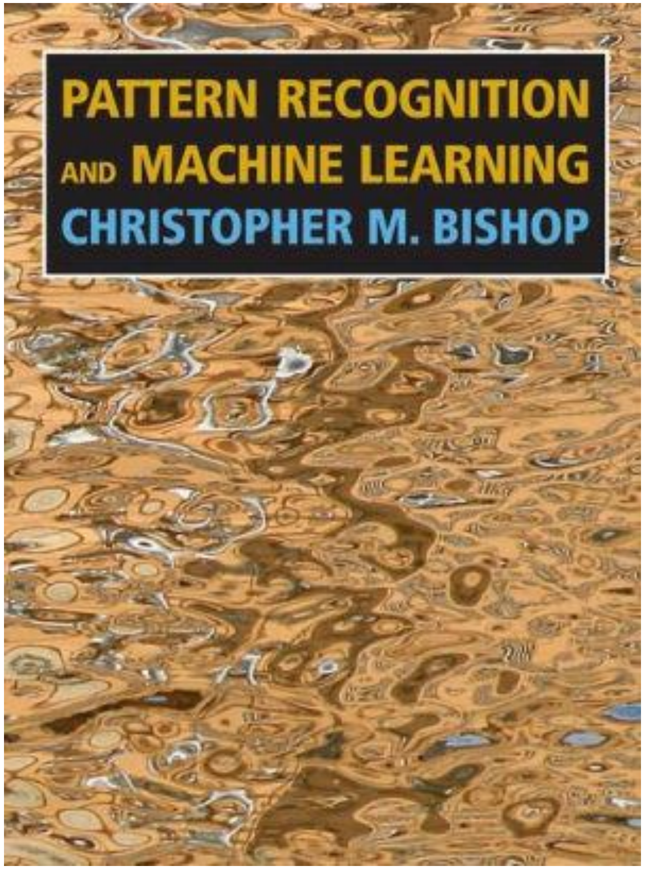
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer



PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP