# *Model Assessment and Selection*

**Rainer Spang**

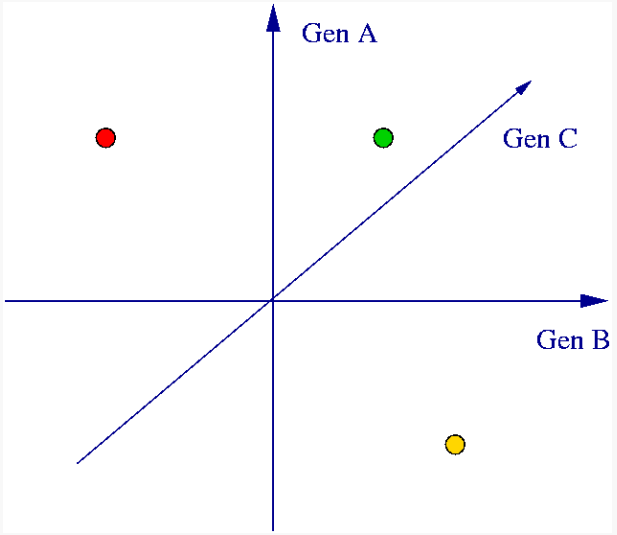
Courses in Practical DNA Microarray Analysis

NGFN
Nationales Genomforschungsnetz

# Which model is best ?

**Experience: Linear models work fine**

**Sparse data:**
**Expression data in high dimensions is sparse. The data does not contain information to identify non linear structures adequately, even if they exist.**



# Which type of regularization is best?

**Experience:** The are all the same, except for some stupid ideas

**Theoretical consolidation:** The challenge is more to unravel the theoretical relationships between the methods

**The important question is not which regularization but how much of it**

# Adaptive model selection

**Choose a family of models with varying regularization strength**

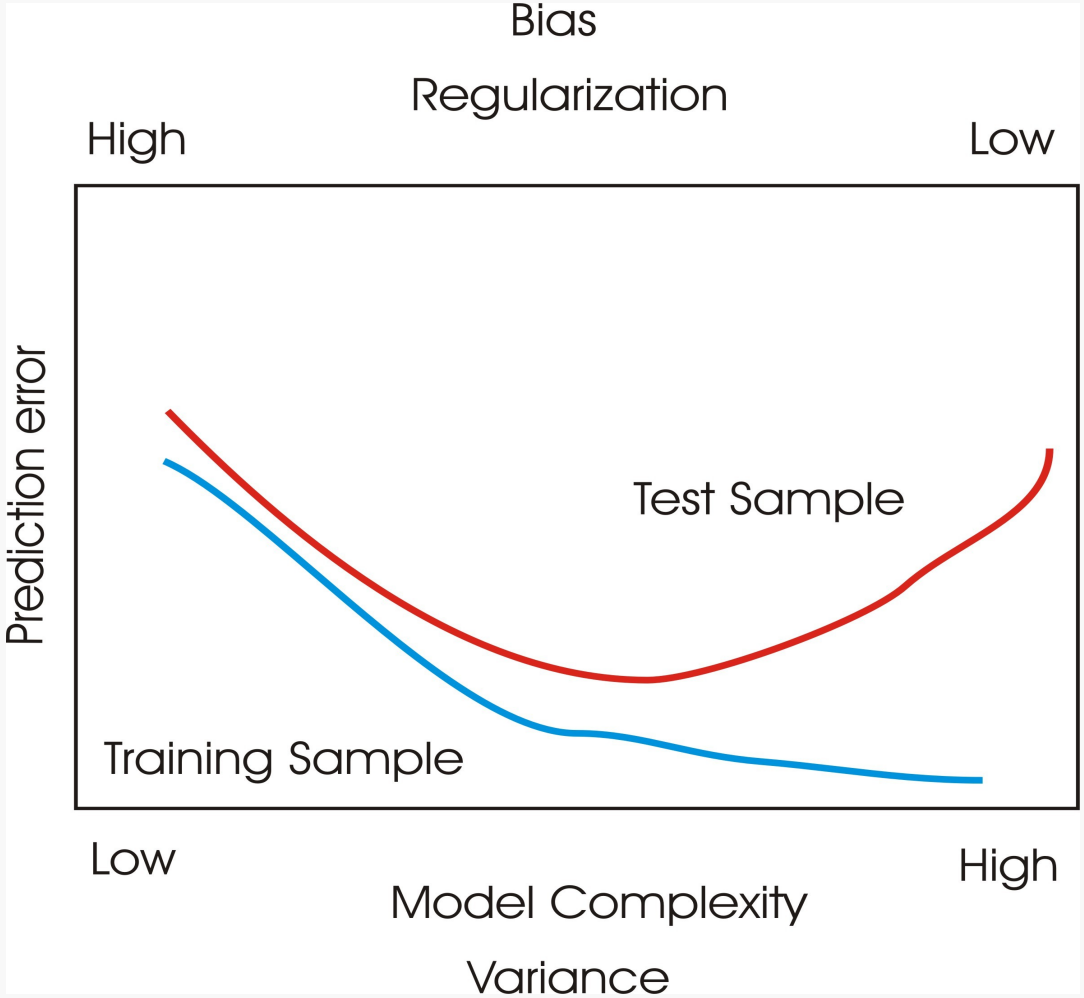
**- tune the number of genes**

**- use a parameter calibrating likelihood and penalty**

**Use cross validation on the training data to optimize regularization strength**

**This can be very data dependent!**

# The bias variance trade off



# *How much shrinkage is good in PAM ?*

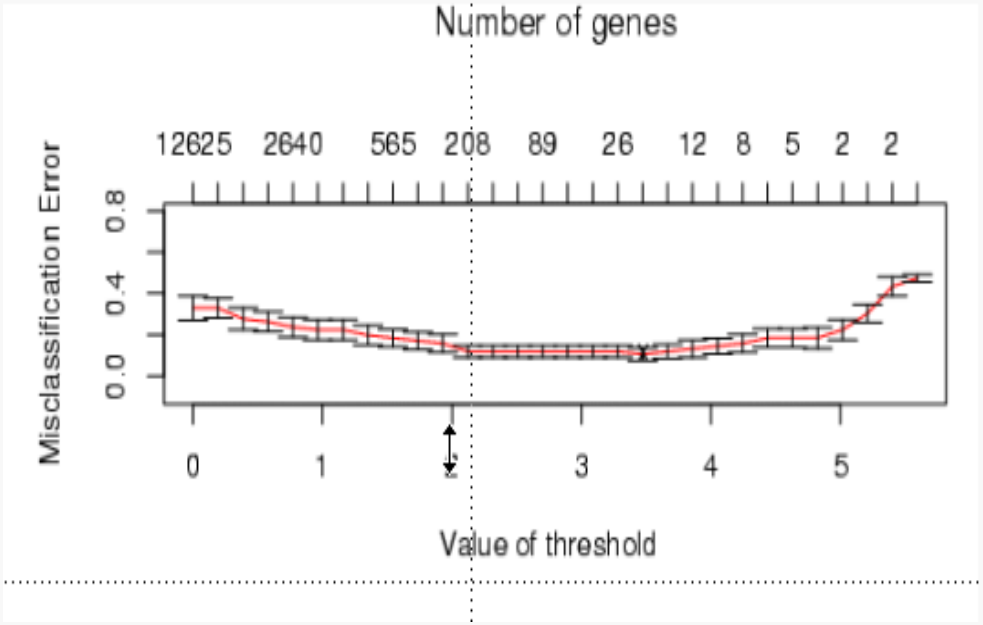

**Compute the CV-Performance for several values of $\Delta$**

**Pick the $\Delta$ that gives you the smallest number of CV-Misclassifications**

***Adaptive Model Selection***

**PAM does this routinely**

# Model Selection Output of PAM



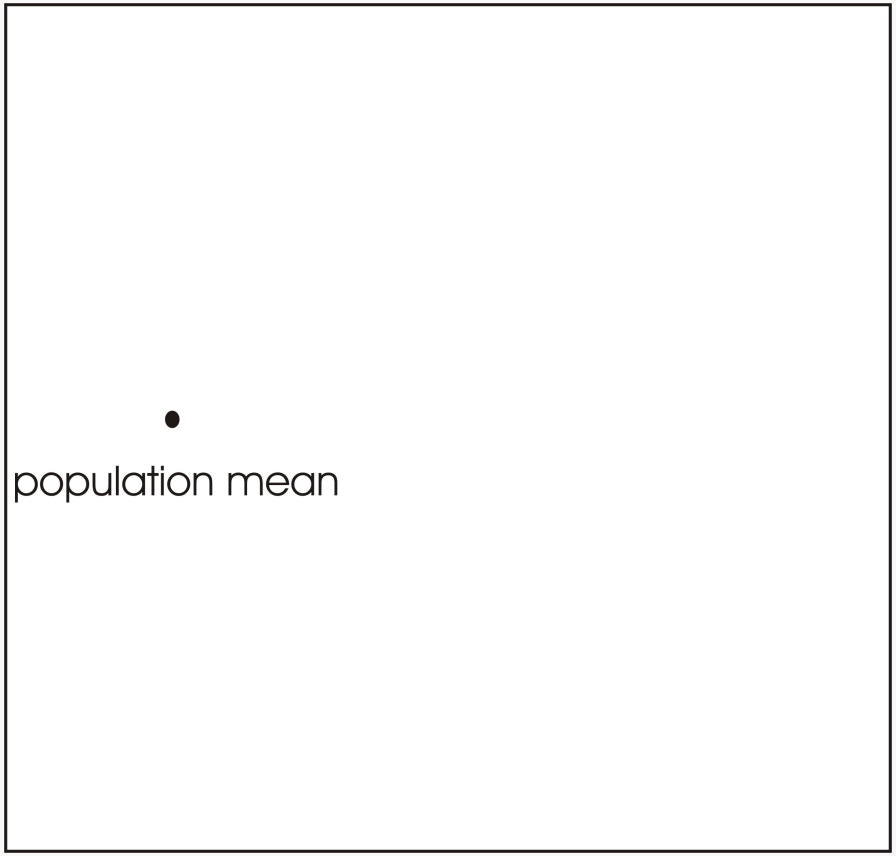

**Small Δ, many genes poor performance due to overfitting**

**High Δ, few genes, poor performance due to lack of information – *underfitting* -**

**The optimal Δ is somewhere in the middle**

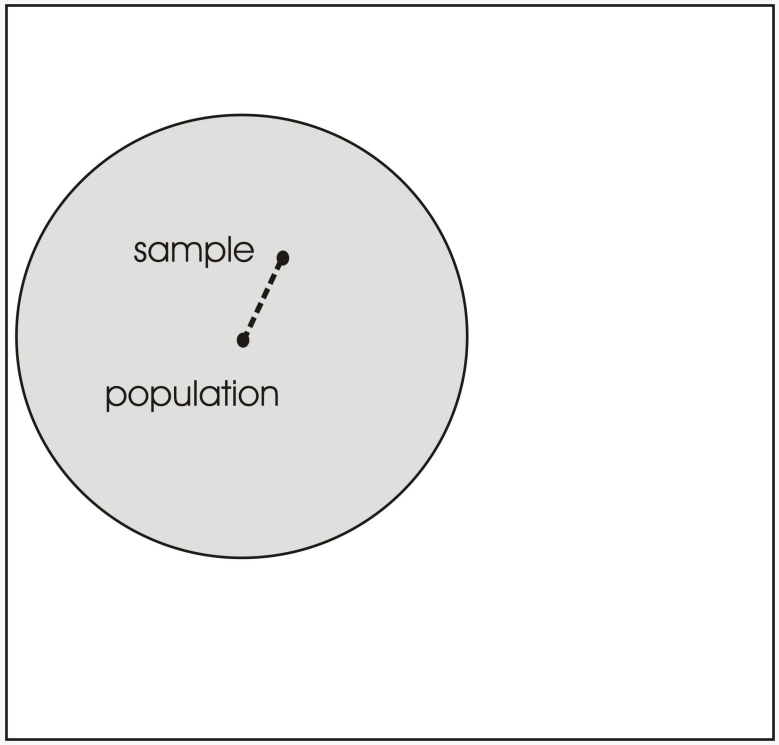
## *Population mean:*

**Genes have a certain mean expression and correlation in the population**

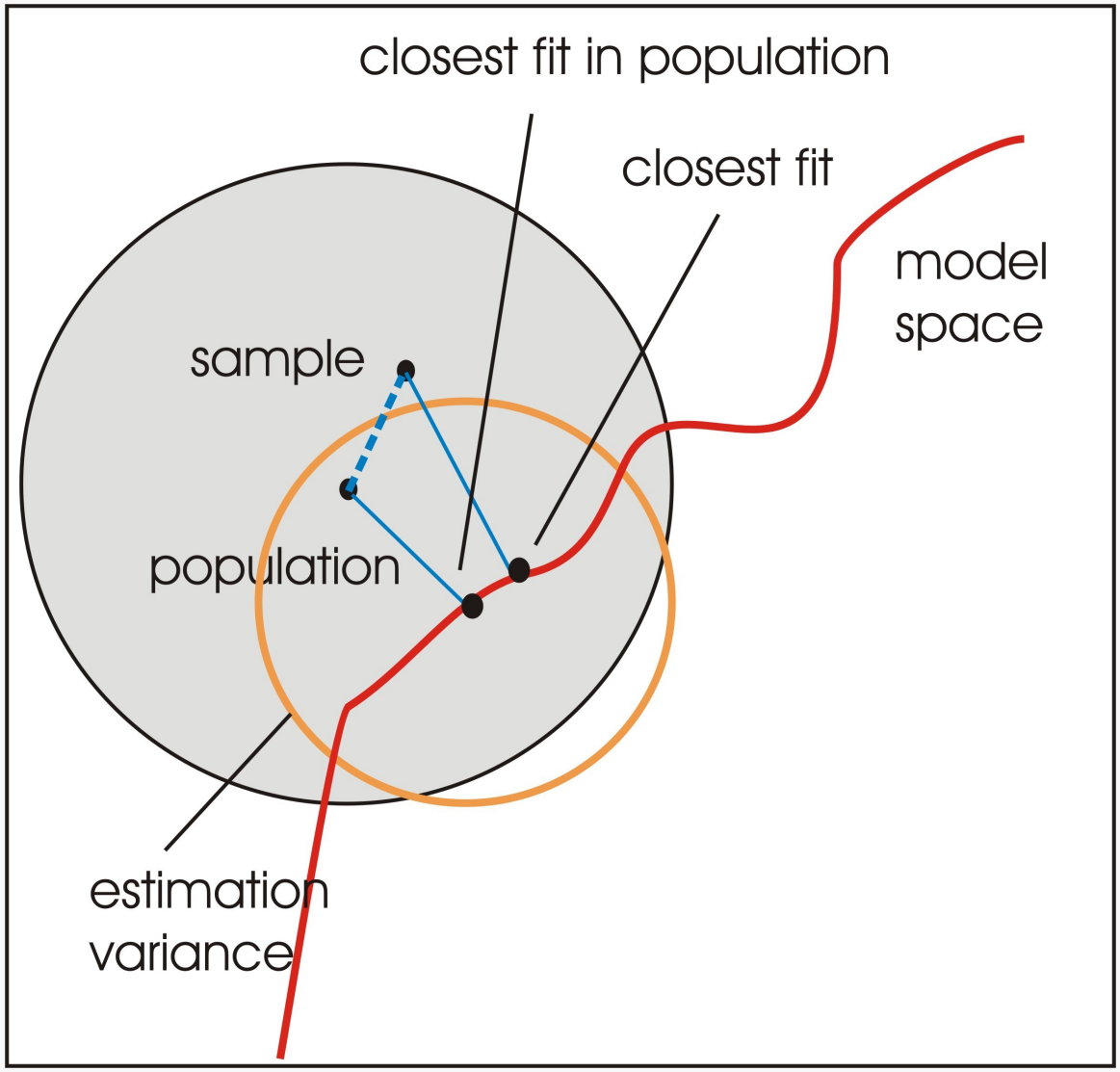


## *Sample mean:*

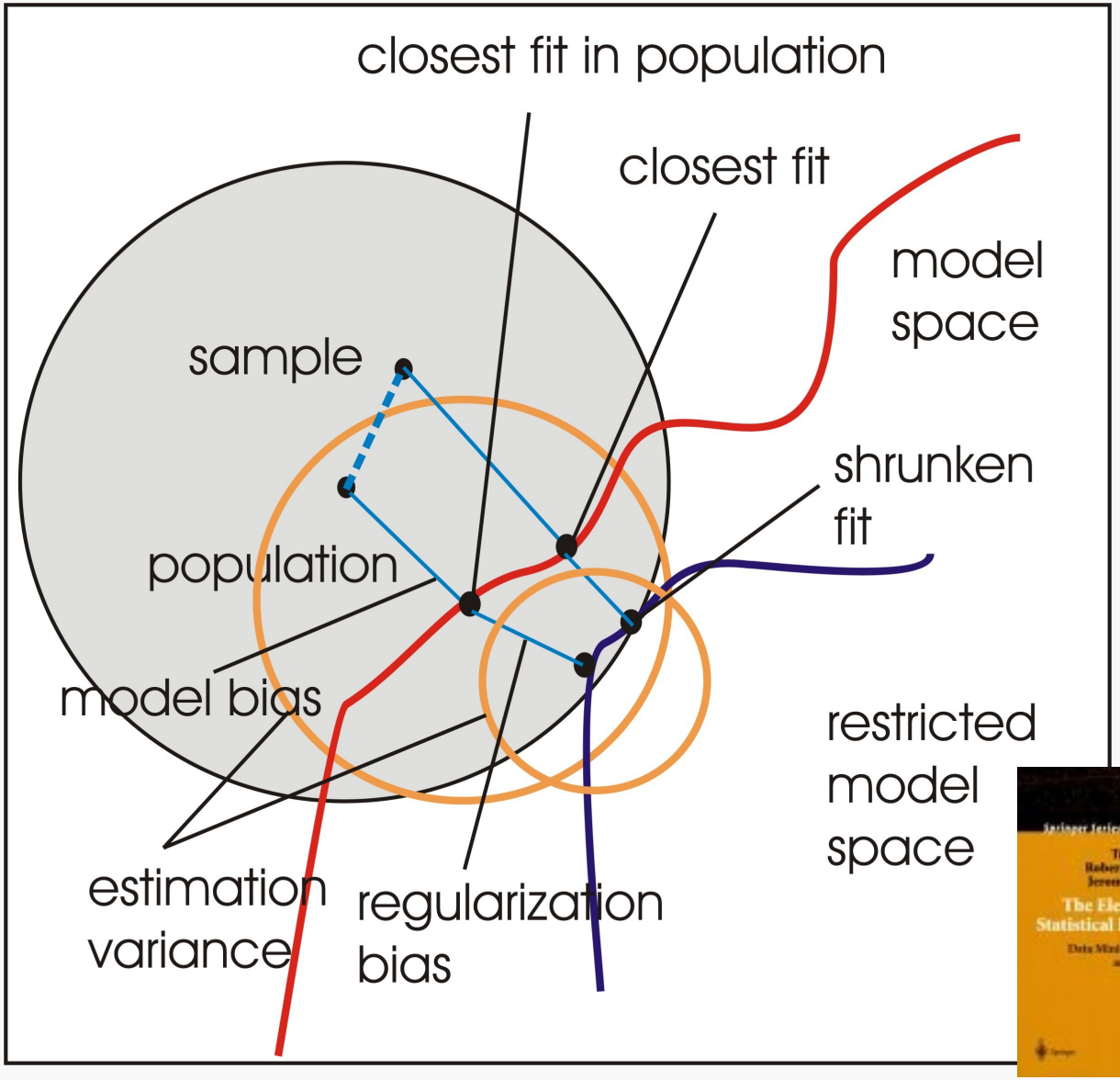
# We observe average expression and empirical correlation

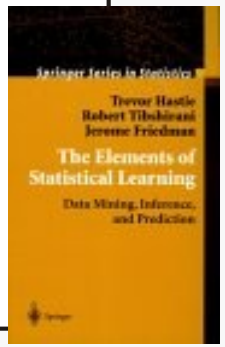


# Fitted model:



# Regularization



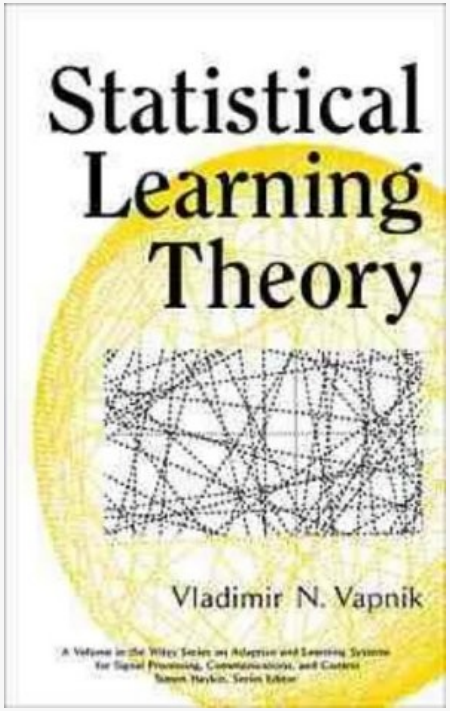



# *Adaptive Model Selection of SVM*

**SVM optimize the margin of separation**

**There are theoretical results connecting the margin to an upper bound of the test error (V. Vapnik)**

***- structural risk minimization -***



# Validation

**The accuracy of a signature on the data it was learned from is biased**

**Validation of a signature requires independent test data**

**This test data must not be used for gene selection or adaptive model selection, otherwise the observed accuracy is biased**

**◊ Selection bias**

# Cross Validation


**You can not evaluate a fitted classification model ( = signature ) using cross validation**

**Cross validation only evaluates the algorithm with which the signature was build**

**Gene selection must be repeated for every relearning step in the cross validation**

◊ **In the loop gene selection**

# *Leave one out Cross-Validation*


**Essentially the same**

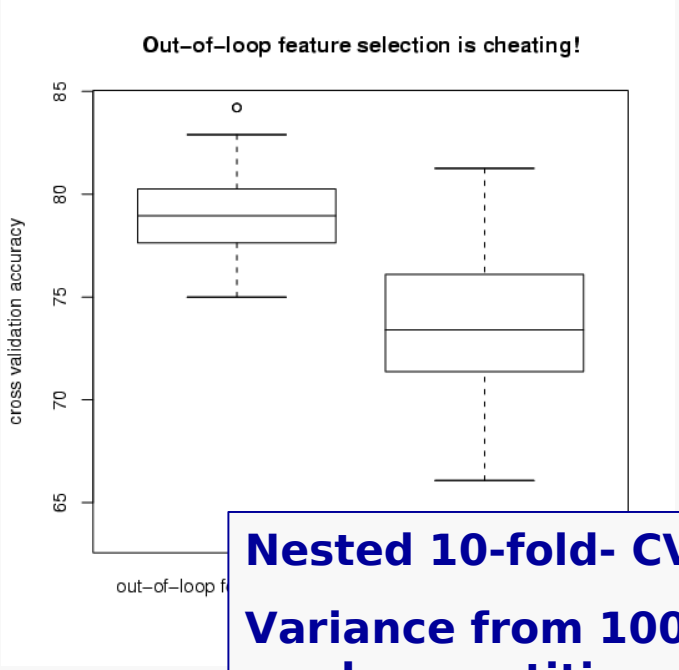
**But you only leave one sample out at a time and predict it using the others**

**Good for small training sets**

# Performance Estimation

**Estimators of performance have a variance …**

**… which can be high. The chances of a meaningless signature to produce 100% accuracy on test data is high if the test data includes only few patients**




**Nested 10-fold- CV**

**Variance from 100 random partitions**

**→ Reuse of the same data … no sample variance**

# *DOs AND DONTs :*

**1. Decide on your diagnosis model (PAM,SVM,etc...) and don‘t change your mind later on**

**2. Split your profiles randomly into a training set and a test set**

**3. Put the data in the test set away ... far away**

**4. Train your model only using the data in the training set**

**(select genes, define centroids, calculate normal vectors for large margin separators, perform adaptive model selection ...)**

**don‘t even think of touching the test data at this time**

**5. Apply the model to the test data ...**

**don‘t even think of changing the model at this time**

**6. Do steps 1-5 only once and accept the result**

# External Validation and Documentation

**Documenting a signature is conceptually different from giving a list of genes, although is is what most publications give you**

**In order to validate a signature on external data or apply it in practice:**

**- All model parameters need to be specified**

**- The scale of the normalized data to which the model refers needs to be specified**

# Establishing a signature

# Thank you