

Classification Models for Molecular Diagnosis

Course in Practical DNA Microarray Analysis

Holger Fröhlich
Div. Molecular Genome Analysis

Basic Research:

Which role plays gene A in disease B ?

Clinical Routine:

Which consequence has expression status X of gene A for patient Y ?

Personalized Medicine

Yesterday the focus was on basic research questions

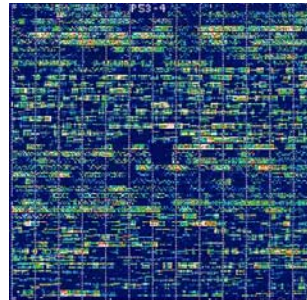
We have investigated genes

- Differentially expressed genes
- Coexpressed genes (clustering)

Today it will be on patients

- Molecular diagnosis
- Predicting survival / therapy response

DNA Chip of Ms. Smith



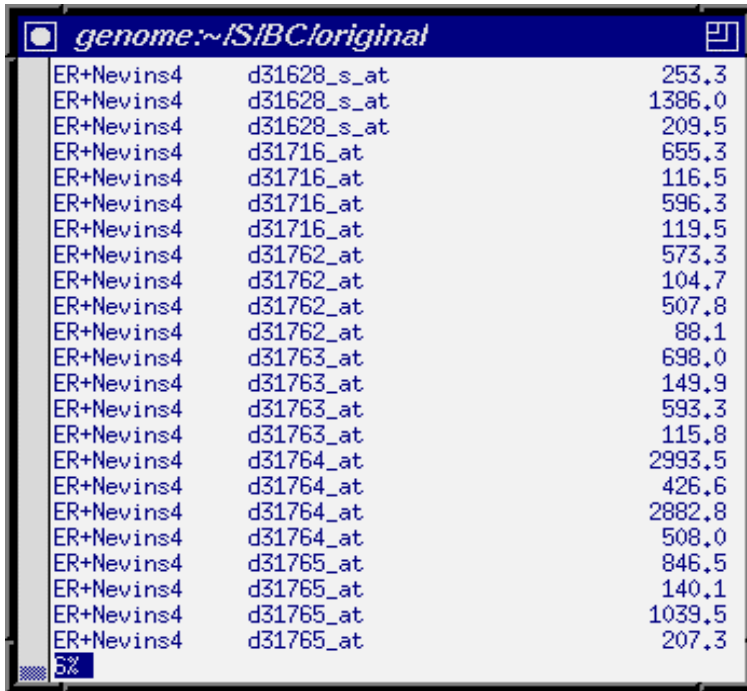
Ms. Smith

```
genome:~/ISIBC/original
```

ER+Nevins4	d31628_s_at	253.3
ER+Nevins4	d31628_s_at	1386.0
ER+Nevins4	d31628_s_at	209.5
ER+Nevins4	d31716_at	655.3
ER+Nevins4	d31716_at	116.5
ER+Nevins4	d31716_at	596.3
ER+Nevins4	d31716_at	119.5
ER+Nevins4	d31762_at	573.3
ER+Nevins4	d31762_at	104.7
ER+Nevins4	d31762_at	507.8
ER+Nevins4	d31762_at	88.1
ER+Nevins4	d31763_at	698.0
ER+Nevins4	d31763_at	149.9
ER+Nevins4	d31763_at	593.3
ER+Nevins4	d31763_at	115.8
ER+Nevins4	d31764_at	2993.5
ER+Nevins4	d31764_at	426.6
ER+Nevins4	d31764_at	2882.8
ER+Nevins4	d31764_at	508.0
ER+Nevins4	d31765_at	846.5
ER+Nevins4	d31765_at	140.1
ER+Nevins4	d31765_at	1039.5
ER+Nevins4	d31765_at	207.3

**Expression profile
of Ms. Smith**

The expression profile



```
genome:~/ISIBC/original
ER+Nevins4 d31628_s_at 253.3
ER+Nevins4 d31628_s_at 1386.0
ER+Nevins4 d31628_s_at 209.5
ER+Nevins4 d31716_at 655.3
ER+Nevins4 d31716_at 116.5
ER+Nevins4 d31716_at 596.3
ER+Nevins4 d31716_at 119.5
ER+Nevins4 d31762_at 573.3
ER+Nevins4 d31762_at 104.7
ER+Nevins4 d31762_at 507.8
ER+Nevins4 d31762_at 88.1
ER+Nevins4 d31763_at 698.0
ER+Nevins4 d31763_at 149.9
ER+Nevins4 d31763_at 593.3
ER+Nevins4 d31763_at 115.8
ER+Nevins4 d31764_at 2993.5
ER+Nevins4 d31764_at 426.6
ER+Nevins4 d31764_at 2882.8
ER+Nevins4 d31764_at 508.0
ER+Nevins4 d31765_at 846.5
ER+Nevins4 d31765_at 140.1
ER+Nevins4 d31765_at 1039.5
ER+Nevins4 d31765_at 207.3
5%
```

...

... a list of 30,000 numbers

... that are all properties of Ms. Smith


... some of them reflect her health problem (a tumor)

... the profile is a digital image of Ms. Smith's tumor

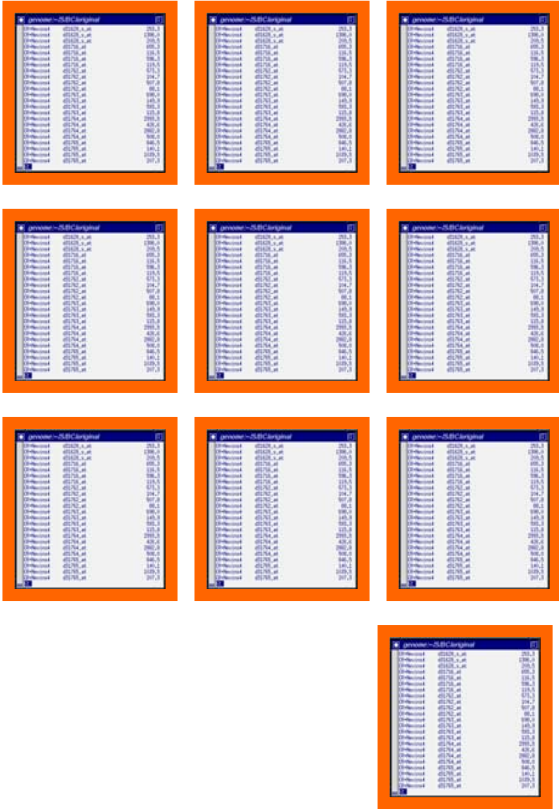
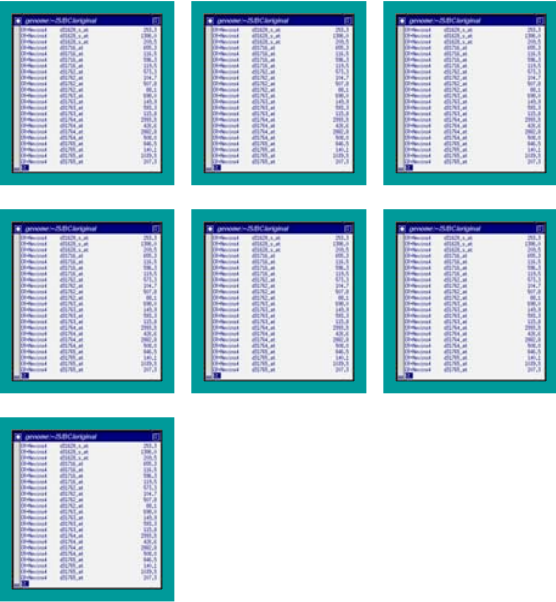
How can these numbers *tell us (predict)* whether Ms. Smith has tumor type **A** or tumor type **B** ?

Comparing her profile to profiles of people with tumor type A and to patients with tumor type B

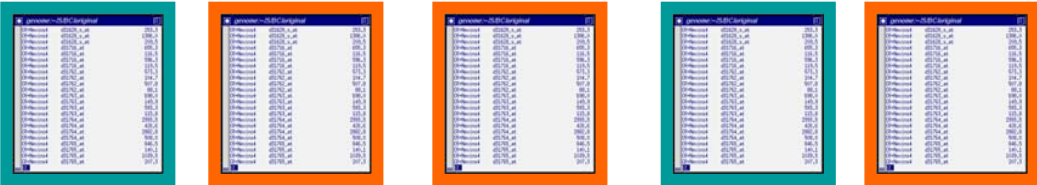
?



Ms. Smith

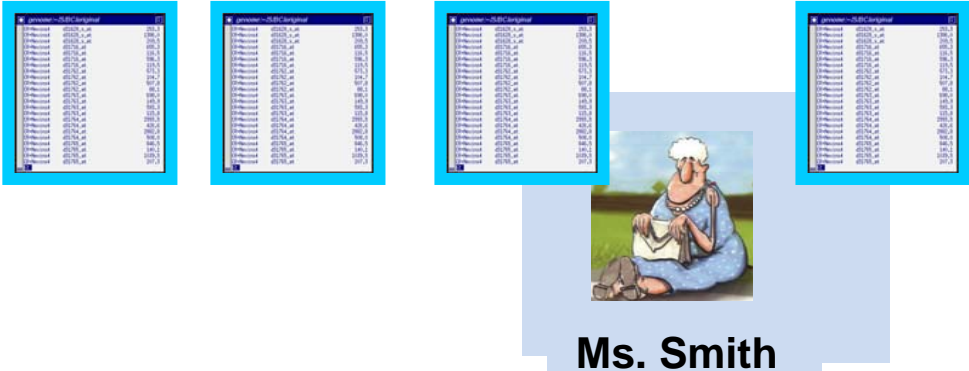


The setup for predictive data analysis



There are patients with known outcome

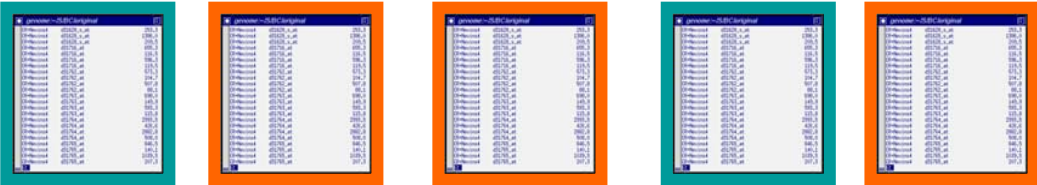
- *the trainings samples* -



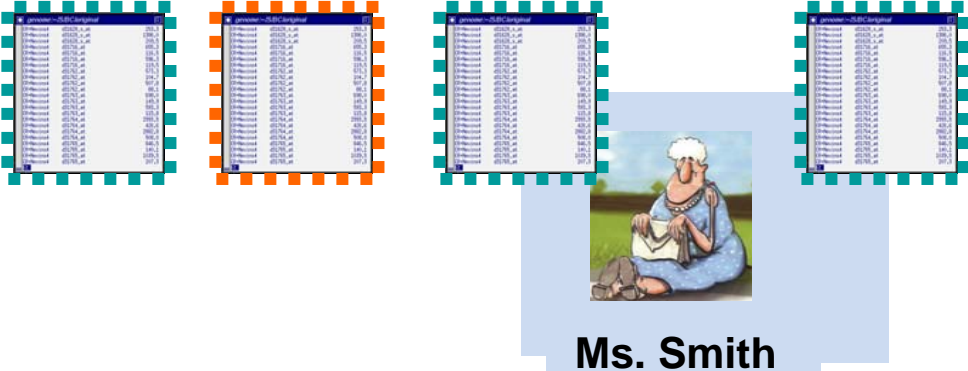
There are patients with unknown outcome

- *the „new“ samples* -

The challenge of predictive data analysis

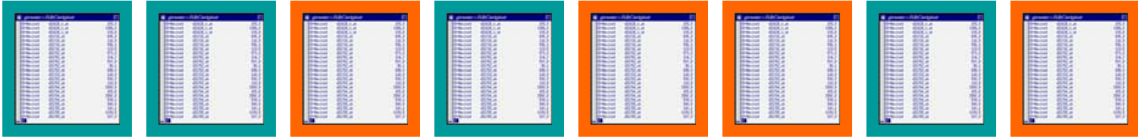


Use the trainings samples ...



... to learn how to predict „new“ samples

How can we find out whether we have really learned how to predict the outcome?



Take some patients from the original training samples and blind the outcome



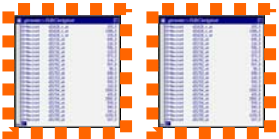
These are now called **test samples**



Only the remaining samples are still training samples. Use them to learn how to predict



ok

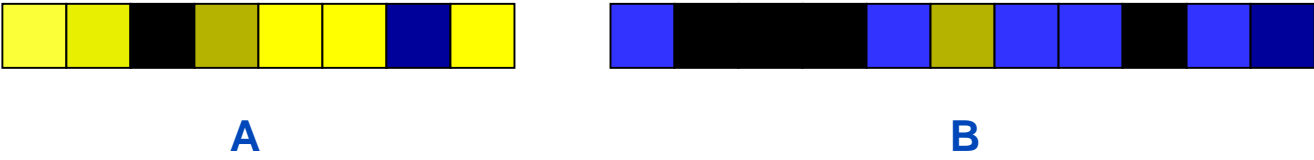





ok mistake

Predict the test samples and compare the predicted outcome to the true outcome

Prediction with 1 gene

Color coded expression levels of trainings samples



- Ms. Smith  → type A
- Ms. Smith  → type B
- Ms. Smith  → borderline

Which color shade is a good decision boundary?

Approach

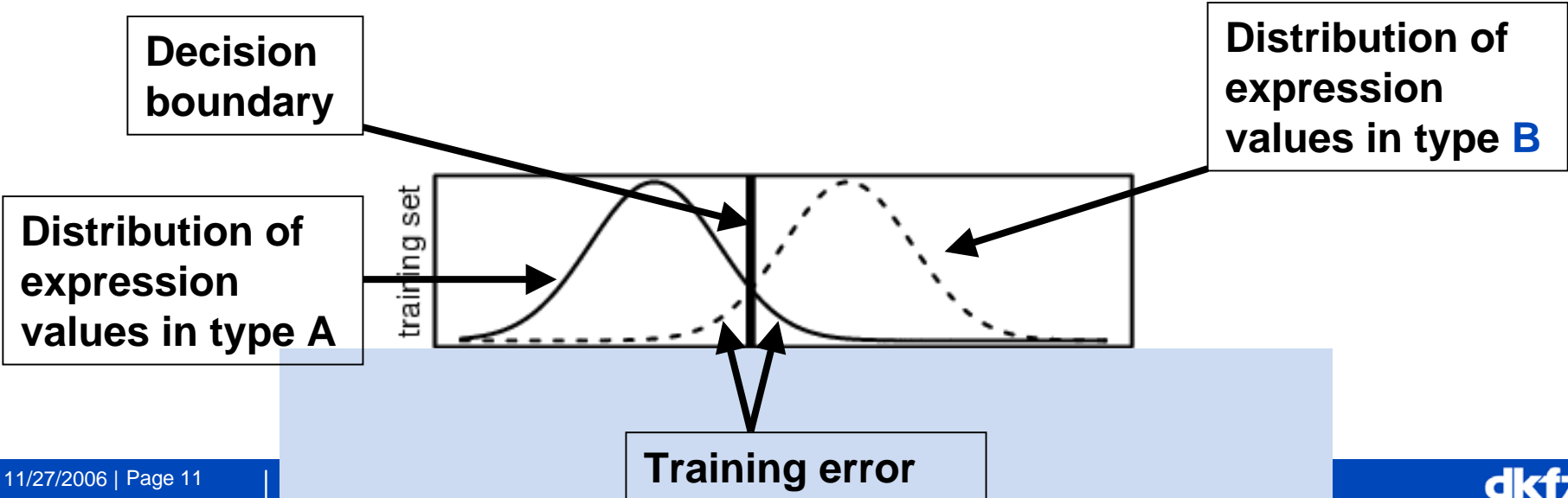
Use the decision boundary with the fewest misclassifications on the trainings samples

„ Smallest *training error* “

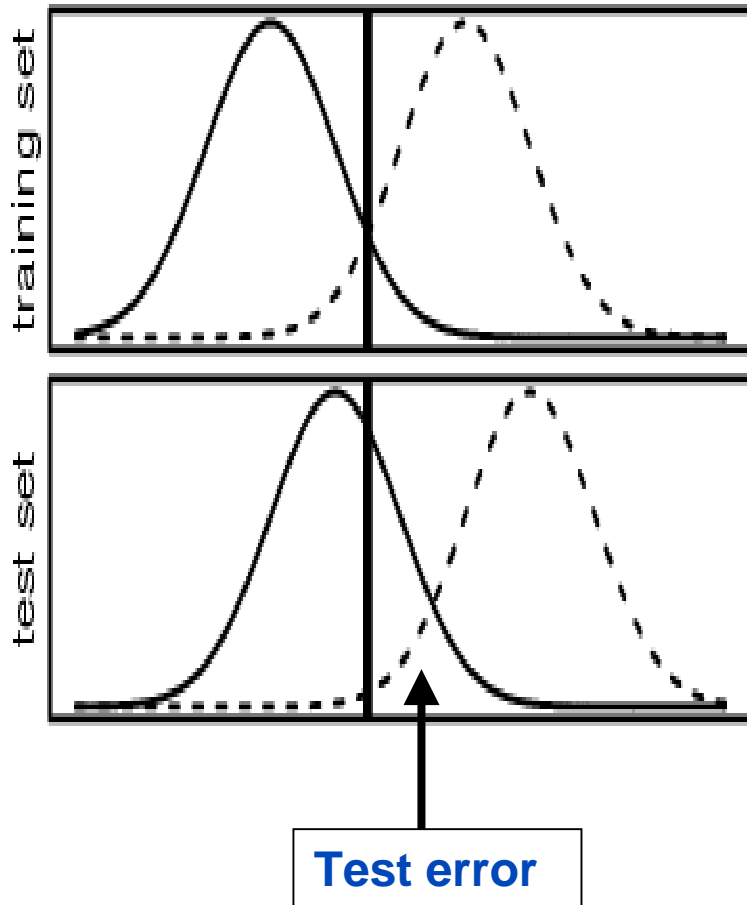


Zero training error is not possible in this case!

A more schematic illustration:



What about the test samples?



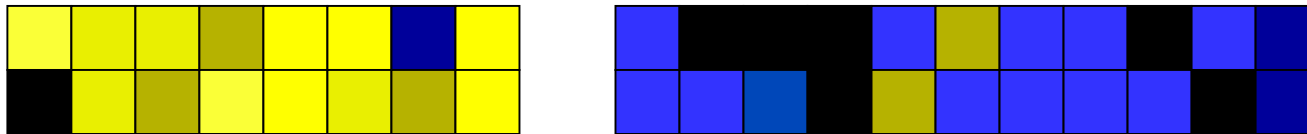
- The decision boundary was chosen to minimize the training error
- The two distributions of expression values for type A and B will be similar but not identical in the test data
- We can not adjust the decision boundary because we do not know the outcome of test samples
- Test errors are in average bigger than training errors
- This phenomenon is called *overfitting*

Prediction with 1 gene

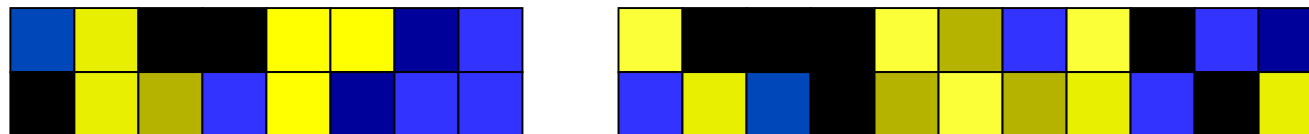


The gene is differentially expressed.

Prediction with 2 genes

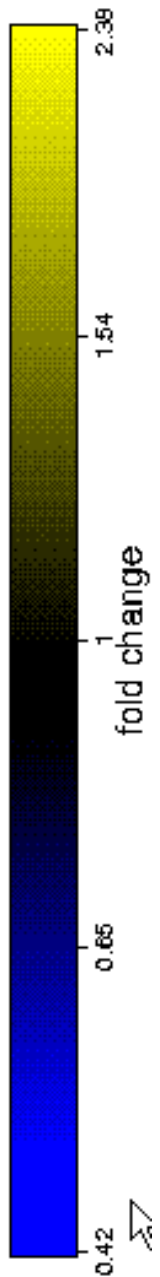


Both genes are differentially expressed.



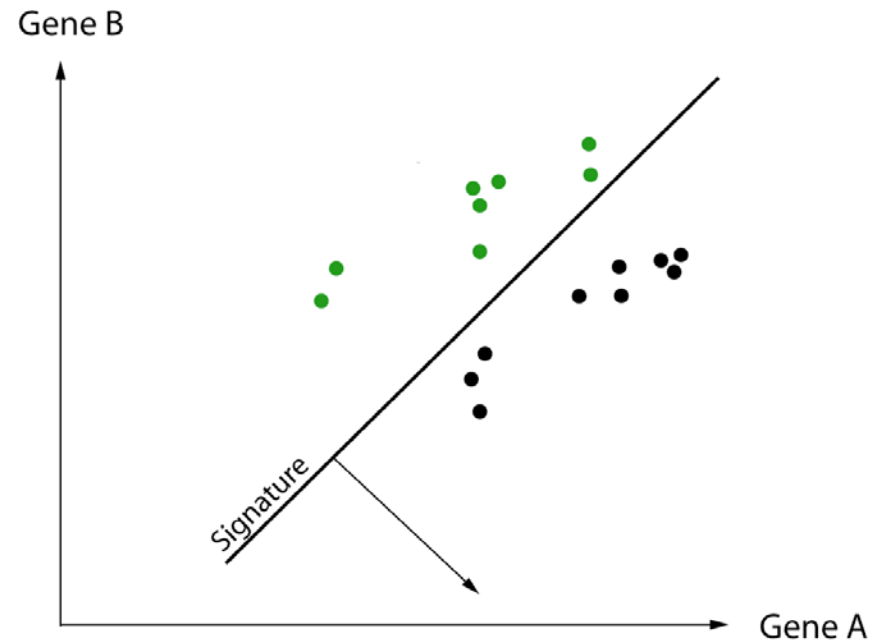
These genes are not differentially expressed.

Can they be of any use?

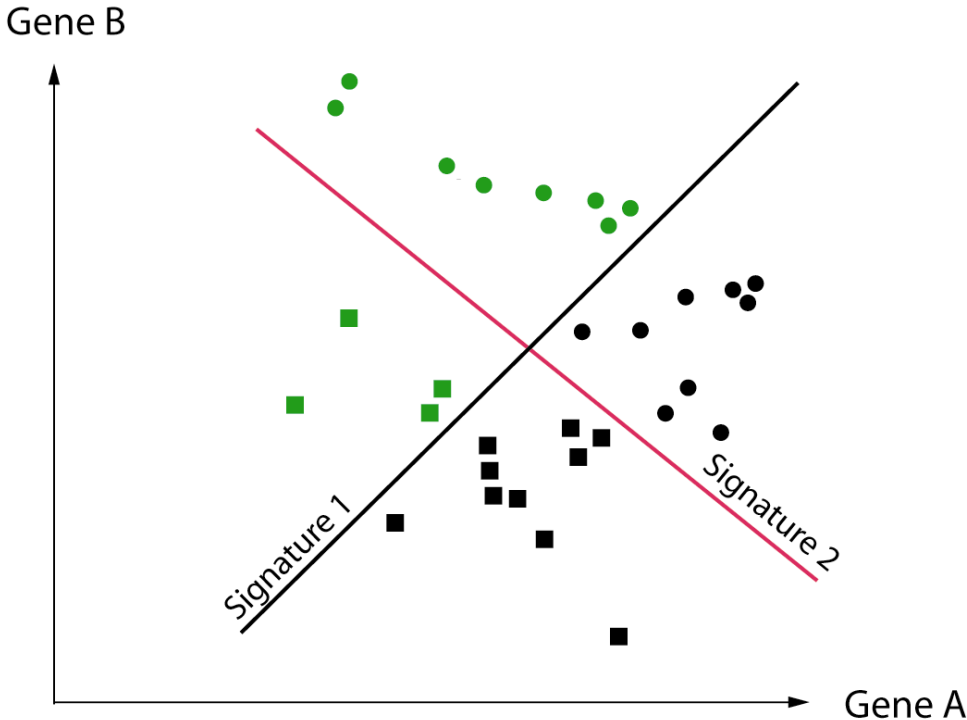


Interacting genes

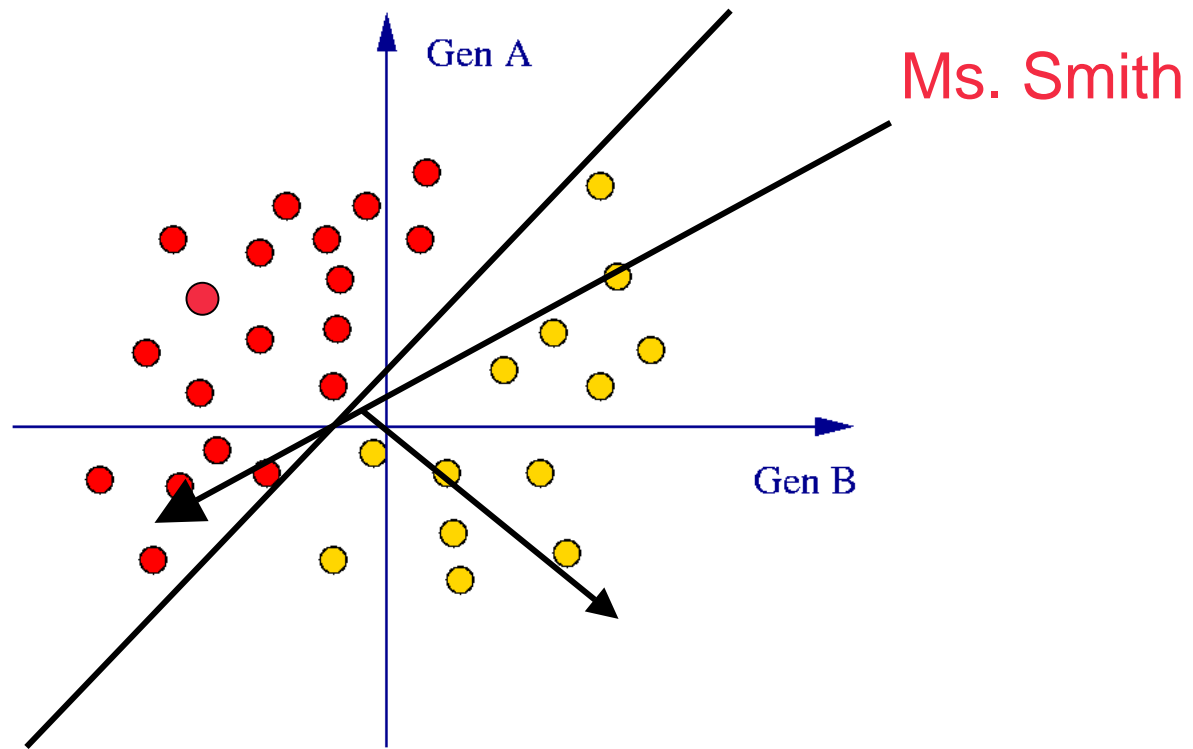
- Assume protein A binds to protein B and inhibits it
- The clinical phenotype is caused by active protein A
- Predictive information is in expression of A minus expression of B



Two different signatures based on the same genes

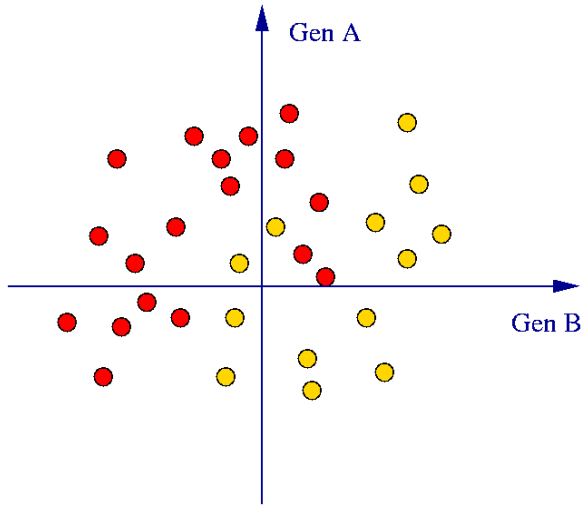


Calling signature genes markers for a certain disease is misleading!



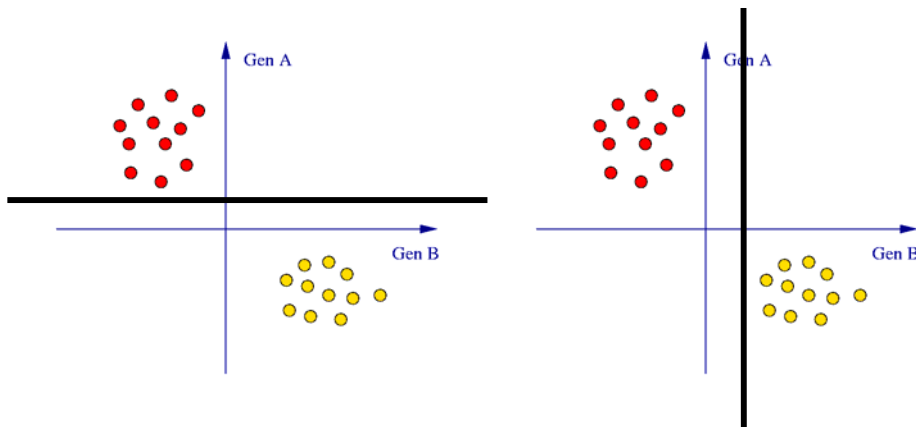
A decision boundary can be defined by a weighted sum (linear combination) of expression values

→ Separating Signature



Problem 1:

No separating line

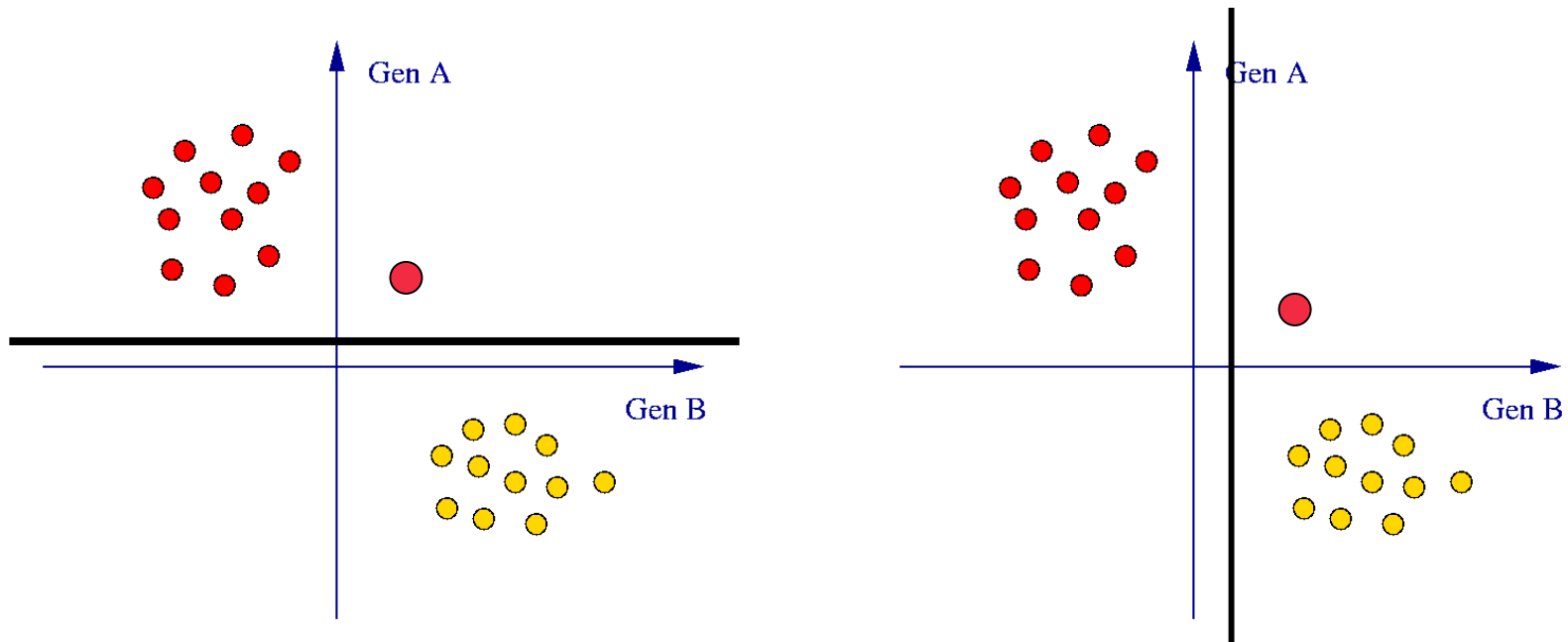


Problem 2:

To many separating lines

Why is this a problem?

What about Ms. Smith ?



*This problem is also related to overfitting ...
more soon*

How many genes

- Is this a biological or a statistical question?
- **Biology:** How many genes carry diagnostic information?
- **Statistics:** How many genes should we use for classification ?
- The microarray offers 30.000 genes or more

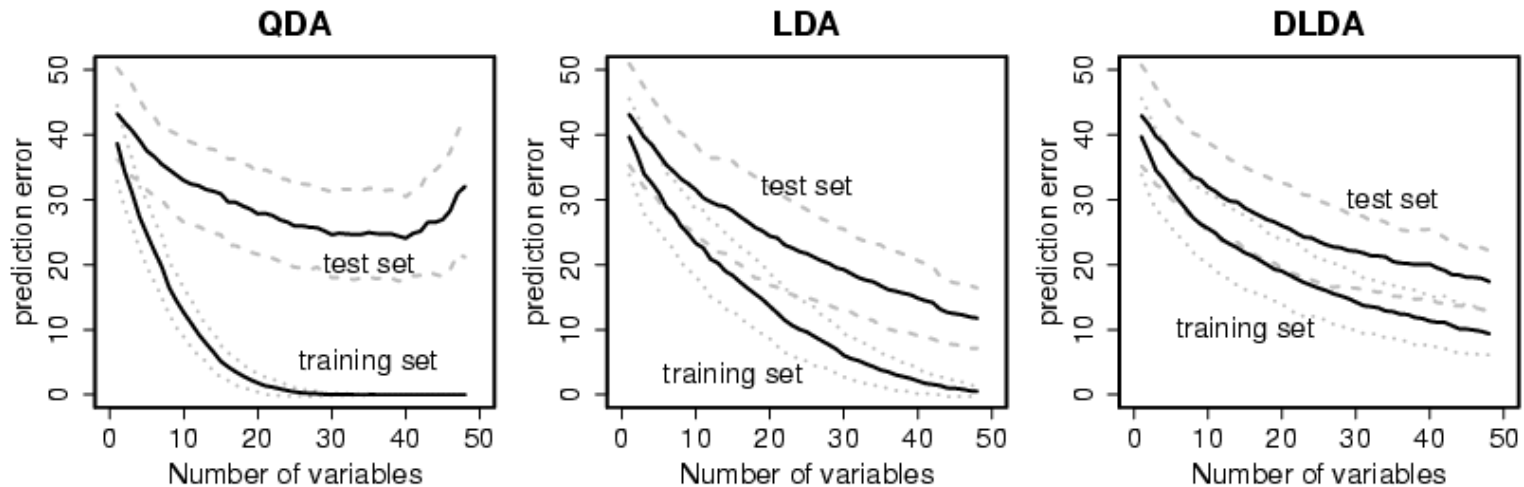
Finding the needle in the haystack

A common myth:

Classification information is restricted to a small number of genes, the challenge is to find them



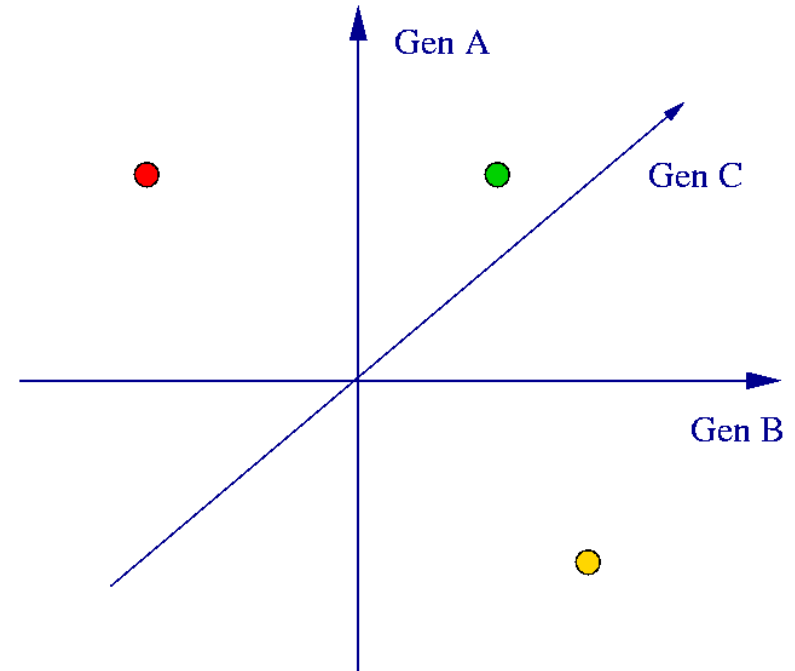
Using more genes



- The gap between training error and test error becomes wider
- There is a statistical reason for not including hundreds of genes in a model even if they are biologically effected

Prediction With 30,000 Genes

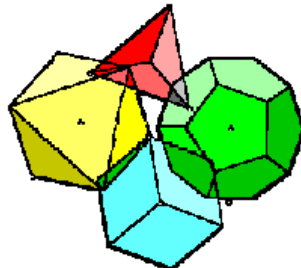
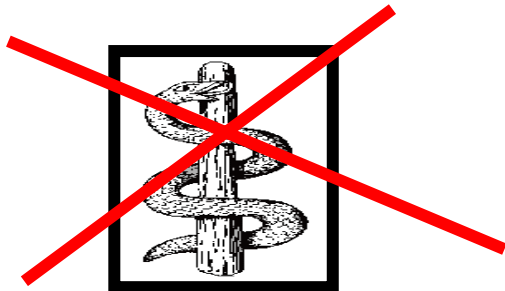
- With the microarray we have more genes than patients
- Think about this in three dimensions
- There are three genes, two patients with known diagnosis (red and yellow) and Ms. Smith (green)
- There is always one plane separating red and yellow with Ms. Smith on the yellow side and a second separating plane with Ms. Smith on the red side



The Overfitting Disaster

From the data alone we can not decide which genes are important for the diagnosis, nor can we give a reliable diagnosis for a new patient

This has little to do medicine. It is a geometrical problem.



The most important consequence of understanding the overfitting disaster:

If you find a separating signature, it does not mean (yet) that you have a top publication ...

... in most cases it means nothing.



Consequences of the Overfitting Disaster

- There always exist separating signatures caused by overfitting
 - *meaningless signatures* -
- Hopefully there is also a separating signature caused by a disease mechanism
 - *meaningful signatures* -
- We need to learn how to find and validate meaningful signatures

How to distinguish a meaningful signature from a meaningless signature?

The meaningless signature might be separating

– *small training error* –

... but it will not be predictive

– *large test error* –

The aim is not a separating signature but a predictive signature:

Good performance in clinical practice !!!

Strategies for Finding Meaningful Signatures

Later on we will discuss 3 possible approaches

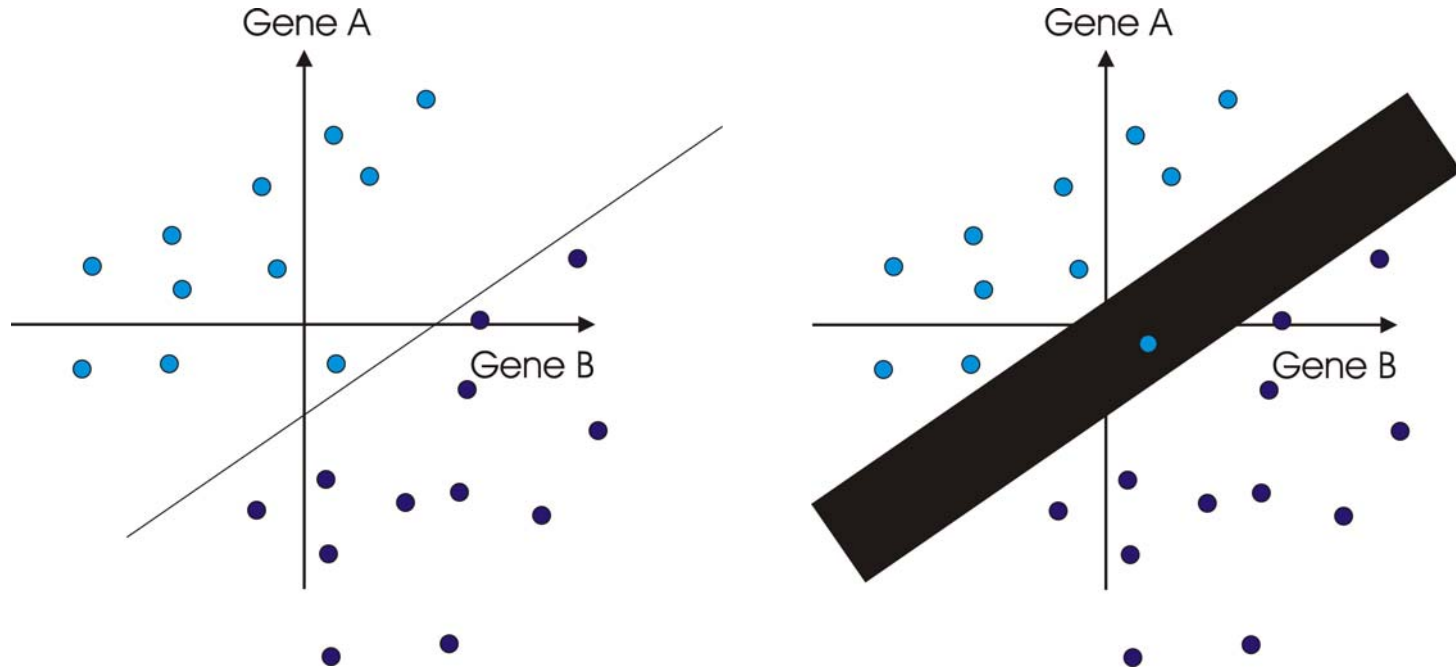
1. Gene selection followed by discriminant analysis (QDA,LDA,DLDA), and the PAM program (Markus' talk)
2. Support Vector Machines
3. Random Forests (Markus' talk)

What is the basis for this methods?

Gene Selection

- When considering all possible linear planes for separating the patient groups, we always find one that perfectly fits, without a biological reason for this.
- When considering only planes that depend on maximally 20 genes it is not guaranteed that we find a well fitting signature. If in spite of this it does exist, chances are good that it reflects transcriptional disorder.

Support Vector Machines



Large Margin Classifiers: If a large margin separation exists, chances are good that we found something relevant.

Regularization

- Both gene selection and Support Vector Machines restrict the set of a priori possible signatures. However, using different strategies.
- Gene selection wants a small number of genes in the signature - *sparse model* -
- SVMs want to maximize the distance between data points and the separating plane - *large margin models* –

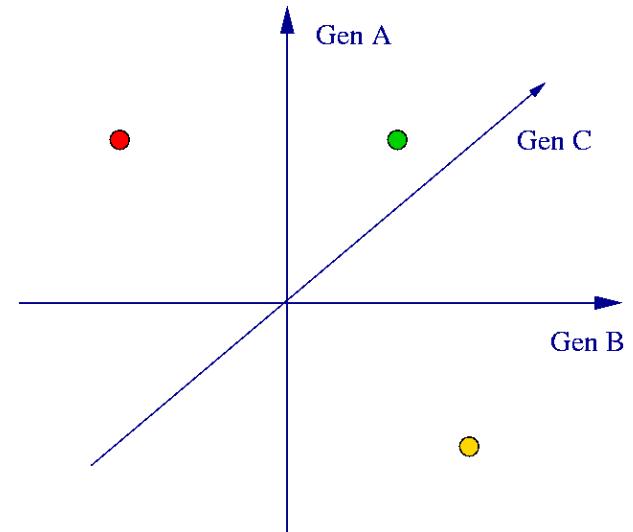
Literature Recommendations

- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2001
 - Ridge regression, LASSO, Kernel based methods, additive models, classification trees, bagging, boosting, neural nets, relevance vector machines, nearest-neighbors, transduction etc. etc.
- R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley-Interscience, 2001
 - Bayesian decision theory, nearest-neighbors, discriminant analysis, SVM, neural nets, CART, clustering, learning theory
- J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge Univ. Press, 2004
 - Kernel methods: SVM, SVR, kernel perceptron, kernel PCA, kernel based clustering, kernel functions for structured domains
- B. Schölkopf, A. Smola, Learning With Kernels, MIT Press, 2002
 - Kernel methods, regularization, optimization, statistical learning theory

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.



Evaluation

- The accuracy of a signature on the data it was learned from is biased
- Evaluating a signature requires **independent test data**
- → **Never ever think about reporting the training error instead of the test error**
- **The training error is more or less meaningless**
- **This test data must not be used for gene selection or model selection, otherwise the observed accuracy is biased**
- The test dataset should be as large as possible (just a few patients is not sufficient)

How Much Data?

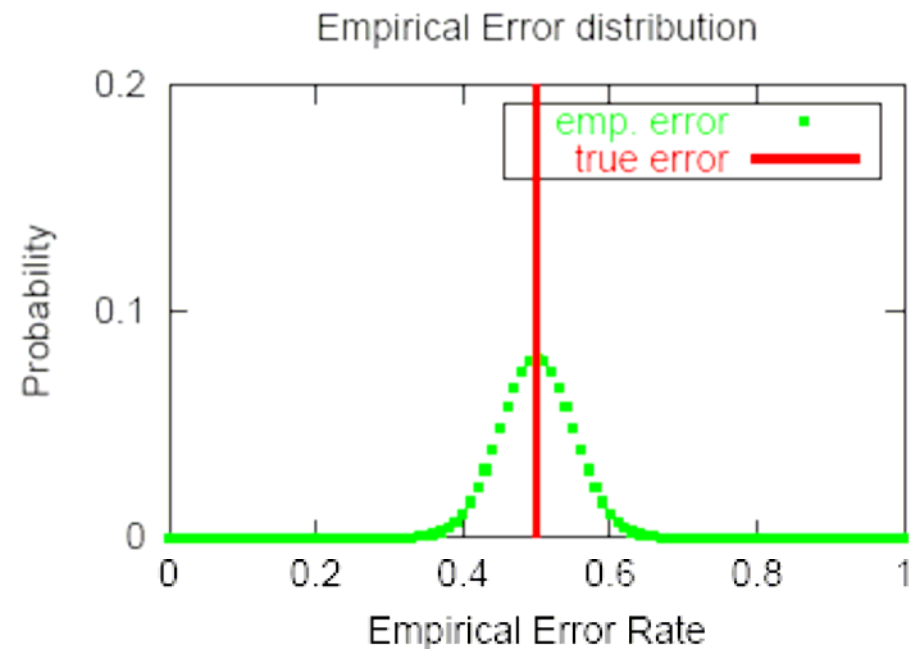
- Difficult to say, only probabilistic statements possible
- Distinguish between:
 - True error rate, if we had infinitely many data (unobservable)
 - Empirical error rate (observed)
- *Chernoff* bound for given classifier:

$$\Pr(|E_{emp} - E_{true}| \geq \varepsilon) \leq 2 \exp(-2n\varepsilon^2)$$

Example: How much data do we need to ensure $\Pr(|E_{emp} - E_{true}| \geq 0.1) \leq 0.05$?

→ $n \geq 184$

For $\Pr(|E_{emp} - E_{true}| \geq 0.05) \leq 0.05$ $n \geq 738!$



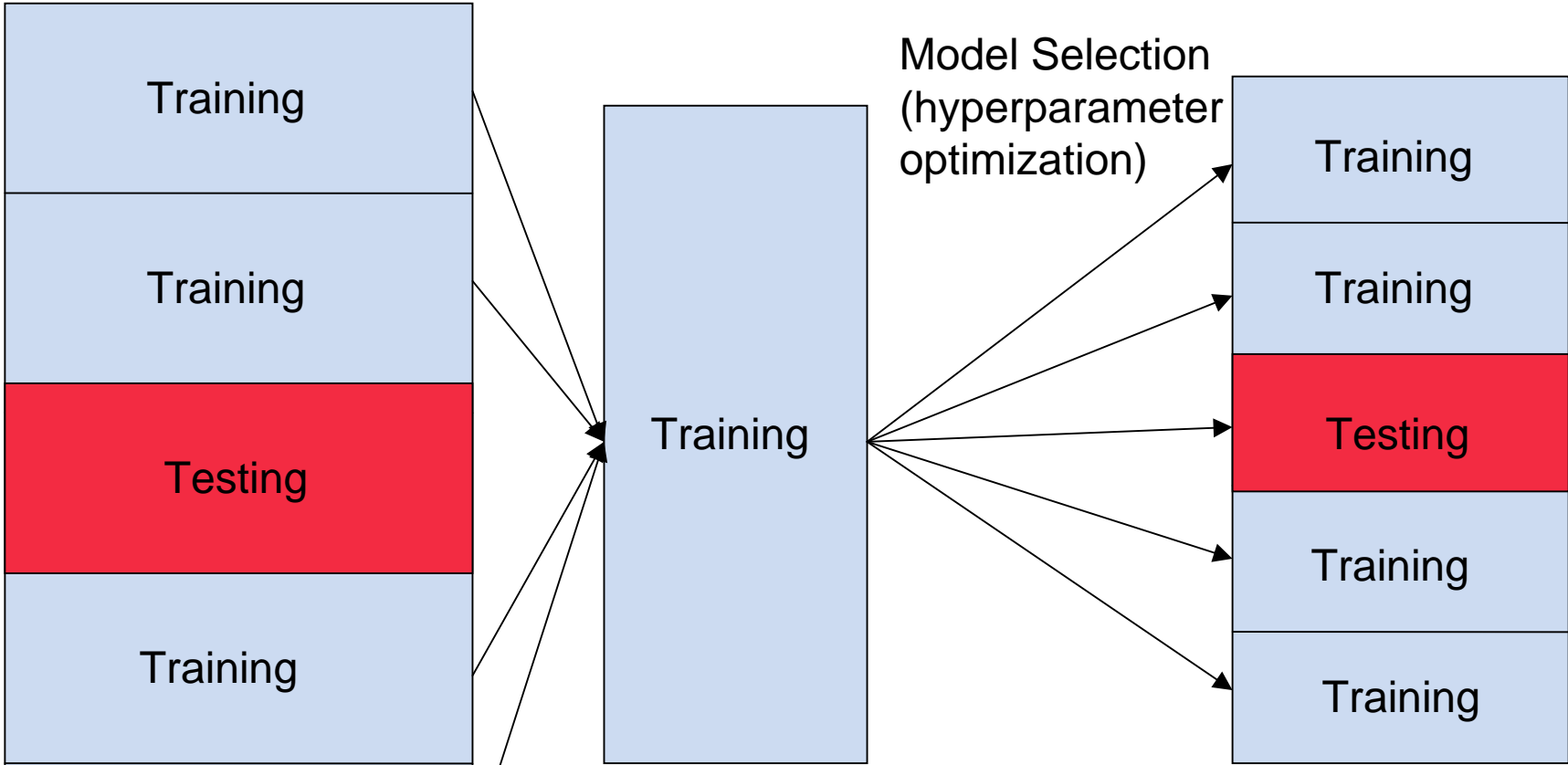
- Large test set needed to estimate true error rate relatively precisely!
- **Selection bias:** Result can depend heavily on the specific choice of the test data unless it is extremely large!

Cross-Validation



- Idea: reduction of selection bias
- Usually 5 or 10-fold cross-validation (i.e. 20%, 10% data left out for testing)
- Stratified (equal proportion of classes in all test sets) vs. non-stratified cross-validation
- **Important:**
 - 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

Nested Cross-Validation



Examples:

- Shrinkage parameter Δ in PAM
- Gene selection

Leave-One-Out Cross-Validation

1

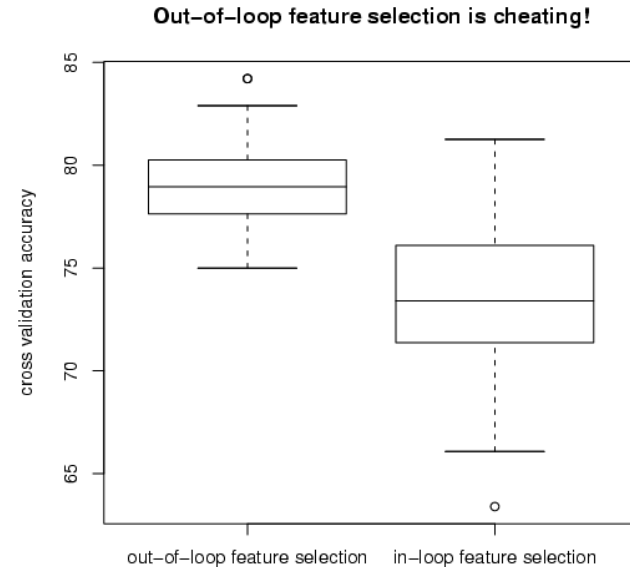


1

- Essentially the same
- But you only leave one sample out at a time and predict it using the others
- Good for smaller data sets

Variance of the Cross-Validation Estimator

- The CV estimator of prediction performance is by itself a random variable
- It thus has an expectation and a variance
- Repeating the nested CV procedure gives an estimate of expectation and variance



Nested 10-fold- CV

Variance from 100 random partitions

Comparison of Classifiers

- *Is classifier A better than classifier B?*

- => cannot be answered in general

Always situations where A is better than B and B is better than A

- **But: on a specific dataset we can indeed give an answer**

- => *Nested cross-validation procedure (repeated n times)*

- => A and B need to be trained and tested in exactly the same way!

- Same separations in training and test data

- Same method to measure classification accuracy

- **Example:**

CV repeat	A	B
1	90%	85%
2	80%	75%
3	85%	100%
4	70%	80%
5	100%	70%

Average A: 85% (Std. Err. 5%)

Average B: 82% (Std. Err. 5.14%)

→ Two-tailed paired t-test: $p = 0.7215$

No significant difference between A and B!

Measuring Classification Accuracy

- **Possibility 1:** count number of correctly classified examples

- Problem:

Class 1	Class -1
500	50

→ correctly classified:

Class 1	Class -1
400/500	0/50

→ **accuracy (%)**: $400/550 = 72.72\%$

- **Possibility 2:** *contingency table*

<i>true</i> class	<i>predicted</i>	
	Class 1	Class -1
Class 1	400/500	100/500
Class -1	50/50	0/50

true pos. rate (sensitivity) = $400/500 = 80\%$

true neg. rate (specificity) = $0/50 = 0\%$

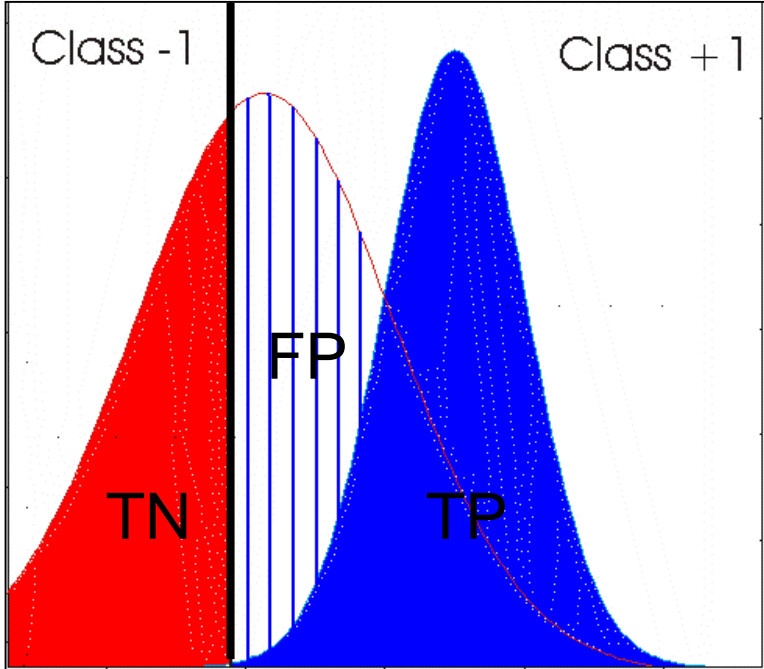
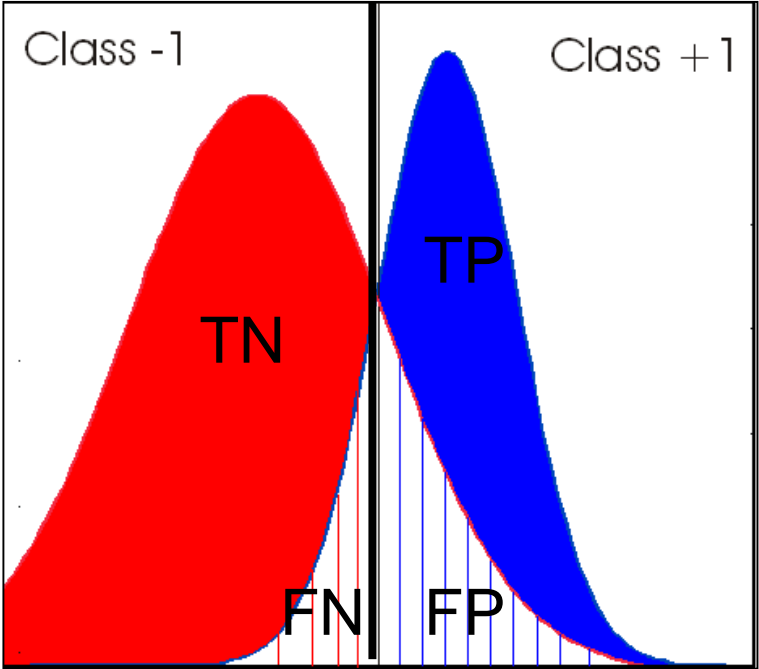
false pos. rate = $50/50 = 100\%$

false neg. rate = $100/500 = 20\%$

- $(\text{sensitivity} + \text{specificity}) / 2 = (80\% + 0\%) / 2 = 40\%$ (**balanced accuracy**)

Measuring Classification Accuracy (2)

- *What happens, if e.g. a false positive is worse than a false negative?*
 - → Assumption so far: sensitivity and specificity are equal goals
- General threshold classifiers:



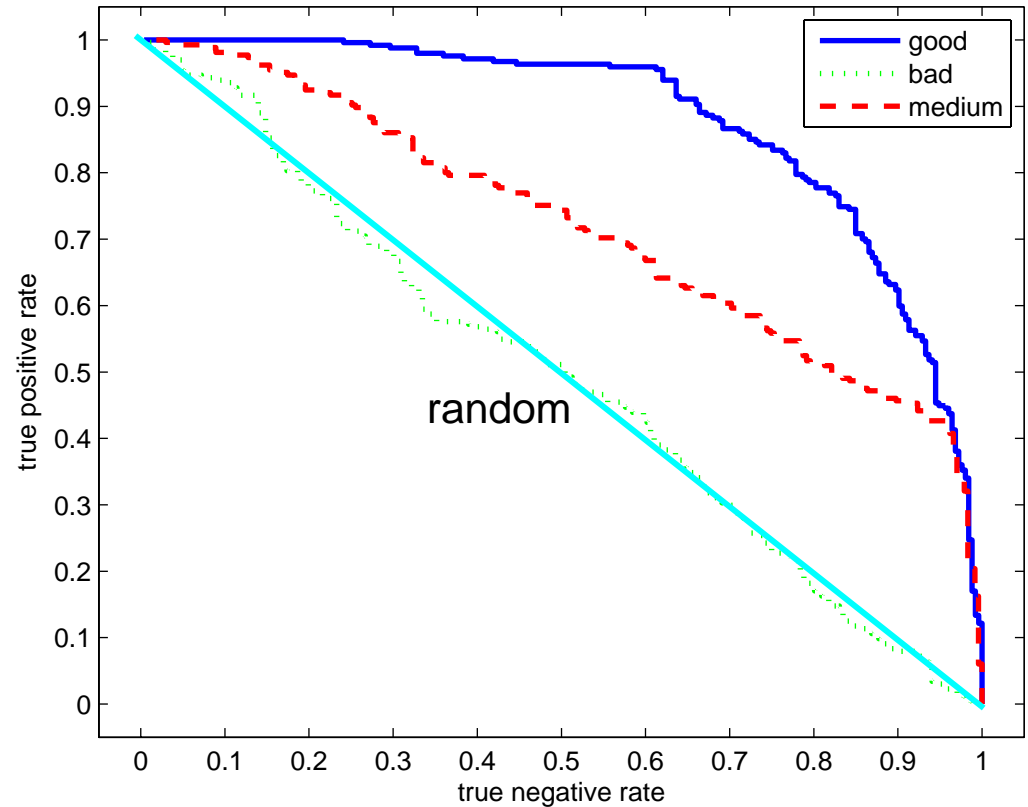
$$f(x) := \begin{cases} \text{class 1} & \text{if } g(x) \geq t \\ \text{class -1} & \text{otherwise} \end{cases}$$

Receiver Operator Characteristic (ROC) Curves

- Trade-off between sensitivity and specificity
- Each point on the ROC curve corresponds to a specific weighting of sensitivity vs. specificity
- *Area under ROC curve* summarizes classifier performance **for all possible thresholds** (and hence weightings of sens. vs. spec.):

$$AUC = \Pr(g(x^+) \geq g(x^-))$$

$$\approx \frac{1}{n * m} \sum_{i=1}^n \sum_{j=1}^m \mathbf{1}(g(x^+_i) \geq g(x^-_j))$$



DOs AND DON'Ts :

1. Decide on your diagnosis model(s) (PAM, SVM, etc...) and **do not change your mind later on**
2. Think about how you want to measure classification accuracy
3. Use nested k-fold cross-validation procedure (repeated n times) to assess prediction performance:
 - Train and optimize your model using the data in the current **training set** only → (select genes, define centroids, calculate normal vectors for large margin separators, perform model selection ...)
 - Put the data in the current **test set** away ... far away
 - **Do not even think of touching the test data at this time**
 - Apply the model to the current **test** data ...
 - **Do not even think of changing the model at this time**
4. Do steps 1-3 only **once** and accept the result ...
 - **Do not even think of optimizing this procedure**

External Validation and Documentation

- Documenting a signature is conceptually different from giving a list of genes, although 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
 - → Add on normalization

Establishing a signature



Split Data into
Training and
Test Data

Test data only:
Internal validation
Full quantitative
specification

External
Validations

Training data only:
Machine Learning

- select genes
- find the optimal number of genes
- learn model parameters

Acknowledgements

- Rainer Spang, University of Regensburg
- Markus Ruschhaupt, DKFZ
- Tim Beißbarth, DKFZ